JDSE2017
Junior Conference on Data Science and Engineering
2nd edition

JDSE2017
PARIS-SACLAY

Junior Conference on Data Science and Engineering

2ND edition

contact@junior-data-science.org

DATABASES AND ONTOLOGIES

TALK SESSION

On the Automatic Distribution and Parallelization of a miRNA Prediction Algorithm using Spark and Mesos frameworks

Talk submission

Alexandre PROTAT, Laurent POLIGNY, Nazim AGOULMINE, Fariza TAHI

IBISC, Univ Evry, Universit Paris-Saclay, 91025, Evry, France.

Abstract. miRNAFold is a fast *ab-initio* program for miRNA precursor prediction. In order to apply this algorithm to human genomes in reasonable amount of time, a scalable implementation has to be made to run on HPC. Here we present a version of miRNAFold integrated in a Spark workflow to run on a Mesos cluster, parallelizing the miRNAFold algorithm. We found this Spark version get more significant speedup by launching one spark job per chromosome than by launching one spark job for whole genome.

Keywords: HPC, miRNA, Spark, Mesos

1 Introduction

MicroRNAs (miRNAs) are small non-coding RNAs involved in many biological regulation process. Predicting miRNA precursors and their structure from DNA sequences is an important field in biology. MiRNAFold [6] is a fast *ab-initio* software able to predict those miRNAs and their structure. A web server implementation of miRNAFold has been developped [5], available on EvryRNA platform (https://evryrna.ibisc.univ-evry.fr/evryrna/mirnafold).

miRNAFold algorithm predicts miRNA hairpins using a sliding window over the DNA. For each window, a base pairing matrix is built, then the matrix is used for searching the anchor of the hairpin, then extending it and validating using some statistical criteria. miRNAFold is the fastest algorithm in the literature for miRNA prediction. It's complexity is $O(L^2.N)$, where L is the length of the window, and N the size of the sequence. It takes less than 1 second to process on a 1K nucleotide sequence. However, running miRNAFold algorithm on a complete genome takes high execution time (aroud 14h to process a whole human genome on our servers).

In this project we developed a scalable and adaptable implementation of miRNAFold, running on a cluster of servers available in our lab. This algorithm is automatically distributed over several servers and parallelized on available server cores, thanks to Apache Spark [7], Mesos [2] and Hadoop [3] software. Benefits of parallelizing and distributing an application are well known[1][4]. With this implementation, we expected to process a whole human genome in a reasonable amount of time, approximately one hour, which was successfully reached.

2 Methods

The miRNAFold algorithm predicts hairpins over DNA sequences using a sliding window. We parallelized this program by defining a batch as a group of successive overlapping windows over the genome.
The miRNAFold program has been wrapped as Java native function in a Java Spark workflow. This workflow first generates a distributed file on the cluster using the Hadoop distributed filesystem (HDFS) [3], writing one batch per line. Then Spark reads the batches file in a distributed and parallelized fashion using resources provided by Mesos, the cluster's resource negotiator.

To execute this Spark workflow, we have created a cluster of virtual machines on the same bare metal machine. We have installed Spark, Mesos and HDFS on those machines, one node is a Mesos Master and HDFS namenode, others are Mesos Slaves and HDFS datanodes. Spark automatically adapts its parallelization and distribution level to the provided resources by Mesos to execute its workflow. Mesos provides resources depending on cluster occupancy, and queued jobs. Batches files are read as blocks of 32 Mb. If many cores are available on resources provided by Mesos, Spark reads and processes those blocks in parallel.

We have implemented and evaluated two approaches:

- In the first approach we launched one Spark job for the whole human genome. It generates a big batch file stored in HDFS.
- In the other approach, each chromosome is processed on the cluster as a Spark job and scheduled by Mesos.

3 Results

We have executed the two approaches on a cluster of 4 servers, each server has 10 cores. The results show that a total speedup of 7.41 for the first approach (Table 1), with approximately one quarter of spark's tasks that fail. The batches file generation time could be significant (7 minutes in our final cluster configuration) but the decisive criterion was the size of the batch file. If this file is too large it will need more parallelism to proceed it and this will exceed the RAM that our cluster can actually provide. Since this approach handles a whole genome as a single job, if some tasks fail too many times, the job is killed and the successfully intermediary results are lost.

The second approach gives us much better results with a speedup of 7.55 (Table 1) and no task failure in the best case. In addition, this approach has also advantage that all batches files are generated together, simultaneously. However, those results are heavily dependent from the Mesos scheduling, as all chromosome jobs cannot be processed at the same time. Better results (i.e with no failed tasks) are reached when small chromosomes are processed at first (when the occupancy is high), and bigger chromosomes after (i.e when more resources are available) as their parallelism is optimal.

4 Conclusion

MiRNA precursor prediction is of crucial importance for discover novel miRNAs and understand transcriptional regulation networks. However, processing genomes could be very expensive in term of processing and time.

Table 1. miRNAFold execution time comparison, iterative versus two spark approaches.

version	average-time	speedup
Iterative	14h19	1
First approach	1h59	7.41
Second approach	1h18	7.55

All tests have been done on virtual machines located on the same machine bare-metal machine (including cluster nodes).
Iterative miRNAFold virtual machine: 2 CPU cores, 100 Go HDD.
Spark-Mesos cluster: 4 virtual machines of 10 CPU cores, 38 Go RAM and 65 Gb HDD.

Parallelization has of course proved its speedup potential. Nevertheless, the way to parallelize and the design can have a high impact on the result. The whole genome batch locks more resources provided by Mesos, than many unitary chromosome batches, with some task failures. The method using unitary chromosome batches highlights better performances because of its better job adaptability and better usage of the cluster resources.

However, unitary chromosome processing are less predictable in terms of duration and success due to Mesos scheduling strategy. It has also a high ability to recover success job's results. To improve the Mesos scheduling, a program should be developed with a prior knowledge or a machine learning method.

References

1. Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.
2. Benjamin Hindman, Andy Konwinski, Matei Zaharia, Ali Ghodsi, Anthony D Joseph, Randy H Katz, Scott Shenker, and Ion Stoica. Mesos: A platform for fine-grained resource sharing in the data center. In *NSDI*, volume 11, pages 22–22, 2011.
3. Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*, pages 1–10. IEEE, 2010.
4. Lorna A Smith, J Mark Bull, and J Obdrizalek. A parallel java grande benchmark suite. In *Supercomputing, ACM/IEEE 2001 Conference*, pages 6–6. IEEE, 2001.
5. Christophe Tav, Sbastien Tempel, Laurent Poligny, and Fariza Tahi. miRNAFold: a web server for fast miRNA precursor prediction in genomes. *Nucleic Acids Research*, 44(W1):W181–W184, July 2016.
6. Sbastien Tempel and Fariza Tahi. A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic Acids Research*, 40(11):e80–e80, June 2012.
7. Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. Apache spark: A unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, 2016.

Repairing regular expressions by adding missing words

Thomas Rebele[1], Katerina Tzompanaki[2], Fabian Suchanek[1]

[1]Télécom ParisTech, [2]Université Cergy-Pontoise

Abstract. Regular expressions are used in many information extraction systems like YAGO, DBpedia, Gate and SystemT. However, they sometimes do not match what their creator wanted to find. We propose a novel algorithm that automatically adds missing words by creating disjunctions at the appropriate positions, guided by an approximate matching.

Keywords: repairing regular expressions, missing words, improving recall

1 Introduction

Regexes can describe concisely what a user wants to find. They help, e.g., to extract dates or literals in information extraction, to find DNA sequences for families of proteins bioinformatics, or to detect attacks by analyzing log files in network security. Creating a good regex needs expertise, but the regex might still not match all the positive examples. Adapting the regex to include the missed positive examples might need quite some time. We therefore investigated algorithms that change a regex such that it matches a given set of missing words. The user needs to provide a set of missing words, and a set of negative examples, so that our algorithm can avoid generalizing too much.

Regex definition. A regex r is an expression of the following form: an empty expression ϵ, a character a or character class $[a-z]$, a concatenation AB, a disjunction $A|B$, or a Kleene star $A*$, where A and B are regexes. The language of the regex $L(r)$ is defined in the usual way. We use the shorthand notation $A\{n,m\} = \underbrace{A \cdots A}_{n-\text{times}} \underbrace{(\varepsilon|A(\varepsilon|\cdots))}_{m-\text{times nested}}$.

Problem statement. Given a regex r, a set of missing words S, a set of negative examples E^-, and a relaxation factor α, find a regex r', such that $L(r) \cup S \subseteq L(r')$ and $|E^- \cap L(r')| \leq \alpha |E^- \cap L(r)|$.

Related work. Approaches for *regex learning* [3, 2, 7] do not adapt a regex, but learn it from scratch instead. Others which employ *regex transformation* focus on removing false positives [4], or generalizing character classes, and quantifier ranges [5, 1]. They rely on many positive examples. Our approach, however, uses only a few positive examples.

2 Algorithm

Our algorithm applies the following preprocessing steps. It adds a special character to the start and end of the missing words and the regex, which simplifies the border cases. It expands the shorthand notation $A\{n, m\}$ in the regex. It embeds all non concatenations in a concatenation.

To find the gaps, where it needs to change the regex, it applies an approximate matching from the characters of the words to the leaves of the regex syntax tree. Then it recursively descends into the regex to apply the modifications. Finally it applies some basic transformation to simplify the regex.

Finding gaps. For the approximate matching we use [6], which implements a dynamic programming algorithm in a similar way to the edit distance algorithms, approximate string-string matching or longest common subsequence algorithms. We adopted the algorithm to return correspondences between characters of the word and the leaves of the regex, instead of substitutions.

The gaps represent discontinuities in the matching. A gap has a start, and a stop character, and fulfills the following conditions. Start and stop character don't have other matched characters in-between, and both are mapped to regex leaves by the matching. Either they have unmatched characters in-between, or the start regex leaf doesn't directly precede the stop regex leaf, or both.

Recursive fixing. The algorithm takes the list of gaps and recursively descends into the regex tree. At Kleene stars and disjunctions, it filters the list of gaps for every child, such that a child gets only repaired for the gaps concerning its subtrees. At concatenations the algorithm creates disjunctions that embed those children that occur between children containing gap leaves. If such groups overlap, we split them into non-overlapping parts at the borders, and create a disjunction for each such group. We register every newly created disjunction at those gaps that cover them.

Afterwards we add the substring between start and end character of each gap at one of the registered disjunctions. After every modification we check whether the ratio of false positives in E^- is still smaller than α. If not, we try the next registered disjunction of the gap. If we cannot add the substring, we undo all changes for the string s that the gap belongs to, and add the string as $r' = \ldots|s$.

3 Experiments

We conducted experiments on the ReLie *Relie dataset*[1] and *Enron-Random dataset*[2]. For every dataset of the ReLie dataset, we have 5 regexes, provided by our colleagues. For the Enron-Random dataset we use the regexes of [1]. As baselines we take the original regex, a disjunction of the original regex and the missing words, and the regex **.***. For every regex we randomly pick 10 missing words. For ReLIE we pick the negative examples as E^-. For Enron-Random we

[1] `http://dbgroup.eecs.umich.edu/regexLearning/`

[2] `http://www.cs.cmu.edu/~einat/datasets.html`

remove the positive matches from the text to obtain E^-. In Table 1 we show the F1 measure for these datasets and two values of parameter α.

task	original	dis-baseline	star-baseline	repaired regexes		
	r	$r\|s_1\|\cdots\|s_n$	$.*$	α =1.0	α =1.1	α =1.2
Relie/phonenum	78.1	77.3	12.1	+4.1	+7.1	**+9.1**
Relie/coursenum	64.7	69.3	70.3	+.5	**+.7**	+.5
Relie/softwarename	10.4	15.2	9.8	+5.1	**+6.3**	**+6.3**
Relie/urls	65.1	65.8	30.8	+4.9	**+5.4**	**+5.4**
Enron/phone	64.6	64.6	.1	**+23.5**	**+23.5**	**+23.5**
Enron/date	70.1	70.3	.0	**+5.2**	+5.0	**+5.2**

Table 1. F1 measure for different values of the parameter α, improvement over the dis-baseline in percentage points.

4 Discussion

We have shown an algorithm for automatically adding missing words to regular expressions. Our algorithm manages to generalize the regexes in the sense, that it increases their recall without deteriorating their precision. For future work we plan to investigate other feedback functions, and work on minimizing the length of the repaired regex.

References

[1] Rohit Babbar and Nidhi Singh. "Clustering based approach to learning regular expressions over large alphabet for noisy unstructured text". In: *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data, AND 2010, (in conjunction with CIKM 2010)*. Toronto, Ontario, Canada: ACM, 2010.

[2] Alberto Bartoli et al. "Automatic Synthesis of Regular Expressions from Examples". In: *IEEE Computer* 47.12 (2014).

[3] Falk Brauer et al. "Enabling information extraction by inference of regular expressions from sample entities". In: *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011*. Glasgow, United Kingdom: ACM, 2011.

[4] Yunyao Li et al. "Regular expression learning for information extraction". In: *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008*. Honolulu, Hawaii, USA: ACL, 2008.

[5] Karin Murthy, Deepak Padmanabhan, and Prasad Deshpande. "Improving Recall of Regular Expressions for Information Extraction". In: *Web Information Systems Engineering - WISE 2012*. Paphos, Cyprus: Springer, 2012, pp. 455–467.

[6] Eugene W Myers and Webb Miller. "Approximate matching of regular expressions". In: *Bulletin of mathematical biology* 51.1 (1989).

[7] Paul Prasse et al. "Learning to Identify Concise Regular Expressions That Describe Email Campaigns". In: *J. Mach. Learn. Res.* 16.1 (Jan. 2015).

Grouping Answers in Ontology-Based Query Answering: the RDFS case

Maxime Buron and Michaël Thomazo

INRIA, Université Paris-Saclay
LIX, École Polytechnique, Université Paris-Saclay

Abstract. We aim at easing the use of semantic and structured data by improving the restitution of answers. Instead of providing a set/ ranked list of answers, we devise techniques to group answers and to label the group in an informative way, easing further exploration of the answers.

Keywords: databases, graphs, ontologies, semantic web

1 Motivations

A wealth of digital resources is available on the Internet, and efficiently exploiting them is still a challenge that has far-reaching applications. An important feature of resources that are at the core of Semantic Web is that they are both structured (links exist between different entities) and semantic (formal relationships exist between different terms of a vocabulary). Despite a large amount of data being under open licenses, we believe that the exploitation of these resources has not reached its full potential, because of hurdles such as the difficulty to express structured queries on these resources or the difficulty of apprehending their results. In this article, we tackle the latter problem: providing the results of a structured and semantic query in an "interesting" way. We consider the framework of evaluating queries on a dataset enriched with an ontology, known as Ontology-Based Query Answering (OBQA, see [MT14]). We consider the answers to be crisp: a tuple is an answer or not, but there is no notion of relevance as it is classically the case in web search. We aim at outputing groups of answers together with a label for each group, providing an explanation about what is in the group. The labels of the groups enable the user to have an overview of the complete set of answers. Each group can then recursively be explored. For example with a dataset containing (among other) information about persons like in DBpedia [LIJ+15] or YAGO [MBS15], we can query for all the politicians. An interesting labellisation for this query could be a set of labels where each label selects only persons that are in a given party, together with a label that selects politicians that are not known to be in any party. An other labellisation might regroup politicians by their occupation : senator, minister, deputy, mayor, ...

In the remaining of this paper, we introduce the technical framework, state the problem and provide a high-level view of the approach we propose.

2 Technical Context

We focus on evaluating conjunctive queries on knowledge bases, which are pairs $(I, \mathcal{R})$ where I is a set of facts and $\mathcal{R}$ describes domain knowledge. There is a variety of formalisms to represent both data (based on graphs or tables, for instance) and knowledge (such as description logics [BCM+03], existential rules [MT14]). In this paper, we will focus on RDF and RDFS [BG14]. Within RDF, data is represented by triples of the shape (subject, predicate, object), such as `:entity1 :name "Peter"`, `:entity2 :children :entity1` or `:entity1 rdfs:type :Person`. RDFS is used to describe domain knowledge, such as "every patient is a person" (`:Patient rdfs:subClassOf :Person`), or "anybody that has a child is a person" (`:hasChild rdfs:domain :Person`).

OBQA is the problem of finding answers to a query that are entailed by a knowledge base. In other words, not only the (explicit) data triples should be taken into account, but also entailed (implicit) triples. The distinction can be seen in the previous example: without considering the ontology, only `:entity1` would be considered as an answer. However, knowing that the domain of `:hasChild` has type `:Person`, we retrieve `:entity2` as well.

Two main approaches have been devised to solve OBQA: the first one is *materialization*, which consists in making implicit triples explicit, by adding them to the set of data triples. The other main approach is call *query rewriting*, and it consists in reformulating the query in such way that it can be directly evaluated on the data as if there was no ontology, while keeping the same semantics. For instance, the query $q(x) := x$ `rdfs:type :Person` would be reformulated under the previously described ontology as a union of two conjunctive queries, $q(x) := x$ `rdfs:type :Person` and $q'(x) := x$ `:hasChild` y.

3 Grouping Answers

Given an instance I, an ontology $\mathcal{R}$ and a query q, our goal is to provide a set $\mathcal{L} = \{(l_1, S_1), \ldots, (l_n, S_n)\}$ of pairs (l_i, S_i) such that *(i)* S_i is a subset of the answers of q over I and $\mathcal{R}$ and *(ii)* l_i is a *label* for S_i, which verify that S_i is the set of answers of $q \wedge l_i$ over I and $\mathcal{R}$. Such a set is a *labellisation*. A labellisation is *covering* if each answer belongs to at least one S_i. It is *simple* if there are no pair (i, j) such that $i \neq j$ and l_i entails l_j when considering $\mathcal{R}$.

There is a trivial covering and simple labellisation: $\mathcal{L} = \{(q, S)\}$, where S is the set of answers of q over I and $\mathcal{R}$. However, this does not provide any more information than the set of answers, and our current task is to formalise the *informativeness* of a labellisation. We currently investigate several ways to define this notion, exploring graph-based notions and information theory. Another question of interest is the categorization of possible approaches. We currently devise such a categorization by studying how changes in the input affect the output: for instance, does the presence of an atom that is not necessary to answer the query affects the output of the algorithm?

We devised an algorithm that builds covering and simple labellisations, and takes as additional input parameter the expected size of the output labellisation.

Labels are sets of queries, each of which being obtained through a classical query rewriting technique. Two queries are put in the same set in order to obtain groups with the most specific labels of comparable sizes. Further theoretical properties of the algorithm are currently under investigation, and a demonstration of the prototype will be proposed, querying both synthetic and real-world data.

4 Related Work

The main approach to help the user to apprehend the answer set is to rank them according to criteria that are not necessarily known by the user. A classical problem is to output the top k answers [IBS08]. We depart from this approach by restituting the totality of the answer set, providing groups of answers that are explainable in terms of domain knowledge. The task of grouping answer is reminiscent from clustering [DHS00]. We do not use such algorithms here because it does not provide semantic explanations of the performed grouping, but such techniques are an interesting point of comparison in a user study. We also do not need to define a distance on query answers. A shortcoming of our approach is that we cannot guarantee to output partitions of the answers.

References

BCM+03. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

BG14. D. Brickley and R.V. Guha. RDF Schema 1.1. Technical report, W3C, Feb 2014.

DHS00. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

IBS08. I. F. Ilyas, G. Beskales, and M. A. Soliman. A survey of top-k query processing techniques in relational database systems. *ACM Comput. Surv.*, 40(4), 2008.

LIJ+15. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.

MBS15. F. Mahdisoltani, J. Biega, and F. M. Suchanek. YAGO3: A knowledge base from multilingual wikipedias. In *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*, 2015.

MT14. M.-L. Mugnier and M. Thomazo. An introduction to ontology-based query answering with existential rules. In *Reasoning Web. Reasoning on the Web in the Big Data Era - 10th International Summer School 2014, Athens, Greece, September 8-13, 2014. Proceedings*, pages 245–278, 2014.

Modeling Spatially-Correlated Cellular Networks by Applying Inhomogeneous Poisson Point Processes

Oral Talk

Shanshan WANG, Marco Di RENZO

CNRS-Laboratoire des signaux et systèmes-CentraleSupelec-Université Paris Saclay

Abstract. The distribution of base stations (BSs) are usually modelled in a Poisson Point Process (PPP) manner. While random deployments are not accurate for macro base stations. The non-PPP based approaches are much less mathematically tractable than PPP-based approach. This paper proposes a new mathematically tractable approach called Inhomogeneous Double Spatially-Thinned Poisson Point Process Modeling Approach. It can be applied to any spatial point process with repulsions & attractions. This approach is as simple to simulate as PPP-based approach. It has the same mathematical tractability and insightfulness as the PPP-based approach as well.

Keywords: non-PPP, spatially-correlated, performance evaluation

1 Introduction

In this paper, a detailed introduction on how to apply stochastic geometry for modeling, analyzing and optimizing 5G ultra dense cellular networks is provided. The ever-rising demand for wireless data implies that conventional cellular architectures based on large macro cells will soon be unable to support the anticipated density of high-data-rate users. The traditional approach of modeling macro cellular networks is not applicable anymore to ultra-dense network deployments. This is due to the large number of parameters and network configurations that need to be analyzed, which make simulation-based approaches too expensive and impractical.

The practical deployments of heterogeneous ultra-dense cellular networks are not totally random or regular. The widely-adapted approaches to model BSs are mostly based on PPP, which is not accurate for real BSs distribution. Base stations are deployed based on coverage, rate & data traffic criteria that make their locations spatially correlated. Currently available research works rely on two assumptions:

1) The base stations are always assumed to be randomly deployed (Poisson point process assumption), regardless of their type. There are plenty of literatures for PPP-based approach, e.g. [1][2][3][4]. However, random deployments are not accurate for macro base stations.

2) The base stations are modeled using some specific point processes (determinantal [5], Ginibre [6], etc.) that are mathematically tractable. However the mathematical frameworks obtained by using non-Poisson point processes are much less tractable

than their Poisson counterpart. And it does not provide any insight for system design. What's more, their computation may take longer time than optimized system-level simulations

The main contribution of this paper is: A new approach is proposed, which is called: Inhomogeneous Double Spatially-Thinned Poisson Point Process Modeling Approach, it models ultra-dense cellular networks with spatial attractions or repulsions; it is validated against real cellular network deployments; the new mathematical framework for system-level analysis is also developed.

2 Main Contribution

2.1 Sampling serving BS

The base stations are sampled according to a distance-depend function that accounts for the shortest distance properties of actual cellular network deployments and identify the serving base station.

The distance-based function used is called contact distance distribution. It can also be called empty space function, a spherical contact distribution function is defined as probability distribution of the radius of a sphere when it first encounters or makes contact with a point in a point process.

2.2 Sampling Interfering BSs

Another homogeneous Poisson point process is generated and sampled based on the location of the serving base station (previous subsection) and on the distance dependent properties of actual cellular network deployments. The resulting base stations constitute the interfering base stations

The obtained system model is a spatially-thinned version of the original Poisson point process that is mathematically tractable

3 Results

The following figure shows coverage probability Cauchy DPP (repulsive point process) using Double Thinning approach:

4 Conclusion

Our approach which is called Inhomogeneous Double Spatially-Thinned Poisson Point Process Modeling Approach, reproduces practical cellular networks generated by advanced statistical software (R). It is validated against real cellular network deployments as well as the new mathematical framework for system-level analysis. It is shown that the Poisson point process is less accurate. And the proposed proposed approach is obtained without loosing in tractability while it models ultra-dense cellular networks with spatial attractions or repulsions.

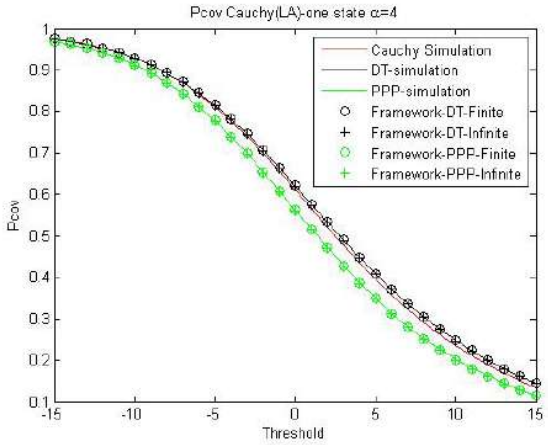


Fig. 1. Coverage Probability for Cauchy LA case with path loss exponent $\alpha = 4$

References

[1] M.DiRenzo,A.Guidotti,andG.E.Corazza,Average rate of downlink heterogeneous cellular networks over generalized fading channels A stochastic geometry approach, IEEE Trans. Commun., vol. 61, no. 7, pp. 30503071, July 2013.

[2] Haenggi, Martin, et al. "Stochastic geometry and random graphs for the analysis and design of wireless networks." IEEE Journal on Selected Areas in Communications 27.7 (2009).

[3] Dousse, Olivier, Franois Baccelli, and Patrick Thiran. "Impact of interferences on connectivity in ad hoc networks." IEEE/ACM Transactions on Networking (TON) 13.2 (2005): 425-436.

[4] Andrews, Jeffrey G., Francois Baccelli, and Radha Krishna Ganti. "A tractable approach to coverage and rate in cellular networks." IEEE Transactions on Communications 59.11 (2011): 3122-3134.

[5] Li, Yingzhe, et al. "Statistical modeling and probabilistic analysis of cellular networks with determinantal point processes." IEEE Transactions on Communications 63.9 (2015): 3405-3422.

[6] Deng, Na, Wuyang Zhou, and Martin Haenggi. "The Ginibre point process as a model for wireless networks with repulsion." IEEE Transactions on Wireless Communications 14.1 (2015): 107-121.

Forecasting bike sharing demand
[Poster demo]

Aurélie Fréchet

Ecole Nationale de la Statistique et de l'Analyse de l'Information
EDF R&D

Abstract. This paper introduces an application based on Open Data sources. The purpose of this application is to forecast in real-time the number of bikes available at each station of each target city. Three main points are considered. The first objective of the study is to collect information from Open Data and optimize an associate database for future requests. Then apply different methods used in electricity load forecast on these data and forecast attendance at stations that weren't in the training set. Finally, we want to create an interactive visualization tool to share the results.

Keywords: Bike-sharing system, count data, generalized additive model, Shiny, sequential aggregation

1 Motivation

This study was requested by the load forecasting research team at EDF R&D challenges in aim to predict bikes availability in a bike-share system (BSS). Data from BSS differ from electricity consumption by their format (counting data instead of continuous ones) but have a similar behavior. Both have peeks in the morning and in the evening and have a different behavior during the weekend. This similarities are explained by the relation with the human daily routine. This study has three main purposes:

- To improve the knowledge of open data environment and find the best processes in order to work with it. Many databases originating in various application domains such as energy systems or bike sharing systems, are now available through the development of data sharing.
- To apply different methods, that actually work for load electricity forecasting, on daily and hourly bike availability forecasting. Forecasting is an important task both for electrical and bike sharing systems. to know in advance the bike or electricity demand allows to optimize system management, to improve customer experience and ultimately to be more cost-efficient. For example, predict when a station will be empty in order to move bikes from full stations to this empty station.
- To visualize the results by an interactive application. In order to facilitate the understanding of the results, an interface will offer the user several menus to get the information he needs.

Fig. 1. Capture of the application

2 Open Data management

Working with large scale data implies management of a database. Our data come from two sources : from a closed Open Bikes challenge[1] for Paris, Toulouse and Lyon and from day-to-day upload from an API for Rennes[2].

SQL Database Creation

For the training data set, historical BSS data from April 2016 to October 2016, for Paris (1 152 bike stations), Lyon (348 bike stations) and Toulouse (277 bike stations) are used. The final database contain 23 868 434 rows. The stations information (number of bikes, number of space) were updated at each new arrival or departure. A pre-processing of the data was necessary to homogenize the time step to 10 min and thus make the data usable as time series.

The database is constructed using RSQLite [1] library, an R interface to SQLite. It contains four types of variables : geographical information (like altitude, latitude, longitude, district and city), time information (like day, month, year, holiday and day of the week), weather information (like temperature, humidity, pressure and wind) and station information (like number of bikes, of spaces, arrivals and departures).

API communication

BSS data for Rennes will be requested directly by an API called from R every ten minutes. These data will be stored in a SQL database the same way as the training set.

3 Modeling bike availability

The purpose of this part is to test different forecasting methods. Three different prediction methods are chosen: Generalized Additive Model [2], Regression tree [3] and Prophet [4]. Finally, the predictions obtained from these three methods will be combined using sequential aggregation of experts [5].

We want to be able to forecast bike availability for every station, regardless of whether it is present in the training set or not. This would allow providers to anticipate the future use of a new station.

Generalized Additive Model

The number of entries or departures depends linearly on unknown smooth functions of variables. These models have been successfully applied to related forecasting problems, especially in load forecasting.

Simple regression tree

Regression trees are a type of algorithm for predictive modeling. Each node of the tree corresponds to a decision (temperature superior or inferior 15 $^\circ$ C), and the leaf to the value of the prediction (if temperature superior 15 $^\circ$ C, 5 bikes). This approach is often very efficient and comprehensive.

Prophet

Prophet's package implements a procedure for forecasting time series data based on an additive model where non-linear trends are fitted with yearly and weekly seasonality plus holidays, so seems quite adapted to our task.

Mixture of experts

It is possible that no method will be better than the others for every station and all the forecasting horizons. That is why we will use aggregation algorithms to dynamically combine them and get the best from each method. This will be done with the opera library [6].

[1] https://maxhalford.github.io/blog/openbikes-challenge/
[2] https://data.explore.star.fr/explore/dataset/vls-stations-etat-tr/

4 Visualization interface

An RShiny [7] application is used to visualize the results on a map (*Figure 1*) thanks to leaflet package [8]. This application allows the user to dynamically consult the recent history and forecasts for each station.

The application shows a map of France with different icons for the cities in the scope analysis. When clicking on the icon, the data from the corresponding city are uploaded in the application. Then the district map of the city is shown to the user.

The user can click on a district to zoom in on this district, in order to have the plan of the district with icons for the bike stations. Clicking on a station will open a window that contains daily and hourly predictions. These options are also available in the right panel, where the user can select the city, the district and the station with select menus.

There is also a panel to change display. The user can choose which information to show on the map: the number of bikes available in the stations, the number of bikes that left the stations in the last 10 minutes, or the number of bikes that entered the station in the last 10 minutes. The icons presented on each station depend on the information asked by the user. Finally, the user can change the date/hour in the menu and be presented with the corresponding data.

References

1. Kirill Müller, Hadley Wickham, David A. James, and Seth Falcon. *RSQLite: 'SQLite' Interface for R*, 2017. R package version 1.1-2.
2. S.N Wood. *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC, 2006.
3. Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2015. R package version 4.1-10.
4. Sean Taylor and Ben Letham. *prophet: Automatic Forecasting Procedure*, 2017. R package version 0.1.1.
5. Pierre Gaillard and Yannig Goude. Prévision de la consommation électrique par mélange de prédicteurs: travail sur le jeu d'experts. 2014.
6. Pierre Gaillard and Yannig Goude. *opera: Online Prediction by Expert Aggregation*, 2016. R package version 1.0.
7. Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2017. R package version 1.0.1.
8. Joe Cheng, Bhaskar Karambelkar, and Yihui Xie. *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*, 2017. R package version 1.1.0.

COMMUNITY AND GRAPH

TALK SESSION

Community recovery in stochastic block models using semi-definite programming
[Talk submission]

Youssouf Emin, Ismael Lemhadri

Ecole polytechnique

Abstract. We analyze the community recovery problem in the context of stochastic block models through the lens of a special semi-definite program, the so-called PECOK algorithm. We show why this program interprets as a relaxed version of K-means, and we prove that it achieves exact recovery with high probability as soon as a simple condition on the 'within-between' covariance gap is satisfied. Furthermore, we show that this condition is weaker than that of several previous SDP formulations, at least in the special case of the planted partition model.

Keywords: community recovery, stochastic block models, semidefinite programming

1 Motivation/Introduction

Community recovery refers to the classical problem of estimating individuals' group memberships from the observation of the interactions within the network. This problem is of utmost importance for many modern data-driven challenges, with large amounts of network data now being available from fields as diverse as social networks [7], protein to protein interaction [6] and social science [8], among many others.
The stochastic block model [9] is a simple yet powerful and extremely popular model of network interactions, due to its analytical tractability and connections to fundamental graph properties. However, fitting the SBM is a challenge: the problem of optimizing label assignment over all possibles classes is NP-hard. One common approach relies on maximum likelihood estimation, which suffers great sensitivity to starting points. Alternative methods using spectral clustering [10] are efficient for networks with large blocks but fail for sparse networks.

2 A semi-definite approach to variable clustering

Our approach relies on the so-called PECOK algorithm (for PEnalized COnvex relaxation of Kmeans), first defined in [1] in the context of G-latent models, where it was introduced as a corrected, relaxed version of K-means. We show how this approach can extend to the case of stochastic block models. This leads to a semi-definite program whose solution is, with high probability, exactly the true configuration of SBM groups, as soon as a condition on the 'within-between' covariance gap is satisfied.

3 Optimality of the semidefinite program

We then compare our condition for perfect recovery to the SBM literature, in particular to [3,4,5]. We show that in the special case of the planted clique problem (which yields a very simple and clear condition) our SDP outperforms several other formulations and is information-theoretically optimal in the sense of the Chernoff-Hellinger divergence (defined in [4]) for several regimes of the parameters.

4 Conclusion and discussion

In this paper, we have shown how the PECOK algorithm can be successfully transposed to the context of stochastic block models, leading to exact recovery of the block partition. Our approach allowed us to build an efficient, highly tractable procedure that was proven effective in other probabilistic frameworks as well [1, 2]. An interesting area of investigation would be to complement this approach with numerical experiments. Another area of interest concerns the adaptive estimation of the number of groups when it is unknown, which we leave as a future endeavor.

References

[1] F. Bunea, C. Giraud, M. Royer and N. Verzelen. PECOK: a convex optimization approach to variable clustering. arXiv:1606.05100v1. 2016.

[2] M. Royer. Adaptive Clustering through Semidefinite Programming. arXiv:1705.06615v1. 2017.

[3] Y. Chen, J. Xu. Statistical-Computational Tradeoffs in Planted Problems and Submatrix Localization with a Growing Number of Clusters and Submatrices. Journal of Machine Learning Research 17. 2016.

[4] A. Perry, A.S. Wein. A semidefinite program for unbalanced multisection in the stochastic block model.arXiv:1507.05605. 2015.

[5] A.A. Amini, E. Levina. On semidefinite relaxations for the block model. arXiv:1406.5647v3. 2016.

[6] M.S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A.R. Pico, A. Vailaya, P. Wang, A. Adler, B.R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G.D. Bader Integration of biological networks and gene expression data using cytoscape. Nature. 2007.

[7] X. Zhang, C. Moore, M. E. J. Newman. Random graph models for dynamic networks. arXiv:1607.07570v1. 2016.

[8] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. Foundations and Trends in Machine Learning, Volume 2 Issue 2. 2009.

[9] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. Social networks 5.2 109-137. 1983.

[10] J. Lei, A. Rinaldo. Consistency of spectral clustering in stochastic block models. 2014.

Phase transitions in hierarchical community detection

Talk submission

Stefano Sarao[†], Thibault Lesieur[†], Lenka Zdeborová[†]

[†]Institut de Physique Thorique, CEA Saclay, France

Abstract. In this work we present analytic results concerning the possibility of capturing, using an algorithm, information in a generalised version of community detection. Community detection is a well known problem in machine learning with a variety of applications in the internet and in real word. The limit value of the signal to noise ratio after which detection occurs characterises in physics a phase transition. This was conjectured for standard community detection in 2011 [DKMZ] and then proved rigorously in 2013 [NMS]. In this paper we extend this analysis to a more general case where a number of nested community exists, this scenario includes also the standard case and we find agreement in the results. The problem was studied using tools from statistical physics (i.e. Approximate Message Passing) that provide on one hand theoretical insights on the problem, on the other hand an algorithm for solving it.

Keywords: Community Detection, Statistical Physics, Approximate Message Passing

1 Introduction

This study generalises the results obtained in [TKZ15,TKZ17] on the stochastic block model (SBM) to the hierarchical stochastic block model (HSBM) where the communities are nested. Most of the notation and the methodology adopted are the same as in [TKZ17]. In the HSBM the different levels of hierarchy range from 1 to G, where level 1 is the largest community made of all the individuals and level G is the level of the smallest communities. The probability of two individuals to be connected depends on the community of highest level where both belong, e.g. the usual SBM has $G = 2$ and the probabilities of connection are: p_1 if they are in the same community at level 1, p_2 if they are in the same community only at level 2. Our goal is to determine the phase transition in the thermodynamic limit where the number of individuals, N, tends to infinity. In our study we focus on the scenario where we have a symmetric structure, which means that for each level inside every community the probability of two members to be connected is the same and the number of sub-communities that they contain is the same. We want to remark that for an algorithm this represent the hardest possible scenario for detecting communities, where counting the degree of a node is not informative.

2 Methods

Our goal is to determine the threshold values of the detectability given the structure of the HSBM. In order to avoid a trivial problem in the thermodynamic limit, the probabilities are scaled in such a way that $|p_i - p_j| = O(1/\sqrt{N}) \ \ \forall \ i, j$. Our method consists in the study of the state evolution of the Approximate Message Passing (AMP) algorithm and studies the stability of the trivial solution, i.e. random guess. We are also interested in the different types of phase transition, first or second order, because the presence of a first order phase transition will suggest the presence of three critical points characterising: algorithmic spinodal, information theoretic and dynamic spinodal phase transitions. We studied also what happens to the performance of an algorithm if one level is particularly simple to detect. Finally, the performed simulations are in agreement to our theoretical analysis.

3 Results

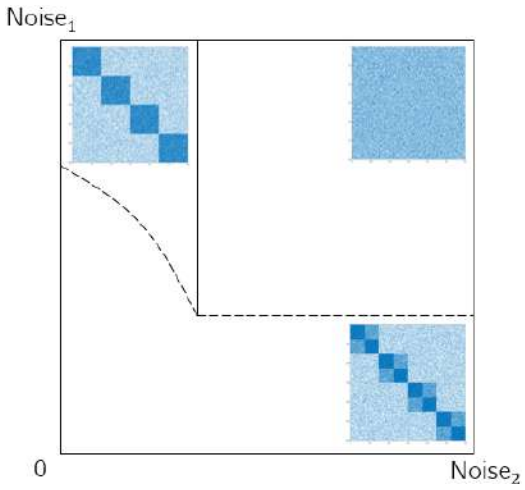


Fig. 1. Phase diagram in the case of three hierarchies, $G = 3$. On the two axis x and y we have respectively the noise level to distinguish between the second and the third level and noise level to distinguish between the first and the second. The diagram highlights three different areas: where both the non trivial level of communities cannot be detect, where only the largest communities can be detected and where the algorithm can reconstruct both the non trivial levels. We represent by a solid line a continuous transition and by a dashed line a discontinuous one. The small images sketch the best results achievable by the algorithm in that region.

In this study we strongly reduce the number of relevant parameter, from $O(2^G)$ to $G - 1$ that quantify the signal to noise ratio in the detection. The number of order

parameters has also been reduced to the only $G-1$ that play a role and quantify the amount of information detected on the communities at a certain level. We identify the threshold value of the detectability and we determine a general criterion to understand whether the transition will be first or second order. We proved that if a given level is easy to detect, the lower ones will be easy to detect as well and problems will be as complicated as determining the remaining sub-communities. This intuitively means that if a level is simple we can solve it and split the adjacency matrix, thus study the remaining communities separately.

Figure 1 represents a phase diagram with $G = 3$, one level more than the usual SBM, that is non trivial and yet visualisable in a plane. In the axis we observe an increasing noise in the detection of level 3 and 2 respectively. In agreement with our results, we observe that it doesn't exist a region where the finer level can be detected and the one in between cannot.

4 Conclusion

In this work we studied analytically the phase diagram of the HSBM and our results agree with the ones present in literature for the standard SBM. In the analysis we were able to evaluate the limiting behaviours of the parameters and identify the consequences on the performance of the algorithms. In the future we will study in detail the information theoretic phase transitions from the state evolution and the free energy.

References

[DKMZ] Decelle, Aurelien, et al. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications." Physical Review E 84.6 (2011): 066106.

[NMS] Mossel, Elchanan, Joe Neeman, and Allan Sly. "Stochastic block models and reconstruction." arXiv preprint arXiv:1202.1499 (2012).

[TKZ15] T. Lesieur, F. Krzakala, and L. Zdeborová. (2015). MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. arXiv preprint arXiv:1507.03857 (2015).

[TKZ17] Lesieur, Thibault, Florent Krzakala, and Lenka Zdeborová. "Constrained Low-rank Matrix Estimation: Phase Transitions, Approximate Message Passing and Applications." arXiv preprint arXiv:1701.00858 (2017).

End-to-end Causal Modelling

[Poster/Talk submission]

Diviyan Kalainathan, Olivier Goudet, Philippe Caillou,
Isabelle Guyon, Michèle Sébag, Paola Tubaro
email:{firstname.name}@lri.fr

TAU, LRI, CNRS, INRIA, Univ. Paris-Sud, Université Paris-Saclay

Abstract. This paper aims at causal modelling and proposes an end-to-end approach including: i) dependency estimation between pairs of variables; ii) (undirected) graph construction; iii) edge orientation, determining the sense of causal relations. The contributions involve an original dependency estimate robust to heterogeneous data, and a rigorous methodology to control the false discovery rate. The empirical validation presents a comparative assessment of state-of-art methods on each one of the above causal modelling steps.

Keywords: Causal inference, Graphical models, Probabilistic methods

Causal modelling, a research topic gaining visibility in the Machine Learning field, opens a wide range of applications in epidemiology, medicine, social sciences and economics. The gold standard methodology in causal modelling is to conduct randomized controlled experiments, subject to feasibility, ethics and cost limitations. Alternatives based on the exploitation of existing evidence are pioneered by [Pea03, SGS00, Chi02, TBA06, HJM+09, ZH09, Fon16, LPMST15] among others. This paper presents a benchmarking study, together with original contributions, to compare existing methods and guide practitioners to rigorously apply them depending on their requirements and conditions of use.

The proposed approach considers the typical setting of cross-sectional survey studies[1], with propositional data described using binary, categorical, and/or continuous features, with iid samples. In medical or socioeconomic studies a sample typically represents a patient, a citizen or a household. Features, also called variables in the following, are generally inter-dependent. The goal of causal modelling is to reveal the mechanism underlying such dependencies, which can expressed as a **functional causal model** (FCM, generalizing Structural Equation Models). An FCM models each variable as a function of (a few) other variables, augmented with an unknown "noise" variable. Formally, if X causes Y (noted $X \to Y$), then $Y = f(X, E)$ with E representing the noise variable.

In this paper, our goal is to construct an FCM capturing the most salient functional dependencies explaining data, aimed at supporting the choice of further experiments in order to assess the influence of given factors on given target variables. The motivating application regards the modelling of the causal relationships between the quality of life at work and the economic performance of

[1] Missing data and time dependencies are left for further study.

the enterprises, exploiting data from the Ministry of Labour.[2] The main criteria for this study include: i) the simplicity of the FCM; ii) the ability to control the rate of false positive (false discovery rate). The proposed methodology is a 3-step process, gradually refining the model to deal with uncertainty and provide practitioners with intermediate outcomes of interest (Figure 1):

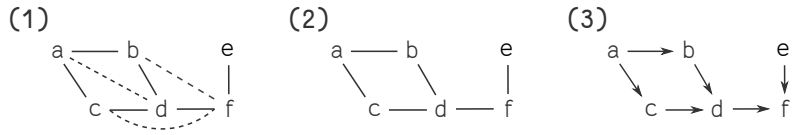


Fig. 1. 3-step causal graph modelling. Indirect dependencies (dashed lines in step 1) are pruned in step 2. Step 3 orients the edges.

Pairwise dependencies

The first step builds a graph skeleton based upon pairwise dependencies between variables (graph nodes), where two nodes are connected *iff* they are not independent. An original non-parametric test based on (binned) Mutual Information is proposed to handle the dependency estimation in an agnostic way, dealing with continuous and categorical variables. The precision/recall performance (and therefore the false discovery rate) is calibrated on the data from the Causal Dependency Challenge [Guy13]. Considering the continuous variable case, where the optimal parametric test is known (based on Pearson test), the performance loss is shown to be negligible for a number of samples greater than 500.

Recovering the Undirected Markovian Graph

As step one retains direct and indirect dependencies, it may result in a cluttered graph. Step two removes indirect dependencies to recover the Undirected Markovian Graph (UMG) via conditional independence tests: Variables get disconnected if they are mutually independent, conditionally to any subset of other variables. No attempt to orient the graph is made at this stage. This step can be very computationally expensive and/or lack of robustness, depending on the number of samples and the number of variables. State-of-art methods have been investigated and assessed. Eventually, a heuristic and computationally efficient method inspired from forward feature selection is retained, together with an embedded halting criterion based on probes (spurious variables), allowing to monitor and control the false discovery rate (number of missed indirect independencies). This approach has been comparatively assessed w.r.t. Bayesian methods, Kernel-based HSIC, Partial correlations and Feature selection methods.

[2] For more detail about the study and the code, the interested reader is referred to github.com/Diviyan-Kalainathan/causal-humans.

Orientation of the edges

Eventually the edges retained in step 2 are oriented to define the FCM. The proposed methodology is based on cause-effect pair algorithms, which overcome limitations of structural methods basing edge orientation structural constraints such as V-structures (unshielded colliders). The advantage of cause-effect pair algorithms is that they can orient graphs belonging to Markov equivalence classes that cannot be resolved with constraints derived from conditional independencies. Experimentally, we compared pairwise algorithms from two families (1) model-based methods (ANM [HJM+09] or PNL [ZH09]) and (2) learning-based approaches, including Jarfo [Fon16] and RCC [LPMST15], that performed well in the ChaLearn Cause-Effect Pairs Challenges [Guy13] and [Guy14].

Discussion

The proposed 3-step methodology has been compared and assessed w.r.t. state-of-art algorithms aimed at direct causal graph reconstruction, including *constraint-based methods* such as the PC algorithm [SGS00], which uses V-structures and constraint propagation to orient the graph, and *score-based methods*, such as GES [Chi02], which use a likelihood metric to improve the structure. Other methods like MMHC [TBA06] combine both approaches. Their robustness has been rigorously assessed with regard to their assumptions (including causal sufficiency and faithfulness).

The comparison establishes the merits of the proposed methodology, in terms of accuracy and friendliness for practitioners. In particular, this method: i) transparently handles numerical, categorical, and binary variables; ii) behaves well (accuracy and computation-wise) in the usual range (from a few hundred to some thousand samples); iii) provides a rigorous and intuitive way of controlling the false discovery rate; iv) avoids parametric modeling assumptions and is robust against violations of simplifying assumptions. Our extensive step-by-step assessment demonstrates its good performance.

Bibliography

[Chi02] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[Fon16] José AR Fonollosa. Conditional distribution variability measures for causality detection. 2016.

[Guy13] Isabelle Guyon. Chalearn cause effect pairs challenge, 2013.

[Guy14] Isabelle Guyon. Chalearn fast causation coefficient challenge. 2014.

[HJM+09] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

[LPMST15] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya O Tolstikhin. Towards a learning theory of cause-effect inference. In *ICML*, pages 1452–1461, 2015.

[Pea03] Judea Pearl. Causality: models, reasoning and inference. *Econometric Theory*, 19(675-685):46, 2003.

[SGS00] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[TBA06] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 2006.

[ZH09] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009.

Weak laws of large numbers for persistence diagrams

Vincent Divol
M2 student at Université Paris-Sud

August 30, 2017

Abstract

Topological Data Analysis (TDA) is an emerging field which aims at extracting useful geometric information from different types of data. Persistence diagrams are one the most used tool in TDA, and consist in a robust summary of the topological properties of point clouds. From a probabilistic point of view, surprisingly few results exist about their asymptotic behavior when the size of the data becomes great. In the framework where point clouds arise as i.i.d. observations, we are able to prove laws of large numbers for a large class of functionals on persistence diagrams. Applications include heuristics on tuning parameters for persistence intensity, a widely used representation of persistence diagrams.

Keywords— persistence diagrams, laws of large numbers, euclidean graphs, Čech complex

1 Framework

In many different situations, datasets are noisy approximations of some topological space $\mathbb{X}$. Even when reconstructing $\mathbb{X}$ is too hard (see [4] for minimax rates in manifold estimation), some meaningful topological properties about the space $\mathbb{X}$ can be efficiently estimated. One elementary topological property of a topological space $\mathbb{X}$ is its number of connected components, and clustering can be seen as the estimation of those features. A higher dimensional generalization of connected components is k-level homology $H_k(\mathbb{X})$, which is a vector space whose dimension, called a Betti number, intuitively counts the number of k-dimensional holes in a space (see [5] for a precise definition of homology).

The k-homology of a finite point cloud $\mathbb{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^d$ will always be trivial, and a good idea to have a grasp on the topology of the underlying space is to look at the ε-neighboorhood $\mathbb{X}_n^\varepsilon$ of the point cloud $\mathbb{X}_n$. Of course, the choice of ε is crucial in this approach. A solution to overcome this problem is to keep track of how the homology of $\mathbb{X}_n^\varepsilon$ changes as ε grows from 0 to $+\infty$, and more precisely at when different features (e.g. k-dimensional holes) born and die. The persistence diagram D_n of $\mathbb{X}_n$ is the collection of the couples $\mathbf{r} = (r_1, r_2) \in \mathbb{R}^2$ where r_1 is the radius at which a feature appears (its birth) and r_2 is its death, the radius at which it disappears. A rigorous definition of persistence diagrams in a more general setting is given in [3]. The persistence

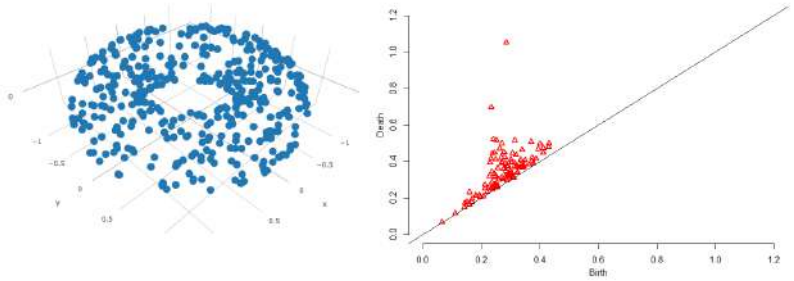


Figure 1: The persistence diagram for 1-homology built on 500 random uniform points on a torus. The two most persistence points on the diagrams correspond to the two holes of the torus.

of a feature $\mathbf{r}$ is defined as $pers(\mathbf{r}) = r_2 - r_1$ and measures intuitively the importance of said feature. For instance, in the persistence diagram of a point cloud sampled on a torus, one would see two much more persistence points in 1-homology, corresponding to the two actual holes of the torus (see figure 1).

2 Laws of large numbers

As seen on figure 1, there are two different kind of points in persistence diagram built on random approximations of manifold: few persistence features correspond to the persistence diagram of the actual manifold and a lot of points with low persistence correspond to noise in the data. We investigated the behavior of the second type of points in the asymptotic setting: When the size n of the data becomes great, how many are they? What is the order of the sum of the persistence of the points? etc. A diagram is a finite set of points, which can equivalently be seen as a measure on the set $\Delta = \{\mathbf{r} = (r_1, r_2) \in \mathbb{R}^2 \text{ s.t. } 0 \leq r_1 < r_2 \leq \infty\}$. Taking this point of view, the quantity $\sum_{\mathbf{r}\in D} f(\mathbf{r})$ is written $D(f)$. We now state two theorems which are representative of the kind of theorems which can be proven with our approach.

Theorem 1. *Let $\mathbb{X}_n = \{X_1, \dots, X_n\}$ be a n-sample of some continuous density κ on $[0,1]^d$ such that $\inf \kappa > 0$. Denote D_n the persistence diagram of $n^{1/d}\,\mathbb{X}_n$ for k-homology. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function with polynomial growth, such that it converges uniformly to 0 close to the diagonal $\{r_1 = r_2\}$. Then, for all $p \geq 1$.*

$$n^{-1}\left(D_n(f) - E[D_n(f)]\right) \xrightarrow[n\to\infty]{} 0 \text{ in } L_p. \tag{1}$$

Theorem 2. *Assume now that κ is the uniform density. Then, there exists a Radon measure ν on Δ such that*

$$n^{-1} D_n \xrightarrow[n\to\infty]{v} \nu, \tag{2}$$

where $\xrightarrow[n\to\infty]{v}$ denotes the vague convergence on Δ.

Note that this last theorem was already proven in the recent paper [2] in the Poisson setting where $\mathbb{X}_n$ is an homogeneous Poisson process of intensity n.

3 Persistence intensity

The difficulty to do statistical inference directly in the space of persistence diagrams motivated some work to propose more tractable representations of them. Some of those representations take the form of a kernel density estimator over the diagram. For $K : \Delta \to \mathbb{R}_+$ a positive kernel and $w : \mathbb{R}^2 \to \mathbb{R}$ a weight function, define the persistence intensity [1] as

$$\rho_D : z \in \mathbb{R}^2 \to D(wK(z - \square)) := \sum_{\mathbf{r} \in D} w(\mathbf{r}) K(z - \mathbf{r}).$$

The persistence surface being a real valued function, classical machine learning methods can be applied to them and for instance can be used to achieve classification.

As one expects a persistence diagram to have a few meaningful points with high persistence and a lot of points with small persistence, the role of the weight function is to suppress sufficiently those small persistence points. Adams et al. [1] advises to use a weight $w(\mathbf{r})$ equals to the persistence $pers(\mathbf{r})$ of $\mathbf{r}$. The weak law of large numbers we proved hints that a weight equal to $pers(\mathbf{r})^\alpha$ for $\alpha = d$ would be more efficient to suppress the noise. This choice is supported by numerical experiments on synthetic data.

References

[1] Henry Adams, Sofya Chepushtanova, Tegan Emerson, Eric M. Hanson, Michael Kirby, Francis C. Motta, Rachel Neville, Chris Peterson, Patrick D. Shipman, and Lori Ziegelmeier. Persistence images: An alternative persistent homology representation. *CoRR*, abs/1507.06217, 2015.

[2] Trinh Khanh Duy, Yasuaki Hiraoka, and Tomoyuki Shirai. Limit theorems for persistence diagrams. *arXiv preprint arXiv:1612.08371*, 2016.

[3] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[4] Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under hausdorff loss. *The Annals of Statistics*, pages 941–963, 2012.

[5] Allen Hatcher. *Algebraic topology*. Cambridge Univ. Press, Cambridge, 2000.

SPATIO-TEMPORAL DATA

TALK SESSION

Bandit algorithms for power consumption control

[Talk (or eventually Poster demo) submission]

Margaux Brégère

EDF R&D

Abstract. To influence power consumption, some of the electricity price's variations could be sent to customers. In order to optimize the choice of these tariffs, stochastic contextual bandit problems are considered. To measure the performance of the algorithms associated with such models, realistic datasets should be generated. Then, algorithms could be tested and extended to more complex cases.

Keywords: Bandit models, Sequential learning, Power consumption prediction.

1 Motivation

Forecasting and adjusting the production in accordance the daily consumption of its customers is a main concern for EDF. Conventionally, EDF adapts its means of production to this expected consumption. Recently, new communication tools (connected electricity meters, smart home sensors...) between the electricity supplier and its customers should allow to influence the users' consumption. This can be done either directly with a servo-control of various equipment (hot water tanks, pumps of swimming pools) or indirectly by sending incentive signals (off-peak hours, price variations). For instance, the consumption could be adjusted to the production of renewable energy, subject to climate variations. The transmission of these control signals has to be optimized: which customers should be targeted? When? For which weather conditions?

Clients' responses are stochastic and as consumption is only observed once the tariff is fixed, the mathematical framework to consider is the theory of multi-armed bandits. Indeed, in the simplest case, the choice of the electricity price is made at each step time depending on exogenous variables (temperature, date...). When the tariff is chosen, only the consumption associated with this tariff is monitored. (However, the consumption associated with another price is obviously unknown). Given that the price impacts the consumption, it should be picked so that this consumption fits with a target consumption based upon the electricity production. There is an exploration-exploitation trade-off. Actually, exploring multiple tariffs allows to deduce good estimations of their effects while exploiting the ones that seem the bests should reinforce the match between the effective consumption and its target. Although various algorithms that perform well in theory exist (UCB, Exp3), they would require some modifications to be used in this context. Since a unique realization is available at each step time (partial information context), algorithms can't be experimented on real data. Therefore, a generator able to compute realistic datasets is essential.

2 Creation of a realistic data generator

The data generator will be based on the public dataset "SmartMeter Energy Consumption Data in London Households " published by UK Power Networks. The consumption of 5,567 London Households has been read at half hourly intervals throughout the 2013 calendar year period. Households have been allocated to a CACI Acorn group (categorization of the United Kingdom's population into demographic types) and were recruited as a balanced sample representative of the Greater London population. Within the data set are two groups of clients. In the first one, custumers were subjected to Dynamic Time of Use energy prices: tariffs (low, high or normal) were given a day ahead. All non-Time of Use customers were on a flat rate tariff. A descriptive analysis of this dataset points out the importance of tariff choices. The focus is firstly made on the simplest case : only two groups of clients are considered and the average consumption of these two groups is observed. The consumption of an unique household is too random to possibly observe the influence of tariffs' variations.

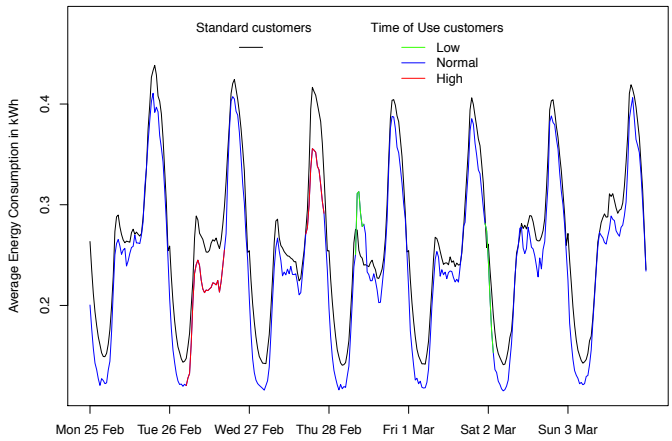


Fig. 1. Average consumption per household and per half hour for both sub-groups of custumers. For the Time of Use customers, colors' variations correspond to the tariffs' changes.

The classical path of the daily consumption is observed each day of the week: the consumption is very low during the night and presents two pics at 12 am and 8 pm. When the tariff is changed, the consumption of Time of Use costumers is influenced: a low tariff provides a raise of the consumption and reversely, a high price allows the consumption decreasing. But these changes are highly dependent on the hour: a variation of tariff during the night has very little consequences. This dataset will be

used to create the generator. Different predictive models (random forests and machine learning methods [3], generalized additive model (GAM) [4], aggregation of specialized experts [2]) will be then tested in order to generate the most realistic datasets possible. A very basic GAM model could be :

$$Y(D,h,T,P) = f_1(D,h) + f_2(T) + f_3(P,h) + \epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0,\sigma), \tag{1}$$

where Y is the consumption which depends on the day of the week D, the hour h, the temperature T and the tariff P and f_1, f_2 and f_3 are functions with a specified parametric form (polynomial, spline).

3 Bandit algorithms for power consumption control

Once this generator has been created, algorithms for contextual stochastic bandit problems should be adapted [1]. As some exogenous variables should be, such as temperature and date, included, the bandit problems to consider are indeed contextual. Each arm of the bandit model is associated to a tariff and the objective is to minimize at each step time t, given a context s_t, a distance $d(C_t, Y_{I_t,t})$ between the observed consumption $Y_{I_t,t}$, once the arm I_t has been chosen, and a target C_t. With $\mathcal{K}$ the set of arms, $\mathcal{S}$ the context set and denoting $g : \mathcal{S} \to \mathcal{K}$ a mapping of contexts to arms, to ensure that the algorithm performs well, the quantity to minimize is the regret :

$$R_T = \sum_{t=1}^{T} d(C_t, Y_{I_t,t}) - \min_{g:\mathcal{S}\to\mathcal{K}} \sum_{t=1}^{T} d(C_t, y_{g(s_t),t}), \tag{2}$$

where $y_{i,t}$ denotes the consumption that would have been observed if the arm i had been picked.

4 Clustering of clients to extend results

An appropriate clustering of the costumers thanks to some relevant features (average consumption at each hour of the day, average consumption per month, etc.) should allow to consider smaller group of clients. Applying a bandit algorithm on each cluster will hopefully provide better results as clients react differently to price variations.

References

1. S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.
2. M. Devaine, P. Gaillard, Y. Goude, and G. Stoltz. Forecasting the electricity consumption by aggregation of specialized experts; application to Slovakian and French country-wide (half-)hourly predictions. *Machine Learning*, 90(2):231–260, 2013.
3. R. Nedellec, J. Cugliari, and Y. Goude. GEFCom2012: Electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting*, 30:375–381, 2014.
4. V. Thouvenot, A. Pichavant, Y. Goude, A. Antoniadis, and J.M. Poggi. Electricity forecasting using multi-stage estimators of nonlinear additive models. *IEEE Transactions on Power Systems*, 31:3665–3673, 2016.

On Modeling and Analyzing Museum Visitor Movements with Semantic Trajectories

[Talk submission]

Alexandros Kontarinis[1,2]

[1] ETIS UMR 8051, Université Paris Seine, Université de Cergy-Pontoise, ENSEA, CNRS, F-95000, Cergy, France
[2] DAVID lab., Université Paris-Saclay, Université de Versailles Saint-Quentin-en-Yvelines, F-78035, Versailles, France
alexandros.kontarinis@ensea.fr

Abstract. Traditional museum audio guides have gradually been enhanced with multiple additional functionalities. The implementation of location-based services in particular, has enabled the collection of large volumes of spatiotemporal visitor movement data, from which individual visitor trajectories can be extracted. These trajectories can be studied to enable museums to better "know" their visitors, by means of trajectory mining. In addition, tapping into context data could make such analyses more expressive and revealing, by adding meaning and intention to trajectories. This work discusses important challenges in studying museum visitor movements with the help of context-aware indoor trajectory modeling and mining, and proposes work directions for each challenge.

Keywords: Data Mining · Indoor Trajectories · Trajectory Mining · Movement Patterns · Museum Experience

1 Motivation

Thanks to the advent of wireless technologies such as Bluetooth and WiFi, the location of museum visitors can be automatically acquired as they move through the exhibition spaces. The goal is to support location-based services offered via the museum's multimedia guides or mobile applications (e.g. automated audio content delivery). These services are ultimately aimed at improving the visiting experience. However, by collecting visitor movement data, museums can also study their visitors' mobility behavior, in an effort to gain a deeper understanding of their needs and expectations.

Useful actionable insight can be more readily extracted from movement data if those are first structured as individual trajectories, which may require some pre-processing steps or even be impossible, depending on the granularity and structure of the raw spatiotemporal data. A trajectory refers to the geometric aspect of the spatiotemporal path of a moving object. In simpler terms, it is a sequence of the object's positions in space and time: $T = (p_1, p_2, ..., p_n)$. Trajectory data-based applications are usually not primarily concerned with the physics of trajectories nor with the geometry of the moving objects. Indeed, museum visitors can be effectively modeled as moving points.

However, trajectories do not only have a geometric aspect. [5] considered trajectories from a semantic point of view, claiming that they should correspond to semantically meaningful travels. Hence, an object's whole movement lifespan consists of potentially many trajectories, each one meaningfully interpreted and defined by its own starting and ending time instants. A semantic segmentation of trajectories into application-specific sub-intervals of moves and stops was also introduced in [5]. But actually, trajectories can be semantically enhanced even beyond the "stops-moves" concept; any type of annotation which adds meaning to trajectories and their subdivisions, gives rise to so-called "semantic trajectories". Thus, a semantic trajectory can be seen as a sequence of spatiotemporal points complemented with annotations containing semantic values related to places, activities, transportation modes, or any other domain or environmental knowledge: $ST = ((p_1, \mathcal{A}_1), (p_2, \mathcal{A}_2), ..., (p_n, \mathcal{A}_n))$.

2 Identifying the Challenges: Current Approaches and Limitations in Trajectory Modeling and Mining

A major challenge in building an indoor visitor trajectory analytics system is designing a formal trajectory data model, which accounts for the complexities of indoor environments. For example, architectural elements can considerably affect movement. In museums, long and narrow hallways often dictate the path taken by the visitors, similar to how transportation networks restrict vehicle movement in outdoor settings. Also, accessibility constraints are far more dynamic (e.g. rooms closed for restoration purposes). Current trajectory models for the most part ignore such intricacies of indoor environments and need to be extended to indoor settings in non-trivial ways. This is partly due to the fact that existing navigation-oriented indoor space modeling standards have so far seen limited application [2]. Equally importantly, the trajectory data model should also account for varying degrees of data quality, because unlike outdoor trajectories which are based on GPS data, indoor trajectories are obtained through a wide variety of positioning technologies and techniques [4]. This leads to a whole range of location perceptions, each of different precision and quality.

Choosing the space and the trajectory model has a direct effect on the types of analysis that are implementable on top of them. For example, indoor distances can not be calculated using the typical 2D euclidean metric, primarily due to walls and multiple floors (vertical movement is also more frequent). This has many implications on the trajectory mining methods, such as rendering conventional trajectory similarity measures ineffective. Similarly, semantic trajectories also raise requirements in the types of trajectory distance measurement (e.g. two visitor trajectories which do not share any spatial characteristics, could still be considered semantically similar if they both start with stops on ancient Greek statues, then Italian paintings, and then finish at an exit).

Finally, semantic trajectories are so far mostly defined at a conceptual level and therefore most existing trajectory data mining methods and techniques deal exclusively with the spatiotemporal dimensions of trajectories. These methods include clustering, classification and mobility pattern extraction (frequent patterns, sequential patterns, association rules, group movement identification), etc., but with few exceptions, they ignore the semantic aspect of trajectories. As a matter of fact, there is still no clear consensus on which types of trajectory semantics would better describe human mobility behavior in general, let alone domain-specific or application-specific behavior.

3 Research Directions

One approach at accounting for the specificities of indoor environments is to represent indoor spaces by graph-based or set-based models consisting of symbolic locations (e.g.

human-readable hall identifiers). Ideally, an indoor trajectory model should combine both symbolic and coordinate trajectory representations. Also, to deal with positioning data quality issues, it should incorporate hierarchical elements that enable the representation of trajectories at multiple levels (e.g. as sequences of exact points, sequences of regions, sequences of rooms). More generally, a separation between a trajectory's abstract perception and its physical encoding is needed. Abstractly, a trajectory can be viewed as a continuous mapping function, defining a position in an indoor space for a visitor and time instant, while physically it can be described by a sequence of discrete predefined spatial cells (in the spirit of how the IndoorGML standard [2] represents indoor space) and temporal intervals, along with movement attributes (e.g. speed).

There exist few algorithms and data structures to support the process of semantically enriching trajectories, and therefore expressiveness and consistency issues in modeling semantic trajectories merit further investigation. Semantic analysis at an arbitrary number of different levels of detail (e.g. stops at different collections of a big museum each consisting of stops and moves at the room level and in turn at the exhibit level) is achievable by exploring enrichment processes based on the hierarchic subdivisions of movement (such as in [1]). Methods for direct domain-specific semantic enrichment of trajectories should also be investigated, as existing works have neglected domain specific datasets (in favor of DBpedia, Open Street Map, Open Weather Map and other popular sources) [3].

With regards to trajectory mining, one direction is to try to extend the existing methods and movement pattern definitions to indoor spaces, or alternatively investigate new movement patterns by considering the context as well, proposing new features of typical context-aware indoor trajectories, and analyzing them to capture visitor behavior and intention. Finally, on-line trajectory mining methods should be explored more, given that most works opt for historical data rather than trajectory data streams.

4 Conclusion

Museums are starting to consider the use of computational data analytics to study the moving patterns of their visitors. In this work, we identify the most important challenges to be resolved, in order to enable an advanced type of museum visitor movement analytics, and provide some general research directions for overcoming them.

References

1. Renato Fileto, Alessandra Raffaet, Alessandro Roncato, Juarez A.P. Sacenti, Cleto May, and Douglas Klein. A Semantic Model for Movement Data Warehouses. In *Proceedings of the 17th International Workshop on Data Warehousing and OLAP*, DOLAP '14, pages 47–56. ACM, 2014.
2. Hae-Kyong Kang and Ki-Joune Li. A Standard Indoor Spatial Data ModelOGC IndoorGML and Implementation Approaches. *ISPRS International Journal of Geo-Information*, 6(4):116, 2017.
3. Luiz Andr P. Paes Leme, Chiara Renso, Bernardo P. Nunes, Giseli Rabello Lopes, Marco A. Casanova, and Vnia P. Vidal. Searching for Data Sources for the Semantic Enrichment of Trajectories. In *Web Information Systems Engineering WISE 2016*, pages 238–246. Springer, Cham, November 2016.
4. Rainer Mautz. *Indoor Positioning Technologies*. PhD thesis, ETH Zurich, 2012.
5. Stefano Spaccapietra, Christine Parent, Maria Luisa Damiani, Jose Antonio de Macedo, Fabio Porto, and Christelle Vangenot. A Conceptual View on Trajectories. *Data Knowl. Eng.*, 65(1):126–146, 2008.

Towards a Scalable Learning and Classification of Multivariate Time Series

Talk submission

Ali Mzahem, Yehia Taher, Karine Zeitouni

University of Versailles UVSQ

Abstract. The rapid development of Internet-of-things (IoT) paradigm has significantly contributed to the raise of a more connected world where data is being generated at high rates; known as data streams or time series data, where traditional processing systems cannot handle them. Predictive analysis, which is gaining momentum nowadays from both research and industrial perspectives, provides a means for prediction in near real time by mining data streams. In this research work we aim at providing a novel platform for learning a scalable learning and classification solution that can deal with highly dimensional time series data.

Keywords: Predictive Analysis, IoT, Scalability, Time Series Data

1 Context and Motivation

Rapid evolution in communication technologies have been resulting with what is known as Internet of Things (IoT). IoT devices, ranging from devices, sensors to machines are now the cornerstone of a smart connected world. A word that is capable of having the required knowledge and information to evolve and learn making use of historical events. These connected things, mainly sensors that can be implanted on everything nowadays, produce what is called data streams or more formally time series data which are time stamped records. This data if exploited efficiently can result with greater erudition and material valuable for improvement or avoidance of possible threats and catastrophic situations. In other words, IoT devices have made proactivity a possibility. For instance, using previous data logs, one can perform analysis to formulate system models that are capable of instantaneous detections, predict events and react upon them in near real time. For example, deploying some sensors to observe the trucking of a historical piece of art from a museum to another will provide us with the necessary data readings, which will be analyzed as it pours into the system to help detecting if there is any violation that will take place before its actual occurrence. Thus, countermeasures can be done to avoid critical situations that cannot be tolerated. Such an example shows the power of proactivity in real world applications.

Many research initiatives have been focusing on the problem of devising predictive models. Some have modeled the sensor data as Time Series and applied some early classification over time series techniques to achieve prediction. These research initiatives have tried to achieve a rule based prediction model, which works by extracting some specific patterns, known as shapelets. Shapelets trigger prediction on their occurrence

in the time series being analyzed on real-time. Other research initiatives used complex event (CEP) technique to achieve proactivity. CEP systems are capable of reacting instantaneously upon occurring events. But, their efforts to apply predictive analytics ended up with event detection instead. For CEP systems to keep up with emerging science of data where predictive analytics is a must, the rules must be written in an automatic manner; they must be learned. Raef et al. [1] have managed to bridge the gap between data mining techniques and complex event processing, and came up with an innovative system, namely autoCEP, that learns rules from historical time series data and deploy them in a CEP engine to carry on with the classification of the new coming data. This work dealt with univariate time series and pushed further to cover multivariate time series by devising two algorithms that perform shapelet extraction, learn the rules, and achieve classification after deploying the rules in the CEP engine in an automated manner. But this approach was effective only to some extent. The learning phase starts to take more relatively a very long time as the dimension number increases, since each dimension is being dealt with alone. The shapelet extraction time; also, started to increase at an exponential rate with the increase in the dimensions of the time series. Also, the complex correlation of extracted shapelets takes a lot of time since it searches for the different permutations among the different dimensions. Moreover, it is not capable of handling Big Data streams.

2 Proposed Contribution

In order to keep up with the Big Data era, the algorithms proposed in [1] need to be re-designed and enhanced to be able to build the required rules from the extracted shapelets out of high dimensional time series in a more efficient manner. So,Particularly, our aim is to design and implement a distributed algorithm for multivariate time series classification. The algorithm must be scalable to handle the huge volume of the data and its high velocity. Rather, it must be capable of extracting the shapelets in a tolerable amount of time. Big Data tools are found to handle the four challenges of the current data being streamed all over the globe: Volume, Velocity, Variety, and Veracity. Using big data technology can help us achieve scalable architecture that is capable of handling high velocity and voluminous amount of data streams. Implementing the algorithm in a distributed manner, or using Big data technologies to implement the algorithm as Storm[1] and Kafka[2] will help us to achieve our goal of scalability resulting with a firm and strong predictive system that can keep up with data explosion. Devising such an algorithm, will result with a highly capable autoCEP engine that predict upon situations in near real time while maintaining the accuracy and the earliness of the classification of multivariate time series.

References

[1] Raef Mousheimish, Yehia Taher, and Karine Zeitouni. 2017. Automatic Learning of Predictive CEP Rules: Bridging the Gap between Data Mining and Complex Event Processing. In Proceedings of ACM DEBS, Spain, June 2017 (DEBS17), 13 pages.

[1] http://storm.apache.org
[2] https://kafka.apache.org

OPTIMIZATION

TALK SESSION

From safe screening rules to working sets for faster Lasso-type solvers

Mathurin Massias*, Alexandre Gramfort*, and Joseph Salmon†

*INRIA Saclay, Université Paris-Saclay, 91220 Palaiseau, France
first.last@inria.fr
†LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013 Paris, France,
first.last@telecom-paristech.fr

Abstract. In this short paper we present A5G [Massias et al., 2017], an efficient working set (WS) algorithm based on an agressive use of Gap Safe screening rules [Fercoq et al., 2015] for solving Lasso-type problems such as the multitask group Lasso.

Keywords: Machine Learning, Convex optimization, Sparsity, High dimension

1 Introduction

Convex sparsity-promoting regularizations are ubiquitous in modern statistical learning. By construction, they yield solutions with few non-zero coefficients, which correspond to saturated constraints in the dual optimization formulation. Working set (WS) strategies are generic optimization techniques that consist in solving simpler problems that only consider a subset of constraints, whose indices form the WS. Working set methods therefore involve two nested iterations: the outer loop corresponds to the definition of the WS and the inner loop calls a solver for the subproblems. For the Lasso estimator a WS is a set of features, while for a Group Lasso it refers to a set of groups. In practice, WS are generally small in this context so the associated feature Gram matrix can fit in memory. Here we show that the Gauss-Southwell rule (a greedy strategy for block coordinate descent techniques [Southwell, 1941]) leads to fast solvers in this case. Combined with a working set strategy based on an aggressive use of so-called Gap Safe screening rules, we propose A5G, a solver achieving state-of-the-art performance on sparse learning problems. Results are presented on Lasso and multi-task Lasso estimators.

2 Working set construction

The generic estimator we consider is defined as a solution of:

$$\hat{\mathrm{B}}^{(\lambda)} \in \underset{\mathrm{B}\in\mathbb{R}^{p\times q}}{\arg\min} \underbrace{\tfrac{1}{2}\left\|Y - X\mathrm{B}\right\|_F^2 + \lambda\Omega(\mathrm{B})}_{\mathcal{P}^{(\lambda)}(\mathrm{B})} , \tag{1}$$

where Ω is a sparsity inducing norm and the non-negative λ is the regularization parameter controlling the trade-off between data fitting and regularization. The associated dual problem reads (see for instance Ndiaye et al. [2015])

$$\hat{\Theta}^{(\lambda)} = \underset{\Theta \in \Delta_X}{\arg\max} \underbrace{\tfrac{1}{2} \|Y\|_F^2 - \tfrac{\lambda^2}{2} \left\|\Theta - \tfrac{Y}{\lambda}\right\|_F^2}_{\mathcal{D}^{(\lambda)}(\Theta)} , \tag{2}$$

where $\Delta_X = \{\Theta \in \mathbb{R}^{n \times q} : \Omega_*(X^\top \Theta) \leq 1\}$. For the multitask Lasso, we have $\Omega = \|\cdot\|_{2,1}$ and $\Omega_* = \|\cdot\|_{2,\infty}$.
Following the seminal work of El Ghaoui et al. [2012], to exploit the sparsity of $\mathrm{B}^{(\lambda)}$, for any dual feasible point Θ, current iterate B and duality gap $\mathcal{G}^{(\lambda)}(\mathrm{B}, \Theta)$, Gap Safe screening rules [Fercoq et al., 2015] safely discard predictor j from Problem (1) as soon as

$$\left\|X_{:,j}^\top \Theta\right\| + \|X_{:,j}\| \sqrt{\frac{2}{\lambda^2} \mathcal{G}^{(\lambda)}(\mathrm{B}, \Theta)} < 1 , \tag{3}$$

or equivalently only consider features such that

$$d_j(\Theta) := \frac{1 - \left\|X_{:,j}^\top \Theta\right\|}{\|X_{:,j}\|} \leq \sqrt{\frac{2}{\lambda^2} \mathcal{G}^{(\lambda)}(\mathrm{B}, \Theta)} . \tag{4}$$

As summarized in Algorithm 1, we propose to solve a sequence of subproblems, periodically considering only features such that

$$d_j(\Theta) \leq r \sqrt{\frac{2}{\lambda^2} \mathcal{G}^{(\lambda)}(\mathrm{B}, \Theta)} \quad \text{with} \quad r \in]0, 1[. \tag{5}$$

When restricting Problem (1) to only a small number of features, it becomes beneficial to use greedy block-coordinate descent solvers [Friedman et al., 2007, Nutini et al., 2015] combined with precomputation of the Gram matrix of the features.

3 Greedy BCD subproblem solvers

We denote $f(\cdot) = \|Y - X\cdot\|_F^2 / 2$ and B_j for the j-th row of B. When considering a block coordinate descent algorithm, one sequentially updates at step k, a single block (here row) j_k of B. For our problem, the block update rule proceeds as follows:

$$\mathrm{B}_{j_k}^k = \mathcal{T}_{j_k, L_{j_k}}(\mathrm{B}^{k-1}) , \tag{6}$$

where for all $j \in [p]$ the partial gradient over the j^{th} block $\nabla_j f$ is assumed to be L_j-Lipschitz (where $L_j = \|X_{:,j}\|^2$ is a possible choice),

$$\mathcal{T}_{j,L}(\mathrm{B}) := \mathrm{prox}_{\frac{\lambda}{L}\|\cdot\|}\left(\mathrm{B}_j - \frac{1}{L} \nabla_j f(\mathrm{B})\right) . \tag{7}$$

Algorithm 1 A5G

input : X, Y, λ
param: $p_0 = 100, \xi_0 = Y/\lambda, \Theta_0 = 0_{n,q}, \mathrm{B}_0 = 0_{p,q}$,
$\overline{\epsilon} = 10^{-6}, \underline{\epsilon} = 0.3$
for $t = 1, \dots, T$ **do**
  $\alpha_t = \max\{\alpha \in [0,1] : (1-\alpha)\Theta_{t-1} + \alpha\xi_{t-1} \in \Delta_X\}$
  $\Theta_t = (1-\alpha_t)\Theta_{t-1} + \alpha_t\xi_{t-1}$
  $g_t = \mathcal{G}^{(X,\lambda)}(\mathrm{B}_{t-1}, \Theta_t)$ `// global gap`
  if $g_t \leq \overline{\epsilon}$ **then**
    Break
  for $j = 1, \dots, p$ **do**
    Compute $d_j^t = (1 - \|X_{:,j}^\top \Theta_t\|)/\|X_{:,j}\|$
    `// safe screening:`
    Remove j^{th} column of X if $d_j^t > \sqrt{2g_t/\lambda^2}$
  Set $(d^t)_{\mathcal{S}^r_{\mathrm{B}_{t-1}}} = -1$ `// keep active features`
  $p_t = \max(p_0, \min(2\,\|\mathrm{B}_{t-1}\|_{2,0}, p))$ `// clipping`
  $\mathcal{W}_t = \{j \in [p] : d_j^t$ among p_t smallest values of $d^t\}$ `// Approximately solve sub-problem :`
  Get $\tilde{\mathrm{B}}_t, \xi_t \in \mathbb{R}^{p_t \times q} \times \Delta_{X_{:,\mathcal{W}_t}}$ s.t. $\mathcal{G}^{(X_{:,\mathcal{W}_t},\lambda)}(\tilde{\mathrm{B}}_t, \xi_t) \leq \underline{\epsilon} g_t$
  Set $\mathrm{B}_t \in \mathbb{R}^{p \times q}$ s.t.$(\mathrm{B}_t)_{\mathcal{W}_t,:} = \tilde{\mathrm{B}}_t$ and $(\mathrm{B}_t)_{\bar{\mathcal{W}}_t,:} = 0$.
return B_t

When precomputing $H = X^\top X$, it becomes beneficial to implement the GS-r rule [Nutini et al., 2015]:

$$\text{Pick} \quad j_k \in \arg\max_{j \in [p]} \max_{\mathrm{B}_j \in \mathbb{R}^q} \left\| \mathcal{T}_{j,L_j}(\mathrm{B}^{k-1}) - \mathrm{B}_j^{k-1} \right\| \ . \tag{8}$$

This rule makes large updates of B and make the bjective decrease quicker than cyclic selection rule. Usually it is more costly to compute, which is no longer the case when applied to small subproblems with Gram matrix computed. this makes greedy rules competitive in terms of time and not in terms of epochs as previsously studied in the literature.

Since it focuses on important features according to Gap Safe screening criterion, and because it uses a greedy solver for the subproblems, A5G (for AGGressive Gap, Greedy with Gram) reaches performance comparable to current state of the art [Johnson and Guestrin, 2015] for the Lasso and multitask Lasso.

4 Discussion

In this paper we have proposed a connection between Gap Safe screening rules and working set (WS) strategies, in particular to tackle more generic learning problems under sparsity assumptions such as the multi-task regression with $\ell_{2,1}$ regularization. We have illustrated the benefit of precomputing the Gram matrix for solving the subproblems. Precomputations allow Gauss-Southwell (GS)

variants to reach comparable performance to cyclic updates, not only in terms of epochs but also in terms of computing time. To our knowledge, our implementation is the first to demonstrate timing performance for GS rules. Last but not least the conjunction of WS strategies with GS methods reached noticeable speed-ups with respect to standard open source implementation available.

Bibliography

L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *J. Pacific Optim.*, 8(4):667–698, 2012.

O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the lasso. In *ICML*, pages 333–342, 2015.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

T. B. Johnson and C. Guestrin. Blitz: A principled meta-algorithm for scaling sparse optimization. In *ICML*, pages 1171–1179, 2015.

M. Massias, A. Gramfort, and J. Salmon. From safe screening rules to working sets for faster Lasso-type solvers. *ArXiv e-prints*, March 2017.

E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *NIPS*, pages 811–819, 2015.

J. Nutini, M. W. Schmidt, I. H. Laradji, M. P. Friedlander, and H. A. Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *ICML*, pages 1632–1641, 2015.

R. V. Southwell. Relaxation methods in engineering science - a treatise on approximate computation. *The Mathematical Gazette*, 25(265):180–182, 1941.

Bi-Objective Integer Programming For RNA Secondary Structure Prediction With Pseudoknots

[Talk submission]

Audrey Legendre, Eric Angel, Fariza Tahi

IBISC, Univ Evry, Université Paris-Saclay, 91025, Evry, France

Abstract. RNA structure prediction is an important field in Bioinformatics, and numerous methods and tools have been proposed. Pseudoknots are specific motifs of RNA secondary structures that are difficult to predict. We develop a bi-objective integer programming based algorithm, called BiokoP, allowing to combine two models and to return optimal and sub-optimal solutions for predicting RNA secondary structures with pseudoknots. BiokoP is compared with other existing tools for RNA pseudoknotted secondary structure prediction proposing several solutions. Considering 161 pseudoknotted secondary structures, BiokoP gives better results. BiokoP is available on EvryRNA platform (https://evryrna.ibisc.univ-evry.fr).

Keywords: RNA, Secondary structure, Pseudoknot, Integer programming, Bi-objective

1 Introduction

RNAs have major roles in the life cycle of the cell, in particular in the transcription and in the translation. Determining the structure of an RNA is an important step in the study of its biological function. Pseudoknots are specific motifs of the secondary structure which are essential in the understanding of the function of RNAs. In this work, we are interested in the prediction of the secondary structure of RNAs with pseudoknots. Two main approaches exist for predicting RNA structures. The thermodynamic approach consists in either to compute the structure of minimum free energy, or to compute the structure of maximum expected accuracy. The comparative approach consists of finding a conserved structure between several species and needs therefore at input several sequences. However, a single model can only approach the real structures. Moreover, it is now established that the real structure has indeed a very low energy, but not necessarily the minimum one. It is therefore interesting to propose approaches able to combine different models and able to return several solutions. To our knowledge, combination of different models for pseudoknotted secondary structure was only used between the comparative and thermodynamic models, meaning several sequences as input are needed. Moreover, only the tools proposed in [4], [2] and [3] can return several solutions.

In this work we propose a bi-objective integer programming based method to combine two models and to return optimal and sub-optimal solutions. Integer programming (IP) is very flexible, and allow to model diverse problems. Our method allows us to combine two thermodynamic models into a single bi-objective integer program (BOIP), from which we can get the optimal and sub-optimal secondary structures having the best tradeoff between the two criteria. The resulting tool, BiokoP (Bi-objective programming pseudoknot Prediction), is available on the EvryRNA platform: https://evryrna.ibisc.univ-evry.fr.

2 Methods

Our work is based on IP [8]. IP consists of optimizing an objective function according to linear constraints over a set of integer decision variables, to obtain an optimal solution. We are interested in optimizing several objective functions, corresponding here to different models for pseudoknotted RNA secondary structure prediction. We thus developed a BOIP, and an algorithm to find the set of optimal solutions, called the *Pareto set*, and sub-optimal solutions to which we will refer as the k-best Pareto sets.

In our BOIP, we combine two different thermodynamic models, one inspired by [6] and the other by [5]. The decision variables are binary and indicate the base pairings of the RNA sequence and the stack of two base pairs. To take into account the pseudoknots, it is assumed that a secondary structure can be decomposed into pseudoknot-free substructures. The constraints define basic rules like the impossibility for a base to be paired with several bases or forbid isolated base pairs. We propose a new generic algorithm to compute the exact k-best Pareto sets for any BOIP.

3 Results

We evaluated BiokoP on a dataset we constructed from the PseudoBase++ database [7]. It gathers 161 sequences of non redundant pseudoknotted sequences of RNAs.

We studied the distribution of the real structures (referenced in the literature) returned by BiokoP over the k-best Pareto set. 79 real structures are found, and the main part (45) belongs to the first Pareto set. The remaining are distributed on the three first sub-optimal Pareto sets. It shows the importance of considering sub-optimal solutions.

We compared BiokoP to two methods of the literature are able to predict sub-optimal pseudoknotted secondary structures: pKiss [4] and McGenus [2]. We computed the sensitivity, positive predictive value (PPV), and F_1-score, weighted by the rank of the solution. The rank indicates the distance to optimality. A solution with a low rank is more interesting since it is closest to the optimum solution. Concerning the weighted mean sensitivity, BiokoP outperforms the other approaches, specifically McGenus; pKiss is competitive when 1 or 2 solutions are returned. Relative to the weighted mean PPV, BiokoP outperforms McGenus and is comparable to pKiss. In Figure 1 we present the weighted mean F_1-score obtained by each tool on our dataset, in function of the number of returned solutions. The weighted F_1-score of BiokoP is never inferior to the weighted F_1-score of pKiss and McGenus. The weighted F_1-score is quite stable for BiokoP. This suggests that the quality of predicted structures is stable when the number of returned solutions increases.

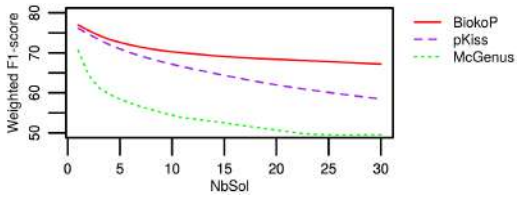


Fig. 1. Weighted F_1-score of the structures predicted with BiokoP, pKiss and McGenus on our dataset, in function of the number of solutions (NbSol).

4 Conclusion

In this work, we provide an original approach for predicting optimal and sub-optimal RNA secondary structures with pseudoknots with respect to two thermodynamic models. We combined two models into a bi-objective integer program (BOIP), and proposed a generic novel algorithm to compute the sets of optimal and sub-optimal solutions for any BOIP. The tests performed on the resulting software, BiokoP, have confirmed the importance of considering sub-optimal solutions. BiokoP was compared with pKiss [4] and McGenus [2]. The results show that BiokoP gives higher weighted mean F_1-score, considering a set of 161 pseudoknotted secondary structures. A track for our work is to propose better mono-criterion models to be combined to increase the global quality of the solutions found.

References

1. Egon Balas and Robert Jeroslow. Canonical cuts on the unit hypercube. *SIAM Journal on Applied Mathematics*, 23(1):61–69, 1972.
2. Michaël Bon, Cristian Micheletti, and Henri Orland. McGenus: a Monte Carlo algorithm to predict RNA secondary structures with pseudoknots. *Nucleic acids research*, pages 1895–1900, 2012.
3. Stéfan Engelen and Fariza Tahi. Tfold: efficient in silico prediction of non-coding RNA secondary structures. *Nucleic acids research*, 38(7):2453–2466, 2010.
4. Stefan Janssen and Robert Giegerich. The RNA shapes studio. *Bioinformatics*, pages 423–425, 2014.
5. Unyanee Poolsap, Yuki Kato, and Tatsuya Akutsu. Prediction of RNA secondary structure with pseudoknots using integer programming. *BMC bioinformatics*, 10(Suppl 1):S38, 2009.
6. Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu, and Kiyoshi Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):i85–i93, 2011.
7. Michela Taufer, Abel Licon, Roberto Araiza, David Mireles, FHD Van Batenburg, Alexander P Gultyaev, and Ming-Ying Leung. PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. *Nucleic acids research*, 37(suppl 1):D127–D135, 2009.
8. H Paul Williams. *Model building in mathematical programming*. Wiley, 2013.

MACHINE LEARNING
AND
DEEP LEARNING

TALK SESSION

Robust deep learning: A case study
[Talk / Poster submission]

Victor Estrade, Cecile Germain, Isabelle Guyon, David Rousseau

Laboratoire de Recherche en Informatique

Abstract. We report on an experiment on robust classification. The literature proposes adversarial and generative learning, as well as feature construction with auto-encoders. When domain-specific a priori knowledge is available, as in our case, a specific flavor of DNN called Tangent Propagation is an effective and less data-intensive alternative.

Keywords: Domain adaptation, Deep Neural Networks, High Energy Physics

1 Motivation

This paper addresses the calibration of a classifier in presence of systematic errors, with an example in High Energy Physics.

An essential component of the analysis of the data produced by the experiments of the LHC (Large Hadron Collider) at CERN is the selection of a region of interest in the space of measured features. Classifiers have become the standard tool to optimize the selection region. In the case of discovery and measurement of a new particle such as the Higgs boson, by definition no real labeled data are available. The classifier has to be trained on simulated data [1].

This introduces two kind of errors: *statistical* and *systematic*. When, as it is the case here, the data model is known, the statistical error essentially comes from the limited size of the training data. Systematics are the "known unknowns" of the data model, in statistical parlance the *nuisance parameters* that coherently bias the training data, but which exact values are unknown.

Formally, for a family of classifiers parameterized by θ (e.g. the architecture and hyperparameters of a neural network), let $h(., \theta)$ be the score function of classifier h and Z be the nuisance parameter. Robustness means that $h(., \theta)$ and Z should be independent (h should be *pivotal*). Of course, h should also be a good classifier, which helps to situate robustness as a regularization objective.

2 Domain Adaptation

Learning with systematics fall under the theory of domain adaptation [4]. Implementations have to choose between two strategies: either a knowledge-free setting, where the invariances are discovered from the data; or the integration of prior knowledge. The knowledge-free adaptation can be supervised, with Generative Adversarial Networks [6],[7], or semi supervised with Domain Adversarial Networks [5]. It requires large training sets, representative of the nominal and perturbed data distributions. In the HEP context, the cost of precise simulations would be too high.

In the second case, the invariances describe the expected robustness properties typically as small geometric transforms in the feature space. The Tangent Propagation

(TP) algorithm, proposed long ago [9] and recently revived [8], provides a principled method to integrate the invariance constraints into the learning of the data model with a classifier. With TP, the systematics are considered as a transformation $f(x, Z)$ of the input. The objective is to have $h(x, \theta) = h(f(x, Z), \theta)$, thus the model is regularized by : $\frac{\partial h(f(x,Z),\theta)}{\partial Z}$ i.e. the partial derivative of the classifier score wrt the nuisance parameter. As usual, a parameter noted λ in the following, controls the tradeoff between the classification error and the regularization.

3 Experimental results

The dataset We use the dataset of the HiggsML challenge [2], `http://opendata.cern.ch/record/328?ln=en`. Data is split between training and test sets with 5-fold cross validation All training is performed at the nominal setting. The systematics are introduced in the **test set** only.

Figure of merit The figure of merit is not the classication accuracy, but a non-linear function of true and false positives related to error propagation in measurement [3]. Let s_0 and b_0 be the number of true and false postives at nominal, and s_Z and b_Z their counterparts with systematics at Z. The figure of merit is $\sigma_\mu = \sqrt{\Sigma_0^2 + \Sigma_Z^2}$, where $\Sigma_0 = \frac{\sqrt{s_0+b_0}}{s_0}$ is the statistical error and $\Sigma_Z = \frac{s_Z+b_Z-(s_0+b_0)}{s_0}$ the systematic error. Because this function is not additive in the examples, we use the regularized classification error as a proxy to train the classifier.

Evaluation methodology The baseline is a DNN without TP (or equivalently a TP-DNN with $\lambda = 0$). As TP constrains the architecture (softmax activations), we also include results for a standard (RELU-based) DNN. In order to make the comparison manageable, the dimensioning hyper parameters are identical for all architectures : 3 hidden layers of 40 neurons each. All networks were trained for 2000 iterations with a mini-batch size of 1024 and optimized with Adam method, and a learning rate of 0.01.

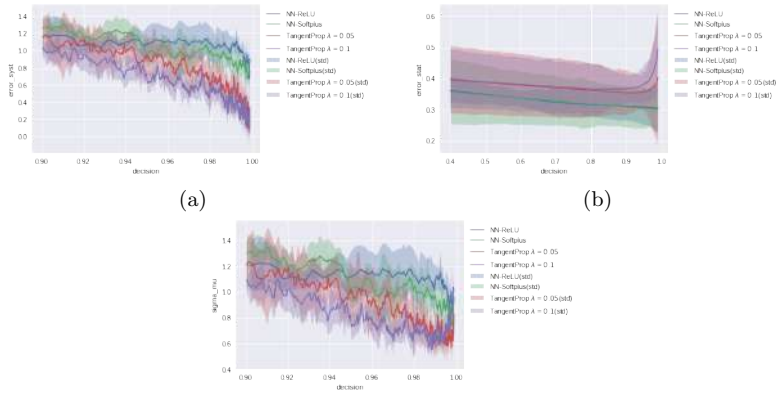
Results Figure 1a shows that TP consistently reduces the systematic error Σ_Z, by 20% on average near the minimum. The narrow confidence intervals support the significance of this result. For all architectures, Σ_Z is very noisy. As a similar behavior is observed with gradient boosting, noisiness is probably intrinsic to the problem.

Figure 1b highlights the complex impact of TP on the statistical error Σ_0. The much wider confidence intervals with TP might be due to the limits of cross-validation. But, as the TP-NN is trained to ignore some variability, this might indicate that this variability crosses the class boundary, i.e. that the gap between the class manifolds is too small. Experimenting with the boostrap may help disentangling this two causes.

Figure 1c shows that TP has a net positive effect on σ_μ. Other experiments (not reported here) show that this remains true for sensible ranges of Z and λ.

4 Conclusion

The positive results of this preliminary work show that the tangent propagation approach can be effective to reduce the systematic error even in the extremely difficult HEP case. Further experiments comparing this methods with adversarial networks and ensemble methods are in progress. We will also refine the implementation with better hyper-parameter selection and explore the bootstrap. As systematics are pervasive



(a) (b)

(c)

Fig. 1: Performance comparison for $Z = -1\%$. The values are the mean and standard deviation of the 5-fold cross validation. The decision threshold range corresponds to the constraints of physics analysis [1]

in scientific measurements, we envision the creation of a *systematics challenge*, in the spirit of the AutoML challenge for future work.

References

1. Claire Adam-Bourdarios, Glen Cowan, Ccile Germain, Isabelle Guyon, Balzs Kgl, and David Rousseau. The Higgs boson machine learning challenge. In *HEPML@ NIPS*, pages 19–55, 2014.
2. ATLAS Collaboration. Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014, 2014.
3. Roger Barlow. Systematic errors: Facts and fictions. In *Advanced Statistical Techniques in Particle Physics. Proceedings, Conference, Durham, UK, March 18-22, 2002*, pages 134–144, 2002.
4. Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, May 2010.
5. Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Franois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *arXiv:1505.07818 [cs, stat]*, May 2015. arXiv: 1505.07818.
6. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.
7. Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to Pivot with Adversarial Networks. *arXiv:1611.01046 [physics, stat]*, November 2016. arXiv: 1611.01046.
8. Salah Rifai, Yann Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The Manifold Tangent Classifier. In *NIPS*, volume 271, page 523, 2011.
9. Patrice Y. Simard, Bernard Victorri, Yann LeCun, and John S. Denker. Tangent Prop - A Formalism for Specifying Selected Invariances in an Adaptive Network. In John E. Moody, Stephen Jose Hanson, and Richard Lippmann, editors, *NIPS*, pages 895–903. Morgan Kaufmann, 1991.

End-to-end Deep Learning Approach for Demographic History Inference

[Poster/Talk submission]

Théophile Sanchez, Guillaume Charpiat, Flora Jay

LRI, Université Paris-Saclay, INRIA, CNRS

Abstract. Recent methods for demographic history inference have achieved good results, avoiding the complexity of raw genomic data by summarizing them into handcrafted features called *expert statistics*. Here we introduce a new approach that takes as input the variant sites found within a sample of individuals from the same population, and infers demographic descriptor values without relying on these predefined *expert statistics*. By letting our model choose how to handle raw data and learn its own way to embed them, we were able to outperform a method frequently used by geneticists for the inference of two demographic descriptor values while using less data.

Keywords: Deep Learning, Population Genetics, Demographic Inference

1 Motivation

Genetic variation within the same population carries signals of the past demography. For instance, a large population with small genetic diversity is likely to have encountered a bottleneck in the past, i.e. a sharp reduction of the population size followed by an expansion. Therefore, the study of genetic variations within a population allows to reconstruct its past history by inferring demographic descriptor values (e.g. effective population size over time or the date of events such as a bottleneck). However, demographic inference remains difficult since evolution is a stochastic process and only a few present-day individuals are sequenced. Thus, different histories can leave similar signals. Moreover, these signals are blurred by other processes such as natural selection.

We aim to infer the demographic descriptor values of a population by comparing genetic sequences from a sample of present-day individuals. The common approach in the literature summarizes the genetic variation among sequences into features called *expert statistics*, such as the site frequency spectrum (SFS), linkage disequilibrium (LD) and identity-by-state (IBS). Then, it uses supervised methods applied on simulated data such as Approximate Bayesian Computation (ABC) [1] or neural networks [2] to train a model that predicts the descriptor values of the demographic history. Here we have developed a new method based directly on raw data (i.e. genetic sequences) using neural networks with convolutional filters. With an approach that does not rely on *expert statistics* we allow more flexibility in the inference model, hoping to improve how the model extracts information and constructs a representation of the data.

2 Methods

Artificial neural networks need supervised training in order to tune their weights. Training our convolutional network cannot be done with real datasets as there are not enough

studies observing directly the effects of known demographic histories on population genetic variations. Instead, we use *fastsimcoal* [3], a program that simulates samples of genetic sequences given a certain demographic history and the values of its descriptors. Each simulation generates a matrix X of 300 successive variant sites found among 20 haploid individuals and a vector d of the variant site positions in the genome. A position is encoded as a distance in base pairs (bp) to the previous variant site. Element $x_{ij} \in \{0, 1\}$ of X stands for the variant version (two possible alleles) of the individual i at the position d_j. Out of 3,000,000 matrices simulated from 30,000 different demographic scenarios (i.e. a set of demographic history descriptor values), we use 90% for training and left out 10% for validation.

For simplification, we consider a particular demographic history consisting of a bottleneck followed by two expansions. It is defined by 8 descriptors: the population sizes before and during the bottleneck (NAncient and NBot), the population sizes after each expansion (NInterm and NCurrent), the date of the bottleneck (TStartBot), its length (BotLeng), the date of the last expansion (TRecentExp) and the sequencing error rate (Err). Descriptors of population sizes are log-transformed so that estimation errors between predicted and true sizes are expressed relatively to the population size. All descriptor values are standardized to contribute equally to the loss computation function.

To estimate these descriptor values, we consider convolutional networks. They have shown their ability to handle large data and recognize patterns, leading to impressive scores for image recognition problems. Also inspired by their ability to make use of spatial information and combine different scales, we developed SPI-DNA (Sequence Position Informed Deep Neural Architecture), a convolutional architecture that takes as input a heterogeneous pair consisting of a matrix X and its associated vector d. Our network outputs predictions for the 8 demographic descriptors used to simulate the input. The first layer of our network consists in 250 convolution filters of 5 different rectangular shapes applied to X. 100 other filters of 5 different shapes are applied to d. The results of the first convolutional layer are then concatenated so that the second convolutional layer will couple information from X and d in a way that emphasizes the original location of X along the genome. The outputs of this second layer are then combined and go through 5 convolutional layers and 2 fully connected layers. Adding convolutional layers one after each other allows our network to combine patterns and reduce the size of the data without adding too many weights to our model. The network also includes the following optimization setups: batch normalization, RELU activation function, adaptive moment estimation (adam) optimizer and mean square error (MSE) loss. It is implemented in *Pytorch*. During the validation, we compute the average of the 100 inferred demographic descriptor values for each scenario and then compute its loss, that we call prediction error.

3 Results

We compared the prediction of the tailored architecture against two other architectures:

- Multilayer perceptron (MLP) with one fully-connected layer of 20 neurons and one output layer of 8.
- 3-layer convolutional network with one layer of convolution, one fully-connected layer of 10 neurons and one output layer of 8. The layer of convolution applies 10 filters of length 5 on d and 30 filters of dimension 20×5 on X.

These two architectures have a number of weights similar to our convolutionnal network. They both use the same optimisation setups as our architectures.

Table 1 shows that from the three architectures, our tailored convolutional network performs the best in average and by descriptors. The MLP makes a prediction error close to 1 for the bottleneck length and the sequencing error rate, meaning that this network outputs random values for these two descriptors. Overall, convolutional networks have better results and our model lead to an improvement of 10% of the average prediction compared to the 3-layer convolutional network and 21% compared to the MLP. ABC still achieves better results on a similar dataset with a higher sequencing error (Err), but our method achieves better inference for two descriptors (NCurrent and NAncient).

Model	# weights	NCurrent	NInterm	NBot	NAncient	TStartBot	TRecentExp	BotLeng	Err	Mean
MLP	126228	.1165	.5307	.3222	.1684	.7886	.6679	.9783	.9605	.5667
3-layer ConvNet	121688	.1009	.5160	.2221	.1319	.6262	.5977	.9739	.8135	.4978
SPI-DNA	128568	**.0764**	.4851	.2043	**.0941**	.4808	.5474	.9445	.7670	.4500
ABC	∅	.0994	**.4073**	**.1380**	.1536	**.3461**	**.4218**	**.8603**	**.0136**	**.3050**

Table 1. Number of weights and prediction error of each descriptor for the three artificial neural network architectures. Prediction errors were computed on a validation set after each 10 batches of 300 inputs of a 2 epochs training. The ones corresponding to the best step (minimum average error) are reported here. For comparison, we added results of the ABC method on a similar dataset with a higher sequencing error (Err) and variable number of variant sites.

4 Discussion/Conclusion

We showed how the choice of architecture can improve the performance of a population genetics inference task. Having multiple convolutional filters of different sizes in the first layer allows our network to capture patterns at different scales that are then combined in the next layers. It is also able to merge heterogeneous large raw data while keeping a low number of weights, preventing overfitting. Even if our method has achieved good results and outperforms the ABC for the prediction of two descriptors, ABC has still overall better results. Despite this, we are confident in the value of SPI-DNA because ABC relies on a tedious work of identifying expert statistics relevant for a task. Understudied population genetics tasks could thus benefit from our new approach. Furthermore, there is room for improvement: the SPI-DNA network currently handles less data than what is available and used by ABC. This is why our next goal is to construct a deep learning architecture that handles a changing number of variant sites, and thus captures the maximum of information from the entire genomic dataset. This is a challenging milestone as most deep learning architectures are not designed to handle data of different sizes, and thus a brand new method is needed.

References

1. S. Boitard, W. Rodríguez, F. Jay, S. Mona, and F. Austerlitz, "Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach," *PLOS Genetics*, vol. 12, pp. 1–36, 03 2016.
2. S. Sheehan and Y. S. Song, "Deep Learning for Population Genetic Inference," *PLOS Computational Biology*, vol. 12, pp. 1–28, 03 2016.
3. L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, "Robust Demographic Inference from Genomic and SNP Data," *PLOS Genetics*, vol. 9, pp. 1–17, 10 2013.

MACHINE LEARNING
AND
STATISTICAL THEORY

TALK SESSION

Supervised layer Self-Organizing Maps with reject options

Talk submission

Ludovic Platon, Farida Zehraoui and Fariza Tahi

IBISC laboratory, UEVE/Genopole/Universit Paris-Saclay, Evry, France

Abstract. We present here a new approach of supervised self organizing map (SOM). We added a supervised perceptron layer to the classical SOM approach. This combination allows the classification of new patterns by taking into account all the map prototypes without changing the SOM organization. We also propose to associate two reject options to our supervised SOM. This allows to improve the results reliability and to discover new classes in applications where some classes are unknown. The results indicate that our approaches are competitive with most popular supervised leaning algorithms like support vector machines and random forest.

Keywords: Self-Organizing Maps, Neural network, TensorFlow, Reject option

1 Introduction

Clustering is the most popular tool for exploratory analysis of data. It can help to guide the analysts to understand the data, if the goal of the analysis is well defined and captured by the clustering model. Most of the time, this goal is expressed by labels representing categories of classes in a given application. For instance, in the biology domain, one can be interested by classifying gene functions, classifying different types of non coding RNA involved in a given biological process, and so on. But in many applications like the cited above, some labels associated to the data are not available and/or some labels are not known in advance.

In this work, we are interested by the self-organizing maps (SOM) [4], which are among the most used connectionist models for data clustering and visualization. They can also be extended to perform the classification task. We propose a two-layer supervised SOM, where the first layer consists of the classical unsupervised SOM and the second layer consists of perceptrons, which are linked to the SOM units. In addition, our approach combines the clustering to classification with reject option, in order to consider partial information on certain labels (some classes must be discovered).

2 Methods

The SOM models comprise an important class of competitive neural models. The main difference between the SOM and standard competitive networks is that the output

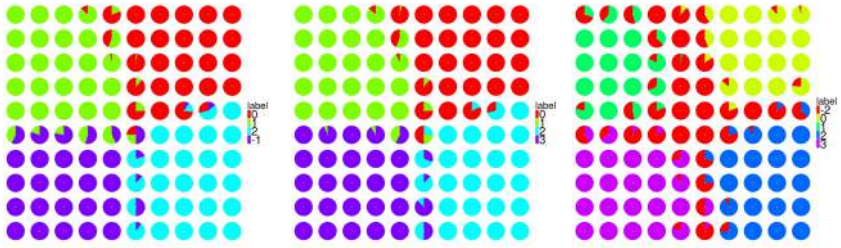
neurons are arranged in specific geometrical forms.
Each SOM unit r is associated with a weight vector $w_r^h = [w_r^{h1}, w_r^{h2}, ..., w_r^{hm}]^T \in \mathbb{R}^m$ with the same dimension as the input feature vector $x = [x_1, x_2, ..., x_m]^T \in \mathbb{R}^m$. The learning algorithm that leads to a self-organization can be summarized in two steps. The first one is the computation of the best-matching unit (BMU) $s(x)$ by using a distance measure (Euclidean distance, for example) between the input and weight vectors ($s(x) = \arg\min_{r \in A} d(x, w_r^h)$). In the second, the BMU and its neighbours weigths are updated towards the input x as follows: $w_r^h(t+1) = w_r^h(t) + \alpha^h(t) h_r[x - w_r^h(t)]$ where $\alpha_h(t) = 1 - \frac{t}{T}$ (with T the total number of iterations) is the learning rate and $h_r = exp(-\frac{-x^2}{\alpha_h(t) * \sigma})$ (with σ the radius of the map) is the neighbourhood function.

The perceptrons in the output layer receive the activation pattern provided by the SOM. The activation function of the unit r can be computed with $a_r = exp^{-\gamma d^2(x, w_r^h)}$. We extent the output layer with two different rejection options, distance rejection and ambiguity rejection. Like in article [1], we assume that the learning step of our model has already been completed. In the distance rejection, the neural network rejects the classification of an input x if the largest probability of the predicted class o^* is lower than a defined threshold β_{max}. In the ambiguity rejection, the neural network rejects the classification of an input x if the difference between the largest probability for predicting the class o^* and the second largest probability for predicting the class o^{**} is lower than a threshold β_{diff}.

3 Results

Our method, called SLSOM, is implemented using the TensorFlow Python API. To evaluate the proposed method, we considered six different datasets. The first one is an artificial dataset (composed of four Gaussian datasets where the clusters 0 and 1 are overlapping), and the five others are extracted from the UCI database (SPECTF, E.coli, Iris WDBC, Cardiotocography). We compared our method SLSOM to several classical supervised learning algorithms (Support Vector Machines, Random Forest, Gaussian Naïve Bayes, K-Nearest Neighbors and Logistic Regression) and existing post labelling supervised SOM presented in the articles [3],[2],[6],[5]. Our method gives better results than all the other existing supervised SOM on all of the datasets. Furthermore, our method is very competitive with the classical supervised agorithms like Gaussian Naïve Bayes, K-Nearest Neighbors, Logical Regression and SVM. On the WDBC dataset, our method outperforms all the other classifiers. SLSOM has an accuracy of 0.97 when the second best classifier (RF) has an accuracy of 0.94 and the remaining methods have an accuracies between 0.83 and 0.90. The result are shown in [7].

To evaluate the performance of our supervised SOM with the two reject options, we used the artificial dataset. To show the interest of the distance rejection, we suppressed the examples of class 3 from the training set and presented the examples of this class to our supervised SOM in the test set. For the ambiguity rejection, we used the four classes in the training and test sets. Rejected examples are labelled (-1) and (-2) respectively in Figures 1a and 1c. In Figure 1a we can see that the examples of the class 3 are rejected by our approach. We can verify that the rejected examples correspond to the class 3 represented by the standard SOM (Figure 1b). In Figure 1c the rejected examples correspond mainly to the examples, which are located in the frontiers of two classes.



(a) Distance reject
accuracy = 0.92
$\beta_{max} = 0.4$
reject = 0.25

(b) Predicted label
accuracy = 0.94

(c) Ambiguity reject
acc = 0.97
$\beta_{diff} = 0.2$
reject = 0.1

Fig. 1: Predicted label repartition for the artificial dataset

4 Conclusion

In this paper, we have presented a new approach of supervised SOM, where we take into account all the map prototypes to classify new patterns. In this approach, we have added a supervised perceptron layer on the top of the SOM. Experiments done on several datasets show very good results comparing to other supervised SOM. These results are competitive (or better) with those obtained using efficient classification methods like SVM. We also added two rejection options to our supervised SOM in order to improve the classification results, discover new classes and analyse examples belonging to these classes. The distance rejection is used to discover new classes when the ambiguity rejection improves the prediction on overlapping dataset.

References

1. H. Ishibuchi and M. Nii. Neural networks for soft decision making. *Fuzzy Sets and Systems*, 115(1):121–140, 2000.
2. S. Kittiwachana and K. Grudpan. Supervised self organizing maps for exploratory data analysis of running waters based on physicochemical parameters: a case study in chiang mai, thailand. *Asia-Pacific Journal of Science and Technology*, 20(1):1–11, 2015.
3. T. Kohonen. The'neural'phonetic typewriter. *Computer*, 21(3):11–22, 1988.
4. T. Kohonen. Self-organizing maps, volume 30 of springer series in information sciences, 1995.
5. K. W. Lau, H. Yin, and S. Hubbard. Kernel self-organising maps for classification. *Neurocomputing*, 69(16):2033–2040, 2006.
6. C. L. Mattos and G. A. Barreto. Artie and muscle models: building ensemble classifiers from fuzzy art and som networks. *Neural Computing and Applications*, 22(1):49–61, 2013.
7. L. Platon, F. Zehraoui, and F. Tahi. Self-organizing maps supervised layer. In *12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM+)*, forthcoming.

On the benefits of output sparsity for multi-label classification

[Poster/Poster demo/Talk submission]

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Joseph Salmon

Télécom ParisTech, Université Paris-Est

Keywords: classification, multi-label, sparsity

Abstract. The modern multi-label problems, where each observation can be associated with a set of labels, are typically large-scale in terms of number of observations, features and labels. Moreover, the amount of labels can even be comparable with the amount of observations. In this context, different remedies have been proposed to overcome the curse of dimensionality. In this work, we aim at exploiting the output sparsity by introducing a new loss, called the sparse weighted Hamming loss. This proposed loss can be seen as a weighted version of classical ones, where active and inactive labels are weighted separately. Leveraging the influence of sparsity in the loss function, we provide improved generalization bounds for the empirical risk minimizer, a suitable property for large-scale problems. For this new loss, we derive rates of convergence linear in the underlying output-sparsity rather than linear in the number of labels. In practice, minimizing the associated risk can be performed efficiently by using convex surrogates and modern convex optimization algorithms. We provide experiments on various real-world datasets demonstrating the pertinence of our approach when compared to non-weighted techniques.

1 Introduction and Motivation

Multi-label classification (MLC) has recently attracted a vast amount of contributions due to the variety of problems that MLC can model: text categorization [Gao et al., 2004], functional genomics [Barutcuoglu et al., 2006], image classification [Li et al., 2014] to name a few. The objective in multi-label classification is to predict a binary vector $Y \in \{0,1\}^L$ for a given observation $X \in \mathcal{X}$, where L is the number of labels available and $\mathcal{X} = \mathbb{R}^D$ is the feature space. Our motivation in this work starts from the observation that for a given $X \in \mathcal{X}$ the output label vector Y is often sparse: for each data point only a few labels are generally active in real-world scenarios (*i.e.,* Y has few non-zero coordinates, say K, with $K \ll L$). We propose a novel sparsity assumption, which restricts a possibility to observe a large amount of active labels. In contrast, Hsu et al. [2009] assumed that the label vector Y is sparse in average, which still may lead to non-sparse observations.

	$\hat{Y}_0 \equiv 0$	$\hat{Y}_1 \equiv 1$	$\hat{Y}_{2K}$	$\hat{Y}_w \equiv 1 - Y$
Proposed loss	$p_1 K$	$p_0(L-K)$	$p_0 K$	$p_1 K + p_0(L-K)$
Hamming loss: $p_0 = p_1 = 0.5$	$\frac{K}{2}$	$\frac{L-K}{2}$	$\frac{K}{2}$	$\frac{L}{2}$
[Jain et al., 2016]: $p_0 = 0, \quad p_1 = 1$	K	0	0	K
Our choice: $p_0 = \frac{2K}{L}, \quad p_1 = 1 - \frac{2K}{L}$	$K - \frac{2K^2}{L}$	$2K - \frac{2K^2}{L}$	$\frac{2K^2}{L}$	$3K - \frac{4K^2}{L}$

Table 1. Loss examples costs for several classifiers with underlying true label being K-sparse, with $K \ll L$: $\hat{Y}_0 \equiv 0$: output no label, $\hat{Y}_1 \equiv 1$ output all labels, $\hat{Y}_{2K}$: output correct active set plus K mistakes on inactive set, $\hat{Y}_w \equiv 1 - Y$: always wrong

2 Contribution

In recent work Jain et al. [2016] proposed to omit inactive labels to build classifiers, this choice of loss function gives a big promotion to classifiers which predict that all labels are active. In contrast, standard decomposable losses promote classifiers which predict that all labels are inactive Balancing between the two choices we manage to construct new loss function, with two main properties:

- it avoids naive predictions like $Y = (0, \dots, 0)$ or $Y = (1, \dots, 1)$, giving promotion to predictions with few mistakes on inactive labels and accurately predicted active labels,
- we prove a generalization bounds, which are linear in the underlying sparsity K instead of linearity in the total amount of labels L.

For a given label vector Y and prediction vector $\hat{Y}$ our loss has the following form

$$\mathcal{L}^w_{0/1}(Y, \hat{Y}) = \sum_{l=1}^{L} \Big\{ p_0 \mathbb{1}_{\{\hat{Y}^l = 1\}} \mathbb{1}_{\{Y^l = 0\}} + p_1 \mathbb{1}_{\{\hat{Y}^l = 0\}} \mathbb{1}_{\{Y^l = 1\}} \Big\} \ ,$$

where Y^l is l^{th} component of the vector $Y \in \{0,1\}^L$. For more insight, let us consider the scenario where Y is exactly K-sparse and let us analyze the following classifiers (a synthesis is also given in Table 1):

- $\hat{Y}_0 \equiv 0$: predicts all labels inactive,
- $\hat{Y}_1 \equiv 1$: predicts all labels active,
- $\hat{Y}_w \equiv 1 - Y$: misspredicts all labels,
- $\hat{Y}_{2K}$: correctly predicts the active set of Y and makes exactly K mistakes on its inactive set.

Intuitively, one would like to build a loss which is able to differentiate between the first three predictions and $\hat{Y}_{2K}$. Indeed, in large-scale problems the predictions similar to $\hat{Y}_{2K}$ provide more valuable insights compared to the first three ones.

With our choice of weights

$$p_0 = \frac{2K}{L}, \quad p_1 = 1 - \frac{2K}{L} \ , \tag{1}$$

the introduced loss function treats $\hat{Y}_0, \hat{Y}_1$ and $\hat{Y}_w$ almost equally and it promotes predictions with small amount of mistakes on inactive sets and correct predictions on active sets. Meanwhile, the Hamming loss does not make any difference between $\hat{Y}_{2K}$ and $\hat{Y}_0$, and the loss considered in [Jain et al., 2016] gives a high promotion to naive classifiers like $\hat{Y}_1$.

Remark 1. In practice, the weights in Eq. (1) rely on the unknown sparsity constant K. Since this quantity is unknown to the practitioner, a simple strategy consists in performing a rough estimation based on the observed labels. Hence, we consider

$$\hat{p}_0 = \frac{2\hat{K}}{L}, \quad \hat{p}_1 = 1 - \hat{p}_0 \ , \tag{2}$$

where we estimate the output sparsity level by the maximal sparsity on the observations:

$$\hat{K} = \max_{i \in [N]} \sum_{l=1}^{L} Y_i^l \ .$$

Additionally, we prove generalization bounds which are beneficial due to their linear dependence on the underlying sparsity constant K, instead of being linear in L. Notice that bounds which are linear in L can fail to provide convergence, since the total amount of labels can grow in the same rate as the number of observations N. We also provide an empirical study of the proposed framework, showing a superior performance in terms of different multi-label classification measures.

Bibliography

Z. Barutcuoglu, R. E. Schapire., and O. G. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

S. Gao, W. Wu, C. H. Lee, and T. S. Chua. A MFoM learning approach to robust multiclass multi-label text categorization. In *ICML*, pages 329–336, 2004.

D.J. Hsu, M. Sham Kakade, J. Langford, and T. Zhang. Multi-label prediction via compressed sensing. In *NIPS*, pages 772–780, 2009.

H. Jain, Y. Prabhu, and M. Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *KDD*, pages 935–944, 2016.

X. Li, F. Zhao, and Y. Guo. Multi-label image classification with a probabilistic label enhancement model. In *UAI*, pages 430–439, 2014.

Scalable Model-based Cascaded Imputation of Missing Data

[Talk submission]

Jacob Montiel[1], Jesse Read[2], Albert Bifet[1], and Talel Abdessalem[1,3]

[1] LTCI, Télécom ParisTech, Université Paris-Saclay, Paris 75013, France,
{jacob.montiel, albert.bifet}@telecom-paristech.fr,
[2] LIX, École Polytechnique, Palaiseau 91120, France,
jesse.read@polytechnique.edu
[3] UMI CNRS IPAL & National University of Singapore
talel.abdessalem@enst.fr

Abstract. Missing data is a common trait of real-world data that can negatively impact interpretability. We present CASCADE IMPUTATION (CIM), an effective and scalable technique for automatic imputation of missing data. CIM is not restrictive on the characteristics of the input data. We compare CIM against well-established imputation techniques over a variety of data sets under multiple test configurations to measure the impact of imputation on the classification problem. Test results show that CIM outperforms other imputation methods over multiple test conditions. Additionally, we identify optimal performance and failure conditions for popular imputation techniques.

Keywords: imputation, missing data, supervised learning, classification

1 Motivation

Missing data is a common phenomenon in real-world applications, with an average estimated in a range between 5% and 20% [6]. Performance of supervised learning methods can be negatively affected by missing data, according to [1], ratios between 5-15% require the usage sophisticated methods while above 15% of missing values can compromise data interpretation .

Missing data mechanisms are classified into [4]: i) Missing Completely At Random (MCAR). The events behind missing values are independent of observable variables and the missing values themselves. ii) Missing At Random (MAR). Missingness can be explained by observable variables. iii) Missing Not At Random (MNAR). The reason for missingness of data is related to the value of the missing data itself.

Manual imputation of MCAR and MAR data is a time consuming process and requires deep understanding of the data and the phenomena that it describes. Current trends in data generation and collection result in larger and more complex data sets. Under this scenario, imputing data manually is impractical, *scalable* automatic imputation solutions are required for real-world applications. One of such real-world applications is *Classification*, a type of Supervised Learning.

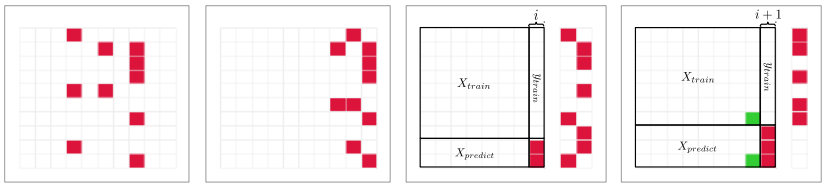
The contributions of this project are:

- A new scalable and effective model-based imputation method that casts the imputation process as a set of classification/regression tasks.

- Different to well established imputation techniques, the proposed method is non-restrictive on the type of missing data to process, including support for:
  - MAR and MCAR missing data mechanisms
  - Numerical and nominal data
  - Small to large data sets, including high dimensional data sets.
- The proposed method does not require additional tools to the ones available in the classification problem, making the integration imputation+classification straightforward.
- We provide a comprehensive evaluation of different imputation methods under multiple scenarios, identifying optimal-operation and failure conditions.

2 Proposed method

In this project we present CASCADE IMPUTATION (CIM), a model-based incremental imputation method. CIM is designed under the assumption that the imputation process can be cast as a set of classification/regression tasks, in the sense that unobserved values are imputed on a supervised learning fashion, in other words, a predictive model for missing data is generated based on input-response samples. The main steps in CIM are shown in Figure 1. Given an incomplete data set $D = (X, y)$ we want to find the corresponding imputed data set $D' = (X', y)$. Original positions of missing data are marked in red while incrementally imputed values marked in green are used in further iterations of the algorithm.



(a) **(b)** **(c)** **(d)**

Fig. 1: CIM Steps. (1a) Original data with missing values marked in red. (1b) Updated positions after sorting attributes by count of missing values. (1c) A classification/regression model is trained and applied for attribute *i*. In the next iteration *i*+1 (1d), a new classification/regression task is performed, previously imputed values are appended to the complete data. (1c-1d) steps are repeated until the data set is complete.

3 Methodology

We are interested in the impact of automatic missing data imputation on classification. In order to thoroughly evaluate our proposed method, we select datasets from a variety of domains, in both multi-class and multi-label problems. For our tests, we use 10 publicly available data sets. The imputation/classification test is composed by three major tasks:

i) **Generate missing data.** We remove all original missing values in the data sets in order to have 'complete' versions. Then, for each complete dataset we generate 10 datasets with 5%, 10%, 25% and 50% missing values using MAR and MCAR mechanisms.

ii) **Impute missing data.** We compare CIM against 4 missing data imputation techniques. *Constant Imputation* (CONSTANT) where a constant value is used to fill missing values. *Simple Imputation* (SIMPLE) where we use the 'mean value' to fill numerical values and the 'most-frequent value' for nominal values. *Expectation-Maximization Imputation* (EMI) [3] is an

iterative methods with two steps. In the expectation step (E), values are imputed based on observed values. In the maximization step (M), imputed values are evaluated and updated if necessary according to the data distribution. The EM algorithm converges to imputed values consistent with the observed values distribution. *k-Nearest Neighbor Imputation* (KNNI) [2] uses the neighborhood of a missing value to estimate the corresponding imputation value. Defining the optimal k value is challenging and has important implications on performance at the cost of computational burden [5]. In our tests we use $k = 3$, as a compromise given the range of data set sizes.

iii) **Use imputed data for classification.** We train classification models using the imputed data and compare the performance of these models against the baseline performance of models generated using complete data. We use 2 popular classification algorithms: Logistic Regression and Random Forest. We use default parameters for each classifier since we are interested on measuring the impact of using imputed data. We perform a total of $80 \times 6 \times 2$ tests and use 10-fold cross validation.

4 Discussion of Experimental Results

Test results show that CIM performs well on a variety of conditions, expanding beyond the operational range of sophisticated methods such as EMI and KNNI which does not scale well to large data sets. We evaluate performance using the Area Under the Receiver Operating Characteristic Curve (AUROC) for binary classifiers and F1-Score for multi-class classifiers. Then measure the Root Mean Square Error (RMSE) between the baseline performance (complete data) and the observed performance (imputed data) to evaluate performance over multiple missing data ratios. We find that CIM-RF is the overall best performer based on the accuracy over the range of missing values, followed by CIM-LR and SIMPLE. Optimal performance is achieved by CIM without using parameter tuning for the internal classification/regression tasks. Although CONSTANT imputes data successfully for all tests, its overall performance is low. KNNI is next as it shows good performance with small to medium size data sets while failing on large data sets. The last overall performer in our tests is EMI given its small operational range. Nonetheless, it is important to remark its good performance on small missing data ratios and small data sets. Also important is the impact of the classification algorithm on performance. Logistic Regression and CIM-RF are a good combination for both MAR and MCAR, while Random Forest improves the performance of CIM-LR and SIMPLE.

References

1. Acuña, E., Rodriguez, C.: The Treatment of Missing Values and its Effect on Classifier Accuracy. Classification, Clustering, and Data Mining Applications (1995), 639–647 (2004)
2. Batista, G.E.A.P.A., Monard, M.C.: A study of k-nearest neighbour as an imputation method. Frontiers in Artificial Intelligence and Applications 87, 251–260 (2002)
3. Dempster, A., Laird, N., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B Methodological 39(1), 1–38 (1977)
4. Little, R.J., Rubin, D.B.: Statistical analysis with missing data. John Wiley & Sons (2002)
5. Maier, M., Hein, M., Von Luxburg, U.: Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters. Theoretical Computer Science 410, 1749–1764 (2009)
6. Rahman, M.G., Islam, M.Z.: Missing value imputation using a fuzzy clustering-based EM approach. Knowledge and Information Systems pp. 389–422 (2015)

Ranking Data with Continuous Labels through Oriented Recursive Partitions

[Talk submission]

Stephan Clémençon and Mastane Achab

LTCI Telecom ParisTech, Université Paris-Saclay

Abstract. We formulate a supervised learning problem, referred to as *continuous ranking*, where a continuous real-valued label Y is assigned to an observable r.v. X taking its values in a feature space $\mathcal{X}$ and the goal is to order all possible observations x in $\mathcal{X}$ by means of a scoring function $s : \mathcal{X} \to \mathbb{R}$ so that $s(X)$ and Y tend to increase or decrease together with highest probability. This problem generalizes *bipartite ranking* to a certain extent and the task of finding optimal scoring functions $s(x)$ can be naturally cast as optimization of a dedicated functional criterion, called the IROC curve here, or as maximization of the Kendall τ related to the pair $(s(X), Y)$. From the theoretical side, we describe the optimal elements of this problem and provide statistical guarantees for empirical Kendall τ maximization under appropriate conditions for the class of scoring function candidates. We also propose a recursive statistical learning algorithm tailored to empirical IROC curve optimization and producing a piecewise constant scoring function that is fully described by an oriented binary tree. Preliminary numerical experiments highlight the difference in nature between *regression* and *continuous ranking* and provide strong empirical evidence of the performance of empirical optimizers of the criteria proposed.

Keywords: continuous ranking, bipartite ranking, ranking tree

1 Introduction

The predictive learning problem considered in this paper can be easily stated in an informal fashion, as follows. Given a collection of objects of arbitrary cardinality, $N \geq 1$ say, respectively described by characteristics $x_1, \ldots, x_N$ in a feature space $\mathcal{X}$, the goal is to learn how to order them by increasing order of magnitude of a certain unknown continuous variable y. To fix ideas, the attribute y can represent the 'size' of the object and be difficult to measure, as for the physical measurement of microscopic bodies in chemistry and biology or the cash flow of companies in quantitative finance and the features x may then correspond to *indirect measurements*. The most convenient way to define a preorder on a feature space $\mathcal{X}$ is to transport the natural order on the real line onto it by means of a (measurable) scoring function $s : \mathcal{X} \to \mathbb{R}$: an object with charcateristics x is then said to be 'larger' ('strictly larger', respectively) than an object described by x' according to the scoring rule s when $s(x') \leq s(x)$ (when

$s(x) < s(x')$). Statistical learning boils down here to build a scoring function $s(x)$, based on a *training* data set $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of objects for which the values of all variables (direct and indirect measurements) have been jointly observed, such that $s(X)$ and Y tend to increase or decrease together with highest probability or, in other words, such that the ordering of new objects induced by $s(x)$ matches that defined by their true measures as well as possible. This problem, that shall be referred to as *continuous ranking* throughout the article can be viewed as an extension of *bipartite ranking*, where the output variable Y is assumed to be binary and the objective can be naturally formulated as a functional M-estimation problem by means of the concept of ROC curve, see [3]. Refer also to [2], [4], [1] for approaches based on the optimization of summary performance measures such as the AUC criterion in the binary context.

2 Contributions

It is the major purpose of this paper to formulate the *continuous ranking* problem in a quantitative manner and explore the connection between the latter and bipartite ranking. Intuitively, optimal scoring rules would be also optimal for any bipartite subproblem defined by thresholding the continuous variable Y with cut-off $t > 0$, separating the observations X such that $Y < t$ from those such that $Y > t$. Viewing this way *continuous ranking* as a continuum of nested bipartite ranking problems, we provide here sufficient conditions for the existence of such (optimal) scoring rules and we introduce a concept of *integrated* ROC *curve* (IROC curve in abbreviated form) that may serve as a natural performance measure for continuous ranking, as well as the related notion of *integrated* AUC *criterion*, a summary scalar criterion, akin to Kendall tau. The paper also introduces a novel recursive algorithm that solves a discretized version of the empirical *integrated* ROC *curve* optimization problem, producing a scoring function that can be computed by means of a hierarchical combination of binary classification rules. Numerical experiments providing strong empirical evidence of the relevance of the approach promoted in this paper are also presented.

References

1. S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *J. Mach. Learn. Res.*, 6:393–425, 2005.
2. S. Clémençon, G. Lugosi, and N.Vayatis. Ranking and scoring using empirical risk minimization. In *Proceedings of COLT 2005*, volume 3559, pages 1–15. Springer., 2005.
3. S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.
4. Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

POSTER SESSION

Modeling Spatially-Correlated Cellular Networks by Applying Inhomogeneous Poisson Point Processes

Oral Talk

Shanshan WANG, Marco Di RENZO

CNRS-Laboratoire des signaux et systèmes-CentraleSupelec-Université Paris Saclay

Abstract. The distribution of base stations (BSs) are usually modelled in a Poisson Point Process (PPP) manner. While random deployments are not accurate for macro base stations. The non-PPP based approaches are much less mathematically tractable than PPP-based approach. This paper proposes a new mathematically tractable approach called Inhomogeneous Double Spatially-Thinned Poisson Point Process Modeling Approach. It can be applied to any spatial point process with repulsions & attractions. This approach is as simple to simulate as PPP-based approach. It has the same mathematical tractability and insightfulness as the PPP-based approach as well.

Keywords: non-PPP, spatially-correlated, performance evaluation

1 Introduction

In this paper, a detailed introduction on how to apply stochastic geometry for modeling, analyzing and optimizing 5G ultra dense cellular networks is provided. The ever-rising demand for wireless data implies that conventional cellular architectures based on large macro cells will soon be unable to support the anticipated density of high-data-rate users. The traditional approach of modeling macro cellular networks is not applicable anymore to ultra-dense network deployments. This is due to the large number of parameters and network configurations that need to be analyzed, which make simulation-based approaches too expensive and impractical.

The practical deployments of heterogeneous ultra-dense cellular networks are not totally random or regular. The widely-adapted approaches to model BSs are mostly based on PPP, which is not accurate for real BSs distribution. Base stations are deployed based on coverage, rate & data traffic criteria that make their locations spatially correlated. Currently available research works rely on two assumptions:

1) The base stations are always assumed to be randomly deployed (Poisson point process assumption), regardless of their type. There are plenty of literatures for PPP-based approach, e.g. [1][2][3][4]. However, random deployments are not accurate for macro base stations.

2) The base stations are modeled using some specific point processes (determinantal [5], Ginibre [6], etc.) that are mathematically tractable. However the mathematical frameworks obtained by using non-Poisson point processes are much less tractable

than their Poisson counterpart. And it does not provide any insight for system design. What's more, their computation may take longer time than optimized system-level simulations

The main contribution of this paper is: A new approach is proposed, which is called: Inhomogeneous Double Spatially-Thinned Poisson Point Process Modeling Approach, it models ultra-dense cellular networks with spatial attractions or repulsions; it is validated against real cellular network deployments; the new mathematical framework for system-level analysis is also developed.

2 Main Contribution

2.1 Sampling serving BS

The base stations are sampled according to a distance-depend function that accounts for the shortest distance properties of actual cellular network deployments and identify the serving base station.

The distance-based function used is called contact distance distribution. It can also be called empty space function, a spherical contact distribution function is defined as probability distribution of the radius of a sphere when it first encounters or makes contact with a point in a point process.

2.2 Sampling Interfering BSs

Another homogeneous Poisson point process is generated and sampled based on the location of the serving base station (previous subsection) and on the distance dependent properties of actual cellular network deployments. The resulting base stations constitute the interfering base stations

The obtained system model is a spatially-thinned version of the original Poisson point process that is mathematically tractable

3 Results

The following figure shows coverage probability Cauchy DPP (repulsive point process) using Double Thinning approach:

4 Conclusion

Our approach which is called Inhomogeneous Double Spatially-Thinned Poisson Point Process Modeling Approach, reproduces practical cellular networks generated by advanced statistical software (R). It is validated against real cellular network deployments as well as the new mathematical framework for system-level analysis. It is shown that the Poisson point process is less accurate. And the proposed proposed approach is obtained without loosing in tractability while it models ultra-dense cellular networks with spatial attractions or repulsions.

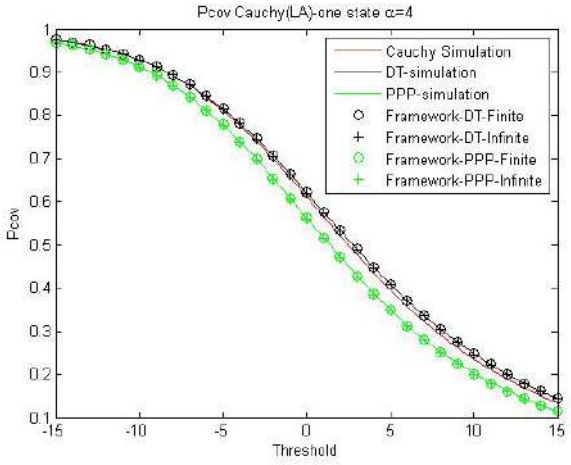


Fig. 1. Coverage Probability for Cauchy LA case with path loss exponent $\alpha = 4$

References

[1] M.DiRenzo,A.Guidotti,andG.E.Corazza,Average rate of downlink heterogeneous cellular networks over generalized fading channels A stochastic geometry approach, IEEE Trans. Commun., vol. 61, no. 7, pp. 30503071, July 2013.
[2] Haenggi, Martin, et al. "Stochastic geometry and random graphs for the analysis and design of wireless networks." IEEE Journal on Selected Areas in Communications 27.7 (2009).
[3] Dousse, Olivier, Franois Baccelli, and Patrick Thiran. "Impact of interferences on connectivity in ad hoc networks." IEEE/ACM Transactions on Networking (TON) 13.2 (2005): 425-436.
[4] Andrews, Jeffrey G., Francois Baccelli, and Radha Krishna Ganti. "A tractable approach to coverage and rate in cellular networks." IEEE Transactions on Communications 59.11 (2011): 3122-3134.
[5] Li, Yingzhe, et al. "Statistical modeling and probabilistic analysis of cellular networks with determinantal point processes." IEEE Transactions on Communications 63.9 (2015): 3405-3422.
[6] Deng, Na, Wuyang Zhou, and Martin Haenggi. "The Ginibre point process as a model for wireless networks with repulsion." IEEE Transactions on Wireless Communications 14.1 (2015): 107-121.

Channel modeling analysis of visible light communication by stochatsic geometry
[Oral talk]

Xiaojun XI , Marco DI Renzo

CNRS

Abstract. The channel modeling of indoor visible light communication based on stochastic geometry is proposed in this paper. This downlink performance of VLC system is analyzed, especially the coverage probability and the achievable average rate. The Poisson point process random cell is considered in order to obtain the lower upper for the practical attocell networks with irregular cell deployment. A tractable mathematical framework of VLC system is proposed and the result shows that it can be matched with computer simulation in high accuracy. Meanwhile, the influence of the density and field of view (FOV) in this model are compared and discussed in the final result.

Keywords: VLC, stochastic geometry, channel modeling

1 Motivation

The higher the transmission frequency, the higher the attenuation that the signals usually undergo. This implies that transmission technologies in the THz and VLC spectrum can be applied to shorter transmission distances. The same applies to MC due to their low propagation speed. This implies that future networks will need to be very ultra dense, much more that current and 5G networks are expected to be. The analysis and design of such networks cannot be conducted by using conventional methodologies because they are not scalable with the network density and size. In addition, approaches based on numerical simulations are not affordable due to the long simulation times, the amount of memory that is needed for simulations, as well as the many parameters that affect the system performance, which would require too many options to be analyzed before identifying the optimal setup. The fact that multiple technologies can be used (based on radio, light and chemical signals) increases the complexity of the problem to a much larger extend. The result is that new approaches need to be used for modeling the locations of the access points and of the mobile devices.

Today, the current approach for handling at least in part this issue is to rely upon tools from stochastic geometry tools and more in particular on the theory of Poisson point processes. Unfortunately, this approach is not applicable anymore and, at the time of writing, there are not tractable and accurate approaches that overcome this limitation. The underlaying assumption of Poisson point processes is that the access points are distributed at random, without spatial interactions. This can serve as a rst

approximation but it is not true in reality and is not acceptable in emerging networks, based on a mixture of radio, light and chemical signals. Let us consider two examples that are related to light and chemical signal transmission. Light-based communication can be used either in indoor or outdoor, the rst being the most promising in terms of revenues. In these cases, LEDs are expected to be deployed in a regular fashion: for example, data can be transmitted from lamp posts, which are regularly deployed in the streets, or data is transmitted by indoor deployments that form regular grids. As for molecular communications, molecular transmitters and receivers are expected to work in teams of nano-machines, rather than being randomly scattered in space. In the rst case, the devices show repulsive characteristics while in the second case they exhibit attractive properties. The problem is that, at the time of writing, there are no tractable mathematical methodologies for handling these spatial structures, which can be efciently used for system optimization and to gain insight for system design. New methodologies are needed. In the sequel, I will discuss an innovative approach that I have been working on for the last year, which is still unpublished and that, so far, has been tested only for application to radio-based networks. It is, however, a promising approach for application to light and chemical communication networks as well.

2 Methods

In this paper, we apply the stochastic geometry to analyze the indoor visible light communication system, which is fast and trackable compared with previous approach. Through our approach, we can obtain an exact mathematical framework without any approximation. In addition, we extend this model by considering the field of view and blockage model to match the practical VCL system.

3 Results

The mathematical framework of coverage probability and average rate can match the simulation results in high accuracy. Meanwhile, through our framework and simulation results, we can prove that there exists an optimal value for the FOV and density.

4 Conclusion

The paper presents a trackable framework for the system performance of an indoor visible light communication system by stochastic geometry, which has a high accuracy compared with the simulation results. Meanwhile, we extend our model by considering the FOV and blockage model, which is practical and useful. In addition, our results show that we can obtain an optimal value of coverage probability and average rate by considering the density and FOV.

References

[1] Chen, Cheng, Dushyantha A. Basnayaka, and Harald Haas. "Downlink performance of optical attocell networks." Journal of Lightwave Technology 34.1 (2016): 137-156.

[2] Di Renzo, Marco, and Peng Guan. "Stochastic geometry modeling of coverage and rate of cellular networks using the Gil-Pelaez inversion theorem." IEEE Communications Letters 18.9 (2014): 1575-1578.
[3] Chen, Cheng, DushyanthaBasnayaka, and Harald Haas. "Downlink SINR Statistics in OFDM-Based Optical Attocell Networks with a Poisson Point Process Network Model." 2015 IEEE Global Communications Conference (GLOBECOM). IEEE, 2015.
[4] Bai, Tianyang, and Robert W. Heath. "Coverage and rate analysis for millimeter-wave cellular networks." IEEE Transactions on Wireless Communications 14.2 (2015): 1100-1114.
[5] Di Renzo, Marco, Alessandro Guidotti, and Giovanni E. Corazza. "Average rate of downlink heterogeneous cellular networks over generalized fading channels: A stochastic geometry approach." IEEE Transactions on Communications 61.7 (2013): 3050-3071.

Graphons for Network Modeling : towards a notion of complexity

Yann Issartel - Université Paris-Sud, École Polytechnique

Key-words : random graph, latent space graph, graphon, non-parametric estimation.

Summary

Networks are ubiquitous in many sciences : genomics, biology, statiscal physics, social science ... In order to extract information from these data repositories, random graphs have proven to be particularly relevant to model real-world networks. Specifically, random graphs with latent space can be characterized by the so-called graphon. Encompassing a wide span of graph models, this non-parametric perspective for analyzing network data has recently gained interest.

However, random graphs constructed on graphon suffer from a lack of interpretation : it can be hard to get simple description from an observation given by sampled graph. This problem poses challenging questions on how to define informative graphon properties and make inferences about them.

As a first answer to this problem, we suggest a suitable notion of complexity for graphons. Roughly speaking, a simple intuition of this graph feature may be the number of explanatory variables (age, job, ...) that are required to understand the presence or absence of connections between people in a social network. Finally, we propose a consistency estimation for this key information in network modeling.

Outline

I/Presentation of the graphon model
II/Construction of a complexity measure for graphons
III/Main lines of the estimation procedure

References

[1] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels : first steps. *Social Networks*, 5(2) :109–137, 1983.

[2] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *Rend. Mat. Appl. (7)*, 28(1) :33–61, 2008

[3] Laszlo Lovasz. *Large networks and graph limits*, volume 60 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2012.

[4] Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *Ann.Statist.*, 45(1) :316–354, 2017.

[5] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv :1509.08588*, 2015

[6] Chao Gao,Yu Lu,andHarrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6) :2624–2652, 2015

[7] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklos Z Racz. Testing for high-dimensional geometry in random graphs. *Random Structures and Algorithms*, 2016.

Approximate Message Passing algorithm for Compressed Sensing: an asymptotic analysis

[Talk submission]

Raphael Berthier

M2 student at Universite Paris-Sud

Abstract. In the last decade, Montanari et al. discovered a new iterative reconstuction algorithm for compressed sensing, named Approximate Message Passing, which exhibits good performances and low computational complexity in high dimensions. Its performance is tracked by a one-dimensional recursion, named state evolution. Our work is to justify rigourously the heuristic of this state evolution recursion in an asymptotic limit.

Keywords: approximate message passing, compressed sensing, denoising, lasso regression

1 Motivation

We consider the compressed sensing setting, where a unknown signal $x_0 \in \mathbb{R}^n$ is to be recovered from linear measurements $y = Ax_0 \in \mathbb{R}^m$, where $A \in \mathbb{R}^{m \times n}$ is called the measurement matrix. If the undersampling parameter $\delta = m/n$ is smaller than 1, this problem seems impossible, since the problem is underdetermined. However, it has been shown that if one makes a *sparsity* assumption on x_0 -i.e. x_0 has few non-zero entries-, and if A is well-chosen, then signal reconstruction is possible.

More precisely, the basis pursuit algorithm estimates

$$\hat{x_0} = \underset{x:Ax=y}{\arg\min} \|x\|_1 . \tag{1}$$

Taking the matrix A to be random, for instance a Gaussian or a Bernoulli matrix, it has been shown that with high probability, $\hat{x_0} = x_0$, provided that the signal is sufficiently sparse. As this minimization (1) is convex, its solution can be easily computed using linear programming methods. However, these methods are too computationaly expensive for very large applications.

Other methods have thus been proposed, one of them being *iterative thresholding* (IT), which provides a sequence of estimates using the following recusion:

$$x^{t+1} = \eta\left(x^t + A^T z^t, \lambda_t\right), \tag{2}$$

$$z^t = y - Ax^t. \tag{3}$$

Here, $\eta(., \lambda)$ is the soft-thresholding function, defined as $\eta(x, \lambda) = sign(x)(x - \lambda)_+$, and λ_t is a sequence of parameters to be determined. At each step, the algorithm does

a gradient step to reduce $\|y - Ax\|_2^2$, and then η forces the sparsity of the estimate. Although this is a natural algorithm, it suffers from slow convergence.

Montanari et al. proposed a very similar algorithm, that adds a term to the second equation:

$$x^{t+1} = \eta\left(x^t + A^T z^t, \lambda_t\right), \tag{4}$$

$$z^t = y - Ax^t + \frac{1}{\delta}\left\|x^t\right\| z^{t-1}. \tag{5}$$

Their inspiration came from approximating a message passing algorithm on a binary complete graph, where the variable nodes correspond to the coordinates of x and the factor nodes correspond to the coordinates of y. Thus they named this algorithm *approximate message passing* (AMP) [1].

While the derivation of this algorithm is rather complex, its asymptotic behaviour (as $n, m \to \infty$, δ fixed) is very simple. Suppose that, as $n \to \infty$, the empirical distribution of the coordinates of x_0 converges to a probability measure p_0 and $X_0 \sim p_0$. Define the state evolution iterates as

$$\tau_0^2 = \frac{1}{\delta}\mathbb{E}\left[X_0^2\right], \tag{6}$$

$$\tau_{t+1}^2 = \frac{1}{\delta}\mathbb{E}\left[(\eta(X_0 + \tau_t Z, \lambda_t) - X_0)^2\right], \tag{7}$$

where Z is a standard Gaussian random variable, independent of X_0.

It has been shown that for A Gaussian matrix, with i.i.d entries normal with mean 0 and variance $1/m$, then the empirical distribution of the coordinates of $x^{t+1} - x_0$, $\frac{1}{n}\sum_{i=1}^n \delta_{x_i^{t+1} - x_{0,i}}$, converges to the law of $\eta\left(X_0 + \tau_t Z, \lambda_t\right) - X_0$. As a consequence, the evolution of the estimator x^{t+1} is well-understood in a asymptotic manner : it is described by a one-dimensional recursion.

Simulations have shown that AMP is an efficient reconstruction algorithm, that the state evolution was a good approximation of reality for medium size applications, and that the state evolution description holds for a wide class of matrices.

In a recent paper [2], Maleki et al. have successfully used the AMP algorithm for different functions than the soft-thresholding function η. Their idea is that if we have another prior than sparsity on the signal x_0, then one can adapt AMP by changing the function η to force this prior. For instance, one can use any image denoiser to enforce a prior on image reconstuction. Another popular application is using an AMP algorithm to find a low rank structure in a noisy matrix [3]. The resulting algorithms are efficient, but we have little theoretical results for them.

2 Main contribution

I am currently working with a PhD student from Stanford, Phan Minh Nguyen, on the analysis of the AMP algorithm using a generic function η. Our goal is to prove that, for (almost) any function η, state evolution equations similar to those in (6), (7) hold. This will enable to develop a rigourous analysis of a wide class of algorithms in this area of research.

During the second time of my internship, we may carry an analysis of the AMP recursion for a particular denoiser η. Thanks to the state evolution equations, we hope to be able to show that this reconstruction method performs well on some class of signals. In that case, some simulation is likely to be carried out. However, things are not fixed yet and I cannot tell now what I will be able to present in September.

3 Conclusion

The AMP reconstruction algorithms come from non-rigourous intuitions, but show very good performance in practice and are very flexible to apply on a wide range of problems. State evolution equations make their analysis very simple. However, these methods are taught in very few classes ; this talk could be a good opportunity to give an introduction to them. I wouldn't plan on describing in detail the proof that I'm working on - it is not well-suited for a first introduction on the subject. I would rather like to give a clear presentation on the general ideas of AMP recursions, including some examples I hopefully will have worked on during the second part of the internship.

References

[1] Montanari. Graphical Model Concepts in Compressed Sensing. arXiv:1011.4328. 2010.
[2] C. Metzler A. Maleki R. Baraniuk. From Denoising to Compressed Sensing. arXiv:1406.4175v5. 2016.
[3] E. Romanov M. Gavish. Near optimal matrix recovery from random linear measurements. arXiv:1705.09958v1. 2017.

Automatic Machine Learning: benchmark and future work
[Talk submission]

Lisheng Sun

UPSud Paris-Saclay, Paris, France

Abstract. The success of Machine Learning applications in many fields relies heavily on well designed models / architectures, namely, selecting a good set of hyper-parameters. Hyper-parameter selection is not only time consuming, but also requires domain knowledge. My PhD thesis aims at developing a methodology to automate this process, and finally produces an "any data", "any time", and "any resource" black-box machine learning algorithm, which automatically examines the input data, selects an appropriate model with its hyper-parameter values, and outputs the desired type of predictions, all by taking into account of the available computational resources and a computational budget. We call this study "AutoML".

Keywords: automatic machine learning, hyper-parameter tuning, model design

1 Introduction and Current work

The AutoML problem is not new. For example, in 2016, Chalearn[1] has organized the Chalearn AutoML challenge[2]. This challenge provided 30 machine learning problems with different task types (binary / multi-class / multi-label classification, regression), diverse data type (image, text, speech, advertising, etc.), task specific objective functions and resource constraints, and asked participants to contribute algorithms that are able to solve all these problems without any human intervention.

As a first step of my PhD project, I am currently doing systematic analyses of the challenge results. This step is crucial for understanding the state-of-the-art solutions to the AutoML problem and identify the remaining difficulties.

The final winning solutions of the Chalearn AutoML challenge can be divided into two main types: 1) Bayesian Optimization techniques such as (autosklearn[3][3] by team 'aad_freiburg' and freeze-thaw Bayesian optimization[4] implemented by JR Lloyd [4], which build the posterior $p(model|data)$ by applying candidate models on the input data and use this posterior distribution to guide the search; 2) Heuristic solutions (abhishek[5]), which suggest a solution to a specific problem based on experiences acquired

[1] http://automl.chalearn.org/
[2] https://competitions.codalab.org/competitions/2321/
[3] https://automl.github.io/auto-sklearn/stable/
[4] https://github.com/jamesrobertlloyd/automl-phase-2
[5] http://blog.kaggle.com/2016/07/21/approaching-almost-any-machine-learning-problem-abhishek-thakur/

before with similar tasks. Most solutions limit their search space to learning machines implemented in the Python-based scikit-learn library[2].

We observed that, among the 30 datasets of the challenge (which are all very different in nature) a given algorithm can be good at solving some of them, but weak at solving others (Fig. 1). We already identified a few factors, which may cause this phenomenon, e.g. 1) the algorithm failed at the data ingestion of preprocessing level because the dataset contained missing values / imbalanced data or other difficulties that require sophisticated feature engineering before the actual learning; 2) the algorithm performed poor time management wasting time on training with unpromising hyper-parameter settings; 3) or conversely the hyper-parameter search space was not large enough. These observations suggest that the current AutoML algorithms are still far from being "any data + any resource + any time" machine learning solvers.

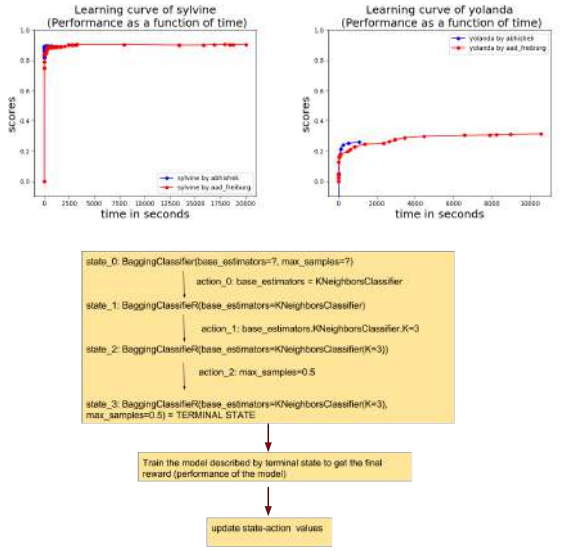


Fig. 1: Upper left and upper right: Performance as a function of time of 2 winning models (aad_freiburg and abhishek) applied on 2 different datasets. These 2 models solve very well 'sylvine', a binary classification task of forest cover type data; but have a poor performance on the other one 'yolanda', a regression task to predict songs' publication year. Bottom: Hyper-parameter selection process as a RL problem. It corresponds to an episode where the model under consideration is a bagging classifier and is built by performing actions in each state to assign values to its parameters. The agent receives reward (i.e. model's performance) only when it reaches the terminal state where the model gets trained on input data. The values of state-action pairs are then updated according to this reward.

2 Reinforcement learning: another possible solution to AutoML problem?

The hyper-parameter selection process can be formulated as a Reinforcement learning (RL) problem: states s are characterized by a sequence of tuples (hyper-parameter, value) that describes the model; in each state, available actions assign values to hyper-parameters; the model's performance plays the role of final reward. Surrogate performance approximations play the role of intermediate rewards. Fig. 1 illustrates a very simple hyper-parameter selection example as a value-based RL instance.

Recently, using RL for model design has started attracting the attention of the machine learning community. For example, B. Zoph et al. (2016) [5] and B. Baker et al. (2016) [1] have respectively used policy based and value based RL to automatically design neural networks to solve some deep learning benchmarks (CIFAR-10, PTB, etc.) and have achieved results comparable to human's best hyper-parameter tuning. However, the success of these efforts greatly relies on the use of huge computational resources, which severely limits their practical deployment in many domains. Therefore, reducing the computational demand will be one of the targets of my future work. I envision several workarounds: 1) smarter search strategy: early stopping can be used to guide the search toward more promising regions without waiting for the convergence of the current model; 2) smarter resource management: when given severe resource limitation, the algorithm must know to reduce its search space and / or training time for each candidate model to obtain an output before the resource limit is exhausted – this can be done, for example, by encoding the resource limit into the reward function.

In my presentation, I will first explain my analyses of the AutoML challenge, present various directions of research that they suggest, and my first explorations of RL algorithms applied to the AutoML problem.

References

1. Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. *arXiv preprint arXiv:1611.02167*, 2016.
2. Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, et al. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013.
3. Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*, pages 2962–2970, 2015.
4. K. Swersky, J. Snoek, and R. Prescott Adams. Freeze-Thaw Bayesian Optimization. *ArXiv e-prints*, June 2014.
5. Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

Deep specification and verification of SQL compilation chain

[Poster]

Léo Andrès, Raphaël Cornet, and Eunice Martins

LRI, Équipe VALS, Université de Paris Sud - Paris Saclay

Abstract. Nowadays, a variety of complex and critical systems manipulate sensitive data. Handling such data is an important issue, since such systems should guarantee its integrity and reliability. Within this context, query languages, such as SQL (Structured Query Language), are used by relational database management systems (RDBMS) as an important vehicule to process data. However, many questions spin around RDBMS concerning the safety guarantees they provide. It is in this respect that we propose to deeply specify, using the Coq proof assistant, the compilation chain of SQL.

Keywords: relational databases, SQL, query compilation, formal methods

1 Context and scientific positioning

1.1 Scientific goals and challenges

Current data-centric applications ranging from e-commerce, health crises' monitoring, to homeland security involve increasingly massive data volumes which are precious and whose availability, integrity and reliability is highly desirable. An important part of such data are handled by relational database management systems (RDBMS) through their query language, SQL, which is *the* standard for such systems. RDBMS, while intensively used in practice, have not yet reached the same high *safety* level guarantees as found in other critical systems, potentially yielding puzzling behaviours or even disastrous situations. Such a lack of strong assurance is problematic. Surprisingly, while formal methods are nowadays widely used to specify critical systems and to ensure that they comply with their specifications, such methods have been poorly promoted for data-centric systems [3, 2, 1].

In this work, we adopt such an approach for the SQL compilation chain, which is an important aspect to increase confidence in RDBMS. To do so, we mainly rely on the Coq interactive theorem prover, in combination with state-of-the-art RDBMS (namely PostgreSQL and Oracle) to provide efficient optimisations that will be checked back in Coq.

Our work is part of the Datacert project involved in the more general setting, while not funded by, the *NSF - Expedition in Computer Science - The Science of Deep Specification*.

1.2 Deep specification

How to specify in a clear and simple way the behaviour of complex and cumbersome systems? The question arises and, no rare, finding an answer is quite tricky. In fact, specifying a system represents a huge challenge. Trying to understand and to translate all the different behaviour patterns into a well-structured description is a rough task.

Unfortunately, despite being considered as an excellent practice in the scope of software development, it is also discarded by many who consider it as a hurdle and an unnecessary process. In truth, undertaking in these activities is far from effortless and straightforward, since such endeavour should result in a rich, accurate and complete portrayal of the system's conduct.

Therefore, a precious tool is the development of abstract interfaces, in which the multiple parts of a system are identified and separately specified. As outlined by the well-known 'divide and conquer' maneuver, widely used in algorithmics, it is wise to subdivide the multiple parts of one system (especially complex ones) and provide simpler and more focused details on each part at a time.

Besides, an interface should be clear but rich from the point of view of its completeness (in terms of the behaviour descriptions); two-sided (both from the implementation side and the final user); written with a formal notation and being able to support automated or machine-assisted tools. Concerning the later, during the last years, it is notable to watch the certification of such specifications gaining ground, by means of proof assistants such as Isabelle or Coq.

1.3 SQL's compilation in a nutshell

SQL compilation consists in four steps. The first two steps that include parsing and semantic analysis, translate the query into an algebraic expression. The last two steps also called the planning phase consist in logical and physical optimisation. The logical optimisation step exploits algebraic equivalences to perform sound query rewritings. The physical optimisation is in charge of producing query evaluation plans which are trees whose nodes are concrete, system-provided, implementations of algebraic operators. This last step is *data dependent* and is achieved based on auxiliary data structures and system maintained statistics.

Let us illustrate the compilation steps on an example. Assume that the following relations have been created:

```
create table movie
  (mid integer,
   title VARCHAR(90) not Null,
   year integer not Null,
   PRIMARY KEY(mid));
```

```
create table role
  (mid integer,
   name VARCHAR(70),
   PRIMARY KEY(mid, name),
   FOREIGN KEY (mid) REFERENCES movie);
```

Let us express the following query `select name from movie m, role r where m.mid = r.mid and year > 2000;` which has the following relational algebraic semantics:

$$\pi_{\texttt{name}}(\sigma_{\texttt{year}>2000}(\texttt{movie} \bowtie \texttt{role}))$$

At that point, any decent system such as PostgreSQL, Oracle etc. allows us through the statement `explain analyse` to see which is the plan chosen for any query. In this case the PostgreSQL plan proposed is:

```
            QUERY PLAN
-----------------------------------
 Hash Join
   Hash Cond: (r.mid = m.mid)
   ->  Seq Scan on role r
   ->  Hash
         ->  Seq Scan on movie m
               Filter: (year > 2000)
```

Given an algebraic operator, the underlying system provides several different algorithm implementations for it. For instance to the relational join correspond at least four such different algorithms: nested loop join, index nested loop join, sort-merge join and hash join (which is the one that has been chosen in this example).

Based on the work in [2, 1] our goal is to formally specify, using Coq, the different phases of SQL's compilation chain. In our work we follow a *skeptical* approach that consists in verifying in Coq that a physical plan given by a SGBD for a given query actually fulfills this query. Combined with the verified specification of the algorithms appearing in these plans (used during query execution) and [2], this work provides **the first fully certified SQL compilation chain**.

2 Contributions

Parsing and semantics analysis: a certified parser from SQL to its Coq representation We first implemented the parsing phase of the `select-from-where-groupby-having` fragment of SQL (with nested queries and aggregates). Then exploiting the results in [1] we obtained strong guarantees that the SQL query and its algebraic counterpart do have the same semantics.

Planning: designing a plan description language While SQL is standardised, plans issued by relational database systems are very ad-hoc as no standard specifies how to present them. Therefore, a preliminary task consisted in designing a plan description language able to host plans provided by systems such as PostgreSQL or Oracle.

Planning: Coq specification of access paths and join algorithms This task consisted in specifying the different main stream join algorithms: nested loop, sort-merge, index-based as well as specifying the so called *iterator interface*; and proving the correctness of the algorithms with respect to this specification.

Parsing PostgreSQL and Oracle plans to physical algebra We designed two parsers from PostgreSQL and Oracle plans to our Coq certified physical algebra.

Planning: formally relating a query plan with its semantics This task consisted in formally proving, using Coq, that any given physical query plan, issued by the system under consideration PostgreSQL, Oracle is correct with respect to the initial query.

References

1. V. Benzaken and É. Contejean. SQLCert: Coq mechanisation of SQL's compilation (formally reconciling SQL and (relational) algebra). Submitted for publication, 2016.
2. V. Benzaken, É. Contejean, and S. Dumbrava. A Coq Formalization of the Relational Data Model. In *23rd European Symposium on Programming (ESOP)*, 2014.
3. Gregory Malecha, Greg Morrisett, Avraham Shinnar, and Ryan Wisnesky. Toward a verified relational database management system. In *ACM Int. Conf. POPL*, 2010.

Breaking boundaries between language and database runtimes

Julien Lopez

LRI. Université Paris-Sud

Abstract. BOLDR is a modular framework that enables the evaluation of queries containing application logic (such as user-defined functions) in databases. BOLDR detects queries present in an application, translates them into an intermediate representation, rewrites them in order to make the most out of database optimizations, and converts the database(s) results back to the application. Experiments indicate that these techniques are applicable to real-world database applications, both in terms of successfully handling a variety of language-integrated queries, and in terms of providing performance benefits.

Keywords: language-integrated queries, databases, user-defined functions

1 Motivation/Introduction

Innovation in data analytics has been largely supported by the adoption of new programming languages that are best suited for statistical analysis, data mining, and manipulation of specific data formats. Built-in support for data analysis in data processing platforms cannot follow the pace of innovation sustained by the ecosystems of these languages. Therefore, it is crucial for databases to support user-provided data analysis methods.

To address this issue, database vendors have already started to equip their databases with support for new programming languages (Oracle R Enterprise [1]; Python in PostgreSQL's PL/Python [2], Amazon Redshift [3], Hive [4], and SPARK [5]; JavaScript in MongoDB [6] and Cassandra's CQL [7]; ...). These solutions share the common problem of exposing to the host language (language from which the query comes from) a low-level, ad-hoc API, that heavily limits the range of host language expressions that can be used in queries.

On the other hand, Microsoft's LINQ framework [8] proposes to extend application programming languages with built-in querying syntax and represent externally stored data in the data model of the application programming language. However, LINQ also suffers from the limitation that query expressions can only execute if the host language expression can be translated into the standard set of operators – which makes a dramatic negative impact on the expressivity and the performances of the approach.

As more programming languages gain support for in-database execution, it is now possible to consider a bolder approach: exploit the ability to execute host

expressions in the database. BOLDR is a language-integrated query framework that allows host language expressions and functions to occur in queries that are to be evaluated in the database. BOLDR translates queries in a host language into a Query Intermediate Representation (QIR) that might contain user-defined expressions of the host language, thus lifting the common limitation of other similar solutions.

2 The BOLDR framework

BOLDR is composed of three layers as shown in figure 1a: the runtimes (interpretors) of high-level, dynamic programming languages (R, Python, Ruby, Javascript, ...); the different database engines (relational databases such as PostgreSQL, HBase (table over Hadoop's HDFS), ...); and, in the middle, QIR acting as a universal language that subsumes the query languages of the different datastores (SQL, MapReduce, ...). QIR is a small functionnal language that supports records and sequencial data structures (both possibly nested), and that contains relational operators (projection, selection, join, ...).

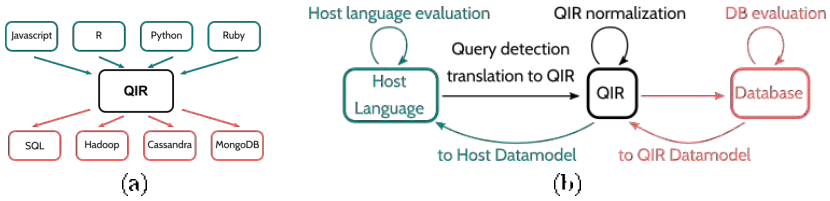


Fig. 1. The BOLDR framework

Figure 1b describes the evaluation of a host language program in BOLDR: the host language runtime evaluates the code that may include queries to databases. During evaluation, these queries are dynamically detected and translated into QIR terms. When one of these terms needs to be evaluated, it is normalized by a middle-layer optimizer and translated into a query for the target database (e.g. in SQL). This normalization ensures that the QIR term is translated as much as possible into an idiomatic query that will be optimized and executed by the database. The parts that cannot be translated are executed by the QIR runtime.

Here is an example in a JavaScript-like language as the host language:

```
1  function printPeopleFromCity (D, city) {
2    q = Filter(fun(p) {return p.city == city;}, Scan(D_PEOPLE));
3
4    while ((row = next(q)) != null) {
5      println(row.lastname + " " + row.firstname);
6    }
7  }
```

The query defined on line 2 is translated during evaluation as a QIR expression. When the results are accessed on line 4, the query is sent to the database, the results are fetched and translated into expressions of the host language, and the evaluation resumes.

3 Implementation and results

BOLDR consists of the QIR language, host languages, and target databases. The full stack has been implemented with the inclusion of R and SimpleLanguage (a subset of JavaScript) as host languages, and PostgreSQL and HBase as database back-ends. Development was done in Java and made use of the Truffle framework created for the implementation of programming languages by Oracle Labs. Truffle, in which languages such as Python, Ruby, R, and JavaScript have been implemented, is packaged as a simple jar file making the integration of Truffle languages to databases easy; thus allowing the execution of untranslatable code from a programming language inside a database: the query calls the runtime of the language as an external function.

BOLDR was tested notably using the TPC-H benchmark [9]. The results show that queries generated by BOLDR are in most cases as efficient as hand-written queries, and is better than PostgreSQL at inlining queries in some cases.

4 Discussion/Conclusion

BOLDR is a framework that allows programming languages to express complex queries that can contain application logic in the form of user-defined functions. These queries can then target any source of data, provided that this source is interfaced with the framework, and make the most of database optimizations. Future work includes the definition and implementation of a domain-specific language (DSL) to define the translation of QIR to a database representation, the implementation of more interfaces to languages and databases, an extension of QIR to more generic data operators, a type system for QIR using gradual typing, and optimizations of the execution of foreign code inside the database.

References

1. Oracle Corporation. Oracle R Enterprise. `http://www.oracle.com/technetwork/database/database-technologies/r`.
2. PL/Python - Python Procedural Language. `https://www.postgresql.org/docs/9.5/static/plpython.html`.
3. Python Language Support for UDFs. `http://docs.aws.amazon.com/redshift/latest/dg/udf-python-language-support.html`.
4. Hive Manual - MapReduce scripts. `https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Transform`.
5. Pyspark documentation - pyspark.sql.functions. `http://spark.apache.org/docs/1.6.2/api/python/pyspark.sql.html`.
6. MongoDB User Manual - Server-side JavaScript. `https://docs.mongodb.com/manual/core/server-side-javascript/`.
7. User Defined Functions in Cassandra 3.0. `http://www.planetcassandra.org/blog/user-defined-functions-in-cassandra-3-0/`.
8. Microsoft. LINQ (Language-Integrated Query). `https://msdn.microsoft.com/en-us/library/bb397926.aspx`.
9. TPC. The TPC-H benchmark. `http://www.tpc.org/tpch/`.

JOSE ESTRADA, ERIC GAUTHIER, Orange/C3S - Internship Tutors

Big Data Internship: Exploration & analysis of mobile communication's technical data - Bypass Fraud Detection

[Poster]

OUMOUSS EL MEHDI

Télécom Paristech, Paris-Saclay University

Abstract. Telecom operators like Orange suffer from huge revenue losses caused by fraud. As reported in (CFCA, 2011), these losses can reach millions of dollars per year. As a result, counter-measures are developed to detect fraudulent behavior as well as trying to avoid or minimize these losses and, if possible, help arresting fraudsters. However, taking in consideration the huge amount of data (Call Detail Records) to be processed as well as the constant change of fraudsters' behavior, traditional solutions fail to detect a big chunk of these fraudsters. This is due especially to the static nature of these solutions (rule-based systems). Therefore, Big data and Machine learning techniques could represent a feasible and attractive solution to explore.

Keywords: Telecom Fraud Detection, Bypass Fraud, Anomaly Detection, CDRs, Deep Learning

1 Motivation

Telecom fraud is a serious problem that affects all telecom operators. These companies usually are equipped with rule-based (Threshold) systems which have certain business rules that define abnormal traffic and then, basically, flag an anomaly whenever these rules are triggered. Such systems have as advantages being straightforward and easily adapted. However, since they are based on rules, fraudsters can soon or later bypass these rules by changing their behavior and consequently not being detected by the anti-fraud system. Therefore there is a constant search from the side of telecom companies for more efficient solutions capable of reducing the impact of such fraudsters.

Here we provide the big lines of the work we conducted as part of a PoC, during my internship at Orange, a company that is in full Digital transformation. Joining the $OLS/C3S$[1] entity with expertise in **fraud detection**, the goal of the PoC in development can be summarized as follows: Given a set of call detail records (CDRs), we want to construct a **deep learning based** unsupervised anomaly detection system. In simpler words, we want to be able to detect the anomalous activity in patterns of user's behavior. Specifically, the system is intended to assist human analysts by flagging potentially fraudulent activity for later review.

Emphasize will be put towards a certain type of fraud called Bypass Fraud. Also known as Simbox fraud, it is a type of telecom fraud that aims at hijacking and routing

[1] $OLS/C3S$: Orange Labs & Services / Country Support & Shared Services

international calls via VoIP, injecting them back into the destination's cellular network. Consequently, these international calls become local calls within the destination network, depriving operators from payments for calls routing and termination.

2 Anomaly detection

Fraud detection using unlabeled data comes under the umbrella of anomaly detection (Chio, 2015). Also known as outlier or novelty detection. The goal of an anomaly detection system is to identify those records or instances that are unusual, in the sense of being considered low probability by an anomaly detection system. Since we are in an unsupervised learning context, we need not know how these usage patterns differ, only that there is some statistical regularities that can be discovered or modelled by a machine learning system. More specifically, our goal is to model normal usage patterns, and not the fraudulent usage patterns.

3 Methodology

Taking as premises the fact that fraudulent behavior is different and smaller (as fraction of total users on the network) from legitimate users behavior. If we model the notion of normality (normal subscriber behavior), fraudulent behavior is then detected by finding deviations from this model of normality. This has the additional advantage that the learned network is not being tuned to a specific fraud type. Figure 1 shows the

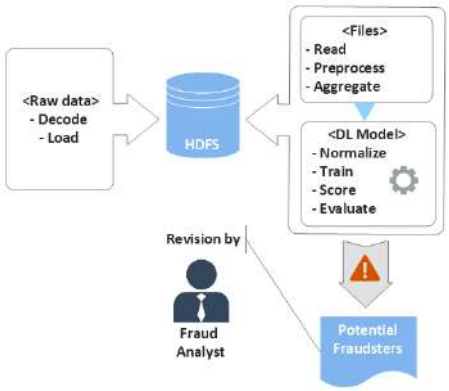


Fig. 1. General Framework

general framework to follow, where once files are stored at HDFS, the data exploitation process begins. CDRs are parsed, cleaned (using filters), then representative features are selected (e.g. aggregated data). Moreover, including additional features derived from other data sources help to better discriminate between normal and fraudulent activity. This is the reason why CDRs are often used in conjunction with other data in

order to improve results. Such features are for instance: customer data, demographic information of subscriber, type of account, current day (holiday, weekend, or special events), etc. Finally data normalization is conducted. All this should be done in well-defined pipeline to allow a better modelling of user behavior patterns.

Once our data is well prepared, neural network learning is conducted. There exist several network architectures (Veen, 2016; Dundar & all., 2015) for unsupervised anomaly detection methods. After further research, we found out that one deep learning architecture, seems to suit best the problematic we are dealing with and that is: **Auto-Encoders (AE)**. Auto encoders are neural nets trained with the goal to reconstruct their input. This is achieved, through an encoder that creates a dense representation of the input, followed by a decoder that tries to reconstruct the original input from the succinct representation. By training an AE, we end up with a model capable of easily reconstructing what is perceived as a normal behavior, which translates into a low reconstruction error, however, when we have a high reconstruction error, then we are facing a potential fraudulent behavior which should be flagged for further analysis.

Additional cleaning is necessary to filter out certain behavior patterns (white-list of known legitimate users), before passing on the network output to fraud analysts. The output at the end of the pipeline is a list of 'most anomalous' records, including relevant statistics for each record (e.g. scoring).

Finally, evaluation of the learned network is done by comparing the results of the network against the rule-based systems at disposal as well as the feedback from fraud analysts. As for technologies used for implementation, we used Apache Spark for data processing, DL4J as deep learning framework and Apache Oozie for job scheduling.

4 Discussion

Interestingly, not many telecom companies do use **real world machine learning based** anomaly detection systems as their core anti-fraud system, they are more of a complementary solution beside threshold systems. The reason behind such a choice has to do with several challenges that machine learning based system do face, namely:

- ML applications in an industry with high Intolerance to errors:
  - High rate of false negatives: the system is not working.
  - High rate of false positives: weak integrity (unusable system).
- Semantic gap : hard to interpret the results (alerts).
- Possible model-poisoning: it is hard to find clean training data.

References

CFCA. (2011). Cfcaś 2011 worldwide telecom fraud survey. Retrieved from `http://cfca.org/pdf/survey/Global%20Fraud_Loss_Survey2011.pdf`

Chio, C. (2015). *Detecting network intrusions with machine learning based anomaly detection techniques.* Retrieved from `https://www.youtube.com/watch?v=c71gt-I8Lik`

Dundar, & all. (2015). Convolutional clustering for unsupervised learning. *arXiv preprint arXiv:1511.06241*.

Veen, F. V. (2016). *The neural network zoo.* Retrieved from `http://www.asimovinstitute.org/neural-network-zoo/`

Modelizing transformation processes with ontology and probabilistic relational models [Poster]

Melanie MUNCH

UMR MIA-Paris, AgroParisTech, INRA, Universit Paris-Saclay, 75005, France

Abstract. Reasoning on a transformation processes requires a framework able to deal with a large amount of heterogeneous data and to model the uncertainty characterizing the biological processes. In this project, we propose a method based on Probabilistic Relational Models (oriented-object Bayesian Network) whose structure is learned from an ontology.

Keywords: modelization, probabilistic relational models, ontology

1 Motivation

A transformation process can be represented as a sequence of operations (or steps), receiving different inputs (such as conducts, mixtures, devices, ...) and designed to obtain a specific output (or product). To understand a transformation process is to understand the **relationship** between its different aspects: we need a characterization of the product at multiple scales (i.e. population, cellular and molecular) studied with different types of measurement (e.g. physiological, biochemical, genetic).

To organize these data and deal with their heterogeneity, an ontology dedicated to transformation processes has been designed by members of the LINK team at AgroParisTech and INRA [1]. This ontology, PO^2, gathers and standardizes experts' knowledge and information coming from different sources, acquired from different domains and at different scales. In this ontology, every step is defined as a class to which a set of descriptor classes is linked: **devices**, **mixtures** and **methods** are classes whose parameters are set by an operator; **observations** are classes whose parameters are measured during the step. Therefore each transformation process is represented by a succession of instances of steps and instances of their associated descriptors. Each step is linked to the one(s) following it according to the chronological order. However, a stabilization process is a dynamic process characterized by **uncertainty**, and despite its efficiency to structure the information, an ontology does not allow reasoning in face of uncertainty. A framework able to deal with this unpredictability while offering an insight of the links between the different observations is necessary.

The aim of this project is to propose and implement such a framework. We propose to combine the representative expression of ontologies with the reasoning possibilities of probabilistic relational models.

2 Our Approach

For a given domain, a Bayesian network (BN) requires a prespecified set of random variables, whose probabilistic relationships to each other has been fixed in advance. However it cannot be used in domains where the number of entities can vary, and cannot model relations between the different random variables. These limitations are a direct consequence of its lack of concept of *object*: while learning its structure or doing inference, we have to "flatten" every attribute, thus loosing their original structural links. For instance, in our case, a device can be defined by its brand and its capacity: if we learn a BN for it we would learn the probabilistic dependencies between these loosing the information that they all belong to the same "device object". Confronted to this problem, *inductive logic programming* can offer some good insight (using logical Horn rules); however it can only draw deterministic conclusions. That is why Probabilistic Relational Models (PRMs) offer a good alternative.

PRMs extend BNs with the concept of **class** connected in a relational structure. A class is a fragment of a BN over a set of inner attributes and a set of outer attributes from other classes referenced by so-called **reference slots** [2].

To face uncertainty in transformation processes, we propose to map the PRM's structure from the PO^2 ontology. In this way, every class in the ontology can be mapped automatically into a class in the PRM. This approach eases, as well, the learning of the PRM itself because it is learned starting from a given structure [3]. We based our work on [4], which proposed a way to pass from an ontology to a PRM in the transformation process domain and we propose an approach to learn a PRM mapped from the PO^2 ontology.

As explained before, a transformation process can be modeled by a sequence of steps with different parameters. In this article, we assume that the step at time t can be linked to one or multiple steps at time *t-1*. We use this dependency between objects to group every attributes with its associated step at time t, in order to create two classes for each step: one with every measured attributes (O_t) and one with every attributes set by the operator (C_t). We discard attributes with a ratio of missing values higher than a certain threshold, and attributes whose value never changes (for instance, if in a step the exact same device is used, the observation about its brand is useless). Then for each step we learn a Bayesian network, using the greedy Hill Climbing algorithm. This model groups attributes of O_t, attributes of C_t and attributes of $O_{t\text{-}1}$. In this way, we can visualize dependencies during time. However, to learn a model that reflects the logic of the process (for instance an observation in a step at time t cannot have an influence on a parameter fixed by the operator at time *t-1*), we have to set an order: attributes from $C_{t\text{-}1} \prec$ attributes from $C_t \prec$ attributes from O_t.

Once these BNs are learned for each step, we gather all of them in a PRM where each class is linked to the other following the structure given by the ontology. An example of PRM modeling a possible transformation process is shown in Figure 1.

3 Conclusion

We presented our algorithm for learning a PRM mapped from an existent ontology. We are now testing the approach on an artificial data set to compare its performance to an approach that learns PRM directly from data. The long terms objectives are the **discovery of new knowledge** in data (such as new links), and the **evolution of both the PRM and the ontology** following the integration of new data.

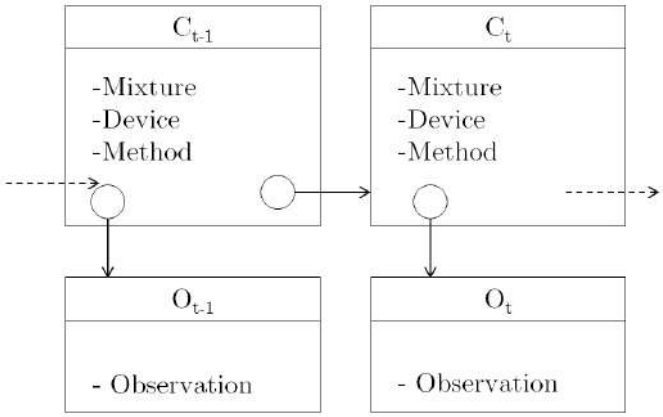


Fig. 1. Organization of two steps in a PRM. Each step is defined by two classes. The arrows represent the reference slots. For example, O_{t-1} has a reference to attributes of C_{t-1}, meaning that it has access to its attributes and values. However, O_t has no access to the attributes of O_{t-1}, according to the logic of the process.

References

1. L. Ibanescu, J. Dibie, S. Dervaux, E. Guichard, and J. Raad, "Po^2 - A process and observation ontology in food science. application to dairy gels," in *Metadata and Semantics Research - 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings* (E. Garoufallou, I. S. Coll, A. Stellato, and J. Greenberg, eds.), vol. 672 of *Communications in Computer and Information Science*, pp. 155–165, 2016.
2. L. Torti, P.-H. Wuillemin, and C. Gonzales, "Reinforcing the Object-Oriented Aspect of Probabilistic Relational Models," in *PGM 2010 - The Fifth European Workshop on Probabilistic Graphical Models*, (Helsinki, Finland), pp. 273–280, Sept. 2010.
3. N. Friedman, L. Getoor, D. Koller, and A. Pfeffer, "Learning probabilistic relational models," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages* (T. Dean, ed.), pp. 1300–1309, Morgan Kaufmann, 1999.
4. C. E. Manfredotti, C. Baudrit, J. Dibie-Barthélemy, and P. Wuillemin, "Mapping ontology with probabilistic relational models," in *KEOD 2015 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Volume 2, Lisbon, Portugal, November 12-14, 2015* (A. L. N. Fred, J. L. G. Dietz, D. Aveiro, K. Liu, and J. Filipe, eds.), pp. 171–178, SciTePress, 2015.

Nodes clustering in a graph under differential privacy constraints
[Poster submission]

Rafael PINOT *

Institut LIST, CEA, Université Paris-Saclay, F-91120, Palaiseau, France

Abstract. We investigate the problem of nodes clustering in a graph representation of a dataset under privacy constraints. Our contribution is twofold. First we formally define the concept of differential privacy for graphs and give an application setting where nodes clustering under privacy constraints allows for a secure analysis of the experiment. Then we propose a theoretically motivated method combining a sanitizing mechanism (such as Laplace or Gaussian mechanism) with a Minimum Spanning Tree (MST)-based clustering algorithm. It provides an accurate method for nodes clustering in a graph while keeping the sensitive information contained in the edges weights of the graph private. We provide some theoretical results on the robustness of the Kruskal minimum spanning tree construction for both of the sanitizing mechanisms. These results exhibit which conditions the graph's weights should respect in order to consider that the nodes form well separated clusters. The method has been experimentally evaluated on simulated data, and preliminary results show the good behavior of the algorithm while identifying well separated clusters. An extended experimental evaluation will be presented at the conference.

Keywords: Graph Clustering , Differential Privacy, Minimum Spanning Tree

1 Motivation/Introduction

Graphs represent a useful representation for many types of data, widely used in e.g. bioinformatics, network analysis, etc. More broadly, any dataset can be converted into a graph by using a well-chosen similarity matrix construction. In that respect, graph clustering [6] appears to be a key tool for understanding the underlying structure of many data sets by locating nodes groups ruled by a specific similarity.

A primary issue to be tackled while using machine learning techniques on a dataset is to protect the private characteristics of the individuals belonging to this dataset. Indeed, previous works have demonstrated that an adversarial use of a machine learning tool can lead to sensitive information release about the database used to train/construct such a tool. This might raise serious issues in medical applications for instance [3].

A widely adopted definition of what a "good" privacy condition should be, called *differential privacy*, has been introduced and theoretically studied in [2]. Since this seminal work [1], several models respecting the differential privacy conditions have

been proposed (e.g. [4]), but most of them do not consider structured types of data such as graphs. Therefore, how to consider differential privacy on structured data types remains an open question. Should one keep private the nodes, the edges and/or the weights of the graph. Those choices are mainly made according to the nature of the data at hand. This work aims at proposing a principled formal framework for privacy-perserving in learning from graph-structured data. The overall goal is to pave the way for applications such as genomics, proteomics, etc. where privacy-preserving is not an option but a strong requirement. In the sequel, after formalizing what one should understand by "privacy-preserving" in such a framework, we will provide a simple and accurate way of studying the structure of the graph by using an MST-based algorithm for graph clustering under differential privacy constraint.

2 Preliminary results

Let us consider the following scenario : an analyser is studying a set of genes $(g_i)_{i\in[n]}$, and she wants to observe the map of those genes interactions on a given population $P = (x_i)_{i\in[m]}$. The gene expression is described as follows: let x be an individual from P then if x has the gene g_i, $g_i^x = 1$ otherwise $g_i^x = 0$ (P can as well represents a family of vectors from $\{0,1\}^n$). For the purpose of this setting, we consider $m >> n$.

According to a population P, a gene-gene interaction graph can be built such as:

$$G = (V, E) \text{ with } V = \{g_i, i \in [n]\} \text{ and } E = \{(g_i, g_j)\ s.t\ i \neq j, \exists x \in P\ s.t\ g_i^x = g_j^x = 1\}$$

with a weight function $w : E \to \mathbb{R}$ such that

$$\forall (g_i, g_j) \in E, w((g_i, g_j)) = \frac{I \times \#\{x \in P\ s.t\ g_i^x = g_j^x = 1\}}{n}$$

where I is a real parameter that allows the analyser to scale the weights.

With such a setting, the sensitive information about an individual is its genes characteristics, i.e. the binary vector representing the individual. To construct a map while respecting *differential privacy* on the genes characteristics of our individuals, one needs to ensure *differential privacy* on the graph weights (i.e on $w(e), \forall e \in E$). The definition of differential privacy [1] mostly relies on a parameter ϵ, denoted privacy degree, that expresses the degree at which one wants to protect the sensitive information in the dataset. Ensuring privacy comes at the price of loosing in accuracy, hence a tradeoff to achieve [2]. Therefore the privacy degree will intervene in the result we present in the following of this work.

The graph weights privacy has been defined recently in [7]. Inspired by this work we prove that if one construct K sets of edges $(E_1, ..., E_K)$ such that:

$$i, j \in [K], i < j \implies \forall e \in E_i, e' \in E_j, w(e) < w(e') \tag{1}$$

Then Eq. (1) hold after adding an i.i.d Laplace noise (similar result for Gaussian noise can be obtained) on the graph's weights with probability greater than $1 - \exp(\epsilon t)(\frac{1}{2} + \frac{t\epsilon}{4})(K-1)$. With $t = \min_{i\in[K-1]}\{\min_{e\in E_i, e'\in E_{i+1}}\{w(e') - w(e)\}\}$, w representing the weights before the noise's addition and ϵ the privacy degree. Thus one can find conditions on the graph such that Kruskal algorithm for building the MST is robust to a sanitizing mechanism (Laplace or Gaussian [2, Chapter 3]).

The MST is known to help recognizing clusters with arbitrary shapes in MST-based clustering algorithms and thus can be used for wider applications than community

detection. It is based on the idea that the structure of a dataset is well represented by its MST. Xu et al. claimed [8] that given a reasonable definition of a cluster, *"if one takes two points c_1, c_2 of a cluster C, then all data points in the tree path connecting c_1 and c_2 in the MST must be in C"*. Due to space limitation, we could not develop this in this abstract but we have rigorously proved this statement, and employed it as a motivation for the use of an MST-based algorithm. Zhou et al. introduced a MST-based graph clustering called MSDR [9] that we enhanced to maintain good performances while ensuring privacy, based on the conditions we found on the Laplace and Gaussian Mechanisms. The preliminary results obtained on the simulated datasets are encouraging and we are looking for good ways of optimizing its parameters.

Future Directions

The continuation of this work will be dedicated to several issues: First, our setting could lead, in certain circumstances to another type of privacy issue: *edge privacy*, this kind of privacy breach is important as well and could be tackled by using some kind of "Threshold-release" [2, Chapter 3] techniques on the graph's edges, for example.

One can also remark that the mechanism we use to ensure *differential privacy* is efficient and widely used, but they may not be really optimal for our setting, therefore we have to look for techniques that will ensure privacy with a mechanism that, by being more specific to our problem, might help us keeping a better accuracy. We will for example consider graph sketching and its incidence on privacy preservation [5] .

References

1. Dwork, McSherry, Nissim, Smith: Calibrating noise to sensitivity in private data analysis pp. 265–284 (2006)
2. Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy, vol. 9. Now Publishers (2013)
3. Fredrikson, Lantz, Jha, Lin, Page, Ristenpart: Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: 23rd USENIX Security Symposium. pp. 17–32. USENIX Association, San Diego, CA (2014)
4. Ji, C.Lipton, Elkan: Differential privacy and machine learning: a survey and review. Cornell University Librairy (2014)
5. Mishra, N., Sandler, M.: Privacy via pseudorandom sketches. In: Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS 06. ACM Press (2006)
6. Schaeffer, S.E.: Graph clustering. Computer Science Review 1(1), 27–64 (aug 2007)
7. Sealfon: Shortest paths and distances with differential privacy. In: Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - PODS. ACM Press (2016)
8. Y.Xu, V.Olman, D.Xu: Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. Bioinformatics 18(4), 536–545 (apr 2002)
9. Y.Zhou, O.Grygorash, T.F.Hain: Clustering with minimum spanning tree. International Journal on Artificial Intelligence Tools 20(01), 139–177 (feb 2011)

*Under Supervision of Anne MORVAN, Florian YGER, Cédric GOUY-PAILLER and Jamal ATIF

Segmentation and Detection in large microscopy for neural development and organization
[Poster submission]

Tania-Marina Bacoyannis, Anatole Chessel, E. Beaurepaire, J. Livet, L. Abdeladim

Laboratory of Optics and Biosciences, Ecole Polytechnique Institut de la Vision, Paris

Abstract. Fluorescence microscopy, through 30 years of research and 2 Nobel prizes, has changed life sciences by allowing us to observe biological processes in vivo. The large compilation of fluorescence based techniques across several conditions, and their study via the latest techniques of image processing, machine learning and data science, will in turn lead us to new insights into complex biological systems. In particular in neuroscience, the 'Brainbow' revolution in labelling allow for the mapping of neuronal circuits and the tracking of neural cell fate and lineages. In this work, we seek to advance toward this goal through the development of detection and annotation machine learning pipelines for large scale fluorescence microscopy images.

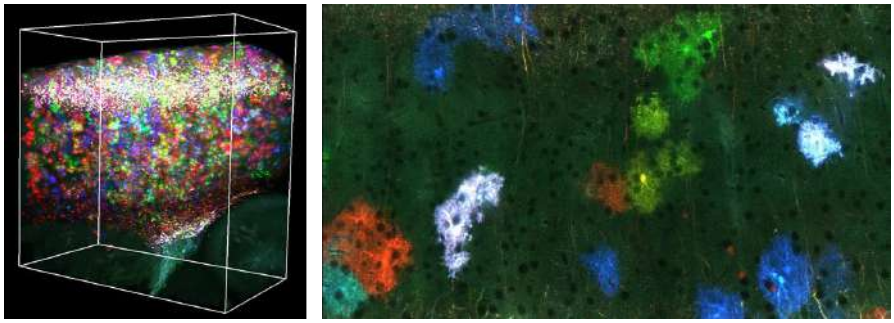
Keywords: Bioimage informatics; Neurobiological data; Fluorescence Microscopy; Deep Learning

1 Introduction

How is the complex neural tissue built from neural progenitors? How do neural progenitors share the genesis of neuronal and glial cells? How is their clonal descent organized in mature neural tissue? These questions are fundamental in neurobiological researches in order to understand a brain's function, neural structure and development. To be able to solve those problems, we need large scale images over cubic millimeter-volumes of cortex tissues with sub-cellular resolution at different stages of development. The Laboratory for Optics and Biosciences (LOB) and the Institut de la Vision (IDV) are collaborating in order to make it possible. Within the LOB, an innovative large volume quantitative imaging based on blockface multiphoton multicolor fluorescence microscopy has recently been developed (article in preparation, see Fig. 1). This new development allows multiple channels, 3D imaging over cubic millimeter-volumes of biological tissues with cubic micrometer-resolution. Combined with novel methods for multicolor fluorescence tagging brainbow developed at IDV, these imaging approaches open very novel possibilities for large scale quantitative studies of brain connectivity and development.

These images routinely weight several hundreds of Giga bites in up to five dimensions (3D plus time and colors) and data analysis is the limiting step to truly take advantage of those datasets for neurobiological projects [1]. With this goal in mind, it will be necessary

to develop specialized methods and software for data management, processing and analysis. The aim of my Master Thesis is to develop a framework for detection and annotation in these large scale fluorescence microscopy images, of two different types of essential neural cells: the astrocytes and the neurons.



Randomly labelled astrocytes imaged by the large-volume multicolor multi-photon microscope developed at LOB in 2015- 2017. Left: 3D rendering of a 1.2 × 2.4 × 2 mm3 volume imaged with 0.4 × 0.4 × 1.5 µm3 sampling; Right: detail of one plane illustrating the subcellular resolution. Each colored group is a clone of astrocytes originating from the same progenitor cell. Confidential unpublished data from L Abdeladim PhD work, collab LOB/IDV; see also [3,4].

2 Methods and Results

Due to their complexity and the fact that they are large scales images, it is not possible to annotate manually datasets, as biologist used to do. In this context, we need to develop an automatic method to annotated datasets. The aim is to automatically annotate and detect our structures of interest (astrocytes and neurons) in a whole large scale microscopy image.

First we develop a process in order to do classification on single tile of well-defined size. We used as training set a manually annotated dataset composed equally of Neurons and Astrocytes. After having extract specifics volumes containing the annotated structure of interest from the whole microscopy raw image, we convert them from RGB to Grayscale space. The initial process developed for classification is based on WND-Charm Method [2]. Specifically, 2919 features were extracted for each specific volume and Principal Component Analysis was chosen to reduce the dimensionality. We applied Random Forest as classification method. Parallelization across tiles is used to speed computation.

The model we developed is relatively efficient, and is able to classify with a sensitivity of 95.8

We are currently trying to use this classifier to a whole image for detection. A possible process will involve sliding a window across the volume to perform point-wise detection.

3 Discussion/Conclusion

The developed process proved his ability to classify astrocytes and neurons. However the classifier has not yet been tested for automatic annotation and detection in a whole unseen image, we are not able to prove is efficacy in the generic case yet. Expected issues include the size of the whole dataset, which may lead us to develop a more multi-scale approach, the instrumental artifacts and biases which need to be modeled or learnt, and the complexity of the whole tissue. Further work include adapting Deep Learning algorithms for segmentation and detection and test them, as they are the current state of the art and should be easier to use and more versatile.

References

[1] Chessel, A. (2017). An Overview of data science uses in bioimage informatics. Methods, 115, 110118.

[2] Orlov N, Shamir L, Macura T, Johnston J, Eckley DM, Goldberg IG. WNDCHARM: Multi-purpose image classification using compound image transforms. Pattern recognition letters. 2008;29(11):1684-1693.

[3] Loulier, K., Beaurepaire, E. Livet, J. (2014). Multiplex cell and lineage tracking with combinatorial labels. Neuron, 81(3), 50520.

[4] Mahou, P.,Beaurepaire, E. (2012). Multicolor two-photon tissue imaging by wavelength mixing. Nature Methods, 9(8), 815818.

Streaming Comparison Benchmark between Spark and Flink using Kafka
[Extended Abstract]

Carlos Perez

Paris-Saclay University

Abstract. Deduplication problem is an interesting real life problem because there are situations when companies need to find duplicates of data in order to reduce the burden of their analytical processing. Given the increasing amount of data available to companies and the potential of the internet of things as an explosion on the quantity and diversity of data solutions such as streaming are interesting to explore.
In this work I present a benchmark to compare the deduplication algorithm using Streaming libraries for Spark and Flink. Publisher subscriber system Kafka is used as a tool to feed streaming data to both deduplication engines. The measurements obtained between the frameworks will also be compared with their batch versions and the results will be analyzed.

Keywords: streaming, kafka, spark, flink

1 Motivation

Deduplication problem is an interesting real life problem because there are situations when companies need to find duplicates of data in order to reduce the burden of their analytical processing.

A batch version of this algorithm has alredy been developed in python. This version uses data stored in HIVE tables that has been preprocessed to clean it. This algorithm uses a set of rules to decide if two rows are duplicated or not. Some of this algorithms are Jaccard similarity measure, Jaro-Winkler similarity masure or Hamming distance. These algorithms are applied to certain row fields chosen by the user. Each pair ¡comparison-algorithm¿ has a weight and emits a score. If the sum of all the scores for two rows surpasses a certain threshold also user defined the records will be considered as duplicated.

The objective of this algorithm is to compare the records of two datasets of sizes m and n to find matches. The total number of comparisons thus would be M x N. To increase the performance of this algorithm a block concept was introduced. Each row was assigned to a specific block. Each block is defined manually by the final user in a configuration file as a substring of a column. For example if we have a client's table with a column name a block of name (0,2) and three records "Deborah" , "Matthew" and "Mark". Deborah would be assigned to block 'DE' and "Matthew" and "Mark" would be assigned to block 'MA'.

In order to explore a more scalable solution and prepare for future challenges of the internet of things a streaming solution of the deduplication algorithm will be coded. Aditionally to explore more ideas Kafka a publisher-subscriber system will be used to feed the data of the algorithms.

The motivation of this paper is to measure performance of the deduplication algorithm using different frameworks and languages. We expect to discover substantial differences in performance times that allow to choose one solution over the others and analywe the main causes of this differences.

2 Benchmark construction

One of the first steps of this process was to translate the python version of the deduplication algorithm to a java batch version. The choice of language was made because of the maintability aadvantages and the wide experience in the industry.

Once a batch version of this qlgorithm is codeed in JAVA dnd tested an streaming version will be created. At the moment of this paper only a functional version for spark streaming is developed. Batch processing can be seen as a special case of streaming processing when all the data is sent in the same window. To adapt the existing batch algorithm to the streaming the structure of the different blocks is kept in memory . For the excercise of benchmarking it was not necessary to define a time window to eliminate old records. However to avoid degradation of performance in a real life streaming scenario a search server such as solr could be added to the solution to store and query the duplicates.

3 Benchmark testing

Once this algorithm is implemented for Spark and Flink time measurements will be made with datasets of 100, 1000, 10000 and 1,000,000 rows to see the progression in execution time.

Also for an incoming row they all have to be preprocessed to assign them to their corresponding blocks and matched against the blocks kept in memory.

The experiments will be carried out on a six node cluster with 64 cores and 128GB RAM per node. Kafka version 2.12 was used. In kafka each engine will have an assigned topic "spark" and "flink". Three different brokers will be defined and deployed in a single machine and two topics "spark" and "flink" will be created created. The dataset to feed the algorithm is a file with 1.000.000 of real client data and will be ingested into Kafka using a console producer. The dedupliation engines will hqve windows of 1 second.

Spark and Flink engines will consume the data from three Kafka brokers and the specific topic assigned to each one of them. After finding the duplicates the engines will persist them directly in Hive in as they are produced.

The choice of the performance measurement tools and methodologies has to be made carefully to give a comparable measure of the different solutions

4 Main contribution

My main contribution is going to be to recode the deduplication algorithm in JAVA for Spark and Flink Frameworks. Using batch and streaming libraries. Aditionnally

the implementation with Kafka and the connection to Spark and Flink to Stream the data ito the engines. Once the development is finished I will measure the execution of times with the same dataset of about 1.000.000 records and I will produce charts to compare these frameworks. I will create charts comparing and execution times, in particular ram memory usage and network consumption.

5 Conclusion

Because of time the results are not conclusive. The deduplication algorithm in Spark has already been coded. The connection between Kafka Spark and Hive has been created and the flow. The deduplication algorithm in Flink is still under development.

References

[ref1] Wang, G., Koshy, J., Subramanian, S., Paramasivam, K., Zadeh, M., Narkhede, N., ... Stein, J. (2015). Building a replicated logging system with Apache Kafka. Proceedings of the VLDB Endowment, 8(12), 1654-1655.
[ref2] Cordova, P. (2015). Analysis of Real Time Stream Processing Systems Considering Latency. University of Toronto patricio@ cs. toronto. edu.
[ref3] Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., Tzoumas, K. (2015). Apache flink: Stream and batch processing in a single engine. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 36(4).
[ref4] Cunningham, D., Subrahmanian, E., Westerberg, A. (2010). User-centered evolutionary software development using Python and Java. In Proceedings of the 6th International Python Conference, http://www. python. org/workshops/1997 (Vol. 10).

Exploring Data Mining Techniques for Opportunistic Mobile Sensing Making Sense of Spatial Multivariate Time Series

Talk submission

Ahmad Mustapha, Yehia Taher, Karine Zeitouni

University of Versailles, DAVID Laboratory

Abstract. With the rise of mobile sensors as low-cost and light devices. Air Monitoring community is utilizing them to shift to new monitoring paradigms. Opportunistic Air Monitoring is gaining mainstream facing new challenges in time series analysis domain. The nomadic nature of sensors requires new data mining techniques in order to exploit the time series on multiple dimensions and different granularity. In this Paper we focus on tackling these raised problems by studying potential solutions. Our proposals will be implemented, evaluated, and applied to a real scenario of opportunistic air quality monitoring.

Keywords: Data Mining, Mobile Sensors, Multivariate Time Series, Opportunistic Mobile Sensing

1 Context and Motivation

Upon the recent development of advanced computing and communication technologies, the world is witnessing the rise of the so-called Internet of Things (IoT). IoT envisions a world where everything is connected - from humans and computing devices to animals, vehicles, and even the smallest appliances. Sensors and actuators are fetched on things enabling them to sense, generate data, communicate, act, and to share information. This is leading to the generation of massive amount of data, now regarded as Big Data or Big Sensing Data in the IoT context. With great embedded potential in this data, both industry and academia are rushing to develop techniques and technologies that not only can handle this large amount of data but can also exploit them in order to mine new knowledge and insights.

Time Series mining or analysis techniques, considered as a specific field in Data Mining, are being used to act on and exploit sensors data. A Time Series is a sequence of time stamped events. It can model sensors data streams where simple or combined observations (temperature, humidity, location, etc.) are being continuously fed along with their timestamp to specialized servers. The aforementioned techniques act on these time series and perform conventional and extended data mining and analysis operations, such as - but not limited to - classification, early classification, clustering, and forecasting.

One application of IoT is monitoring air pollution. Several research initiatives have used fixed air pollution sensors to monitor air quality [5]. However fixed sensors have

been facing shortcomings in modeling air quality because of the high spatiotemporal variability nature of air pollutants. That is why the community is shifting toward new monitoring paradigms which utilize mobile sensors. These new paradigms are enabled with the rise of low-cost and lightweight air pollution sensors. Participatory air quality monitoring is one paradigm where structured monitoring campaigns are being held [1]. Sensors are fetched on pedestrians, cyclists, or on vehicles and specific routes and areas are targeted for monitoring. Another paradigm is opportunistic air quality monitoring. Unlike the previous one, this paradigm doesnt target specific routes or areas; it takes advantage of existing mobile infrastructure or people common daily routines to perform monitoring [4]. These new paradigms have opened the door for new possibilities and challenges.
Opportunistic air quality monitoring has several advantages compared to conventional monitoring techniques. First, it has a lot wider spatiotemporal coverage. Second, it enables insights with high resolutions with a granularity reaching to street level. Third, it promotes personalization where each individual will be able to gain insights related to his exposure to pollutants rather than aggregated ones. Fourth, it measures indoor and outdoor environments (Home, Work, Transportation, Streets, Parks, etc.). This combination of benefits enables air pollution profiling. Extracting and mining this type of profiles represents the motivation for our work as it induces some challenges in the context of mobile sensors.

2 Objectives and Challenges

The nomadic nature of sensors, and their combination (the campaign uses a multi-sensor device) lead to revisiting the traditional methods of data mining and knowledge extraction. Indeed, these sensors produce a multivariate time series where one variable is the geographical position of the device (we call it space multivariate time series). The spatial dimension is essential in mobile sensing, and raises new challenges. The data should be analyzed in multiple dimensionality and granularity, including the spatial dimension and its various scales. Extracted spatial data needs to be aggregated on multiple levels ranging from micro-environment to macro-environment scale (Metro, Subway, Street). Temporal data also must be analyzed to reveal seasonable patterns and trends for specific individual classes or places. However, the best way to process such type of data is by using comprehensive models like multidimensional databases or knowledge bases. Here comes one of the challenges on how to transit from raw and heterogeneous time series data into such a type of models.
Moreover going further in exploiting the personalization aspect of opportunistic mobile sensing enables individuals to relate air pollution to themselves [3], and to act upon gained insights. For example an individual may change his daily routes, his transportation means, even his activities or diet in sake of lesser exposure and lesser health effects. Nonetheless, this requires building individual profiles, comparing them, and correlating these profiles with other data like personal health and activities. This correlation opens the way for highlighting potential relations of causality. .

3 Methodology

To achieve the aforementioned, data mining techniques should be utilized at different stages. This can be enumerated as following:

- Data Uncertainty: Ranging from the calibration phase of sensors to the adjustment of asynchronous data as well as spatial data fusion.
- Data Enrichment: To cope with multidimensional and granular analysis requirement, methods to transit from raw time series into more rich and abstract data should be researched and developed. For example, using Intelligent Data Analysis (IDAI) approaches - which fill the gap between data generation and data comprehension are possible candidates [2].
- Data Correlation: The enriched data opens the way to the discovery of hidden correlations of synchronous or asynchronous phenomena, and to the discovery of potential causality. Functional Data Analysis (FDA) is a possible candidate approach.

4 Conclusion

In this master internship, we aim at developing data mining methods adapted to types of databases including geodated series with associated context on the one hand, and to study potential solutions while detailing different types of analysis in different dimensions one the other hand. Our proposals will be implemented and evaluated in a perspective of large-scale collection of spatial multivariate time series, and applied to a real scenario of opportunistic air quality monitoring.

References

1. Opensense ii. http://opensense.epfl.ch/wiki/index.php/OpenSense_2, accessed June 1, 2017
2. Roda, F., Musulin, E.: Expert Systems with Applications An ontology-based framework to support intelligent data analysis of sensor measurements. EXPERT SYSTEMS WITH APPLICATIONS (2014)
3. Sîrbu, A., Becker, M., Caminiti, S., De Baets, B., Elen, B., Francis, L., Gravino, P., Hotho, A., Ingarra, S., Loreto, V., Molino, A., Mueller, J., Peters, J., Ricchiuti, F., Saracino, F., Servedio, V.D.P., Stumme, G., Theunis, J., Tria, F., Van Den Bossche, J.: Participatory Patterns in an International Air Quality Monitoring Initiative (2015)
4. Van Den Bossche, J., Theunis, J., Elen, B., Peters, J., Botteldooren, D., De Baets, B.: Opportunistic mobile air pollution monitoring: A case study with city wardens in Antwerp (2016)
5. Zheng, Y., Liu, F., Hsieh, H.P.: U-air: When urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2013)

Multi-omics data integration to model iron metabolism in pathogenic yeast species

Poster

Thomas Denecker[1] and Gaëlle Lelandais[1]

[1] Institut de Biologie Intégrative de la Cellule - Orsay

Abstract :

In the case of fungal infections in humans, access to iron resources is a critical element for the relationship between host and pathogens. *Candida* pathogenic yeasts have developed, during their evolution, original strategies for capturing the host's iron and adapting their metabolism to living conditions in low iron environments. This project aims at *in silico* modeling of iron metabolism from "multi-omics" experimental data (genomics, transcriptomics, proteomics). Different *Candida* species will be studied in order to identify their specificity of action, with regard to their modes of infections.

Keywords : Data analysis, Pathogenic Yeasts Species, Omics, Systems Biology

1- Motivation/Introduction

Candida yeast species are responsible for human fungal infections named candidiasis. These infections include cutaneous lesions on different body locations, such as the internal face of the cheeks and the tongue (oral candidiasis), the body folds (cutaneous candidiasis) or the vulvo-vaginal region (genital candidiasis). If candidiasis is usually punctual over time, it can become invasive, mainly in people with severe immunity deficitis [1]. Finally, the use of invasive medical techniques such as the implantation of catheters, is directly related to serious fungal infections hospital [1]. During infections, pathogenic yeasts must adapt their metabolism to very different environments. The case of metals such as iron, is a perfect illustration [2]. As a commensal organism in the intestinal flora, Candida albicans is adapted to an environment where iron is available, whereas in the case of infection, the same yeast species can survive in blood circulation and epithelial tissues, where iron resources are extremely limited [3]. Access to iron resources is thus a critical element in the relationship between host and pathogens [4]. Iron is essential for multiple cellular processes, and hence an effective defence mechanism of host against pathogenic microorganisms is to limit iron access to pathogens [4]. To survive, pathogens have developed original strategies to 1) capture the iron host (using siderophores for example), and 2) adapt their metabolism to life conditions with very low iron concentrations [3].

Over the past two decades, the use of high-throughput (or "omics") experimental techniques has led to a better understanding of the metabolism and iron homeostasis in microorganisms. Initially restricted to the model yeast *Saccharomyces cerevisiae* (for instance [5]), more recent studies were performed in Candida species: *Candida albicans* (for instance [3]) and *Candida glabrata* (for instance [6]). To date, more than twenty scientific

publications present high-throughput experimental data related to the use of iron by pathogenic yeasts. The diversity of the experimental strategies used in our works (genomics, transcriptomics, proteomics or metabolomics) represent a great opportunity to explore the cellular processes at different levels of observations

2- Selection of a list of genes of interest that are good candidate to be involved in iron metabolism in *C. glabrata.*

Two strategies have been put in place:

- Strategy 1: Study the genes of *C. glabrata* that have an orthologous for genes known for their involvement in iron metabolism in yeast models (*C. albicans* and *S. cerevisiae*) Approach with *a priori*
- Strategy 2: Cross-checking lists of genes between different multiomics experiments to achieve to a list of genes of interest verified manually
- Inference of a network to lead to a descriptive modeling of iron metabolism

3- Inference of a network to lead to a descriptive modeling of iron metabolism in *C. glabrata*

Two approaches of network inferences were performed:

- An inference based on bibliographic data: the links between the genes selected in strategy 1 are studied in the literature (for instance [5]) and allowed us to describe a first network
- An inference of co-expression graph: the genes selected in strategy 2 are grouped and linked according of their co-expression between the different experiments studied. A study of these links then carried out.

4- Discussion/Conclusion

A first descriptive model of iron metabolism was performed in *C. glabrata*. It was carried out using an a priori approach using the Yeast models (strategy 1). This model will be further described using the strategy 2 approach with only *C. glabrata* data. Finally, to realize the final model, we will realize a causal inference between the genes.

References

[1] Arias, S. et al. ***Epidemiology and mortality of candidemia both related and unrelated to the central venous catheter: a retrospective cohort study***. Eur. J. Clin. Microbiol. Infect. Dis. 1–7 (2016). doi:10.1007/s10096-016-2825-3

[2] Ding, C., Festa, R. A., Sun, T. S. & Wang, Z. Y. ***Iron and copper as virulence modulators in human fungal pathogens.*** Mol. Microbiol. 93, 10–23 (2014).

[3] Chen, C., Pande, K., French, S. D., Tuch, B. B. & Noble, S. M. ***An iron homeostasis regulatory circuit with reciprocal roles in Candida albicans commensalism and pathogenesis.*** Cell Host Microbe 10, 118–35 (2011)

[4] Sutak, R., Lesuisse, E., Tachezy, J. & Richardson, D. R. ***Crusade for iron: iron uptake in unicellular eukaryotes and its significance for virulence***. Trends Microbiol. 16, 261–8 (2008)

[5] Blaiseau, P.-L., Seguin, A., Camadro, J.-M. & Lesuisse, E. in (2010)

[6] Gerwien, F. et al. ***A Novel Hybrid Iron Regulation Network Combines Features from Pathogenic and Nonpathogenic Yeasts***. MBio, (2016).

Beyond stochastic gradient for maximum likelihood based ICA on EEG and MEG [Poster]

J. Montoya[†], P. Ablin[‡], J-F. Cardoso[*], A.Gramfort[‡]

† : Telecom Paristech, ‡ : Inria, Parietal team, ∗: Institut d'Astrophysique de Paris

Abstract. Independent Component Analysis (ICA) is a technique for unsupervised data exploration widely used in neuroscience. Linear ICA aims at discovering statistically independent sources from multivariate observations. It is a probabilistic generative model for which inference is classically done by maximum likelihood estimation, which leads to a smooth non-convex optimization problem. The gradient is available in closed form so first order gradient methods are often employed despite a slow convergence such as in the Infomax algorithm. While the Hessian is known analytically, the cost of its computation and inversion makes Newton method unpractical for a large number of sources. We show how sparse and positive approximations of the true Hessian can be used to precondition the L-BFGS algorithm. Results on EEG data demonstrate that the proposed technique leads to convergence that can be orders of magnitude faster than algorithms commonly used today.

Keywords: Independent Component Analysis, L-BFGS, Preconditioning

1 Motivation/Introduction

Independent Component Analysis (ICA) is a multivariate data exploration tool massively used in neuroscience [1]. The underlying assumption of linear ICA is that the data are a linear mixture of latent components which are statistically independent. In neuroscience, ICA is typically used to find artifacts in signals [2], and can be a bottleneck for many processing pipelines.

Maximum likelihood estimation is one of the main ways of addressing the ICA problem [3]. The Hessian of the likelihood function has been thoroughly studied and some good and low cost approximation of it has been derived [4]. This approximation has been used for quasi-Newton methods [5]. This method has a quadratic convergence on simulated signals for which the independence assumption holds. However, on real data, the assumption is essentially false and the convergence rate falls back to linear. We address this issue by building on the classical optimization algorithm L-BFGS.

2 A new algorithm for maximum likelihood ICA

Given a set of N signals $x_1, \ldots, x_N$ with T samples each, that we can note as $X = [x_1, .., x_N]^\top \in \mathbb{R}^{N \times T}$, linear ICA aims at finding an unmixing matrix $W \in \mathbb{R}^{N \times N}$ such

that $WX = Y = [y_1, .., y_N]^\top$ contains mutually independent signals. Independence is meant in the statistical sense: the joint probability density function of $[y_1, .., y_N]$, p_Y, is equal to the product of the marginal densities of the y_i's, p_i: $p_Y(y_1, .., y_N) = \prod_{i=1}^{N} p_i(y_i)$. The inference can be done by maximum likelihood in the following way [3]. We postulate that $WX = Y$ has independent rows. Under this model, for each sample t, $Y(t)$ has a factorized density. The density of the corresponding data sample is computed by a linear change of variables, giving $p_X(X(t)) = |\det(W)| \prod_{i=1}^{N} p_i((y_i(t))$. Finally, we obtain the negative averaged log-likelihood of the data:

$$\mathcal{L}(W) = -\log|\det(W)| - \hat{E}\left[\sum_{i=1}^{N} \log(p_i(y_i))\right] \ . \tag{1}$$

$\hat{E}$ denotes the time average. We want to minimize $\mathcal{L}$. We can derive the relative gradient G and Hessian H of $\mathcal{L}$ when expanding the quantity $\mathcal{L}((I+\mathcal{E})W)$ at the second order in $\mathcal{E}$: $\mathcal{L}((I+\mathcal{E})W) = \mathcal{L}(W) + \langle G|\mathcal{E}\rangle + \frac{1}{2}\langle \mathcal{E}|H|\mathcal{E}\rangle + \mathcal{O}(||\mathcal{E}||^3)$. G is a $N \times N$ matrix, H is a $N \times N \times N \times N$ tensor, and we find:

$$G_{ij} = \hat{E}[\psi_i(y_i)y_j] - \delta_{ij} \ , \tag{2}$$

$$H_{ijkl} = \delta_{il}\delta_{jk} + \delta_{ik}\hat{E}[\psi_i'(y_i)y_j y_l] \ . \tag{3}$$

where ψ_i is the i^{th} *score function* : $\psi_i(\cdot) = -\log(p_i(\cdot))'$. We want to focus on the optimization procedure and not on the density/ score estimation, so we will take a fixed score $\psi_i(\cdot) = \tanh(\cdot/2)$, as in the standard ICA algorithm Infomax [6].

The Hessian is sparse but still has about N^3 non-zeros coefficients, making out of the box Newton method not practical. The Hessian expression simplifies if we assume that the signals y_i are independent. In that case, $\hat{E}[\psi_i'(y_i)y_j y_l] = \delta_{jl}\hat{E}[\psi_i'(y_i)]\hat{E}[y_j^2]$.

Thus, we can define a Hessian approximations:

$$\tilde{H}_{ijkl} = \delta_{il}\delta_{jk} + \delta_{ik}\delta_{jl}\hat{E}[\psi_i'(y_i)]\hat{E}[y_j^2] \ . \tag{4}$$

It has a block diagonal structure, with blocks of size 2×2: for a given pair (i, j), $\tilde{H}_{ijkl} \neq 0$ only when $(k, l) = (i, j)$ or $(k, l) = (j, i)$. It means that it is simple to invert, and order of magnitude faster to compute than H because it has fewer non-zero coefficients.

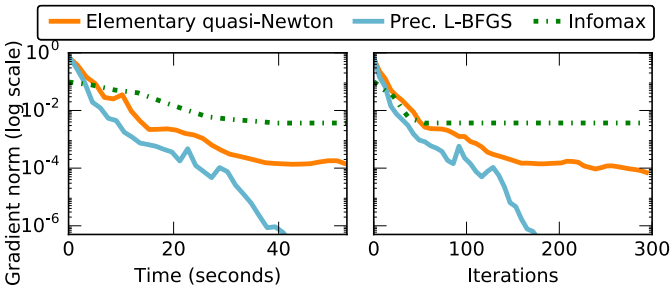
However, it equals H only when the signals are truly independent, which never happens in practice. This is why an elementary quasi-Newton method that goes in the direction $-\tilde{H}^{-1}G$ at each step will only have a linear rate on real data.

To obtain fast converge while taking into account the valuable information carried by those approximations, we introduce a preconditioned L-BFGS algorithm. The standard L-BFGS algorithm [7] builds an approximation of the true Hessian of the objective function by using only the past function and gradient calls. In the standard implementation the initial guess for the Hessian, for lack of a better solution, is taken as a multiple of identity. Our method simply uses the Hessian approximations as a first guess for the Hessian; the rest of the algorithm is the same.

3 Results

On 13 EEG datasets, we have run 3 algorithms: The elementary quasi-newton method, which constitutes the state of the art for maximum likelihood ICA in neuroscience, the preconditioned L-BFGS algorithm and Infomax.

For each of the 13 experiments, we store the gradient norm as a function of time and iterations. In the figure, we display the median of these curves.



We can see that the preconditioned L-BFGS algorithm converges much faster.

4 Discussion/Conclusion

In this work, a Hessian approximation of the objective function of maximum likelihood based ICA has been considered to accelerate estimation algorithms. This approximation can be too rough on real data, so we introduce a preconditioned L-BFGS method. Through analysis of EEG data, we have shown that this method outperforms the elementary quasi-Newton method which relies only on the Hessian approximation.

References

1. S. Makeig, T.-P. Jung, A. J. Bell, D. Ghahremani, and T. J. Sejnowski, "Blind separation of auditory event-related brain responses into independentcomponents," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 94, no. 20, pp. 10 979–10 984, 1997.
2. T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 2, pp. 163–178, 2000.
3. D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Transactions on Signal Processing*, vol. 45, no. 7, pp. 1712–1725, 1997.
4. S.-I. Amari, T.-P. Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345–1351, 1997.
5. J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "AMICA: An adaptive mixture of independent component analyzers with shared components," Tech. Rep., 2012.
6. A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
7. R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

Aircraft Dynamics Identification
[Poster/Talk submission]

C. Rommel[⋆†], J. F. Bonnans[⋆],
B. Gregorutti[†] and P. Martinon[⋆]

CMAP Ecole Polytechnique - INRIA[⋆]
Safety Line[†]

Abstract. Several Maximum Likelihood based approach for aircraft dynamics identification are presented and compared. The motivation is the need of accurate dynamic models for minimizing aircraft fuel consumption using optimal control techniques. Feature selection through the *Bolasso* is used to build the structure of the performance models. Real flight data from 25 different aircraft are used to validate our strategy.

Keywords: multi-task learning, feature selection, Bolasso, Maximum Likelihood

1 Introduction

Aircraft dynamics identification has been a longstanding problem in aircraft engineering, and is essential today for the optimization of flight trajectories in aircraft operations. This motivates the search for accurate dynamical systems identification techniques, the main topic of this study. The application we are most interested in here is aircraft fuel consumption reduction. It is known that this is a major goal for airlines nowadays, mainly for economic reasons, but also because it implies less CO_2 emissions. We limit our study to civil flights, and more specifically to the climb phase, where we expect to have more room for improvement. The techniques presented hereafter are suited for data extracted from the Quick Access Recorder (QAR). They contain multiple variables such as the pressure altitude and the true airspeed, with a sample rate of one second.

According to the literature [3], two widely used approaches for aircraft dynamics estimation are the Output-Error Method and Filter-Error Method, based on the main ideas of measurement error minimization and state dynamics re-estimation. Recent advances include using neural networks for the state estimation part [5]. On the other hand, renewed interest for the older Equation-Error Method has also been observed [4]. We propose in this paper variations of the latter. Adopting a statistical learning point of view, we state several regression formulations of our problem and solve them using Maximum Likelihood based techniques. We illustrate our methods with numerical results based on real data from 10 471 flights.

2 Methods

The main flight mechanics model used in this study is the following:

$$\begin{cases} \dot{h} = & V \sin\gamma, & (1) \\ \dot{V} = & \dfrac{T\cos\alpha - D - mg\sin\gamma}{m}, & (2) \\ \dot{\gamma} = & \dfrac{T\sin\alpha + L - mg\cos\gamma}{mV}, & (3) \\ \dot{m} = & -C_{sp}T, & (4) \end{cases}$$

In system (1)-(4), the elements T, D, L and C_{sp} are unknown and assumed to be functions of the state variables $\boldsymbol{x} = (h, V, \gamma, m)^\top$ and control variables $\boldsymbol{u} = (\alpha, N_1)^\top$. Now, given an aircraft for which a sufficient amount of flight data is available, we aim to identify these four functions with the highest precision possible.

For practical reasons, only parametric inference methods are considered here, which means that parametric models for the functions T, D, L and C_{sp} are needed. Based on flight mechanics knowledge, sets of possible features were determined for each of these functions. Assuming linear models for all of them, feature selections are performed using the Bolasso algorithm proposed in [1]. The robustness of this approach was assessed in some sense.

With the obtained models, a set of regression problems is derived from system (1)-(4). We see that equation (1) does not contain any unknown element to be estimated, which means it is not useful here. After leaving only the functions to be estimated in the r.h.s of the equations, we obtain

$$\begin{cases} Y_1 = X_1 \cdot \boldsymbol{\theta}_1 + \varepsilon_1, & (5) \\ Y_2 = X_2 \cdot \boldsymbol{\theta}_2 + \varepsilon_2, & (6) \\ Y_3 = (X_T \cdot \boldsymbol{\theta}_T)(X_{csp} \cdot \boldsymbol{\theta}_{csp}) + \varepsilon_3, & (7) \end{cases}$$

where

$$Y_1 = m\dot{V} + mg\sin\gamma, \qquad Y_2 = mV\dot{\gamma} + mg\cos\gamma, \qquad Y_3 = C, \tag{8}$$

and

$$X_1 = \begin{bmatrix} X_T\cos\alpha \\ -X_D \end{bmatrix}, \quad X_2 = \begin{bmatrix} X_T\sin\alpha \\ X_L \end{bmatrix}, \quad \boldsymbol{\theta}_1 = \begin{bmatrix} \boldsymbol{\theta}_T \\ \boldsymbol{\theta}_D \end{bmatrix}, \quad \boldsymbol{\theta}_2 = \begin{bmatrix} \boldsymbol{\theta}_T \\ \boldsymbol{\theta}_L \end{bmatrix}. \tag{9}$$

The vectors X_T, X_D, X_L, X_{csp} denote the feature vectors for T, D, L, C_{sp}, while $\boldsymbol{\theta}_T, \boldsymbol{\theta}_D$, $\boldsymbol{\theta}_L$, $\boldsymbol{\theta}_{csp}$ are the parameter vectors to be estimated and $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are random variables accounting for the model and data noise.

While $T = X_T \cdot \boldsymbol{\theta}_T$ appears in (5)-(7), note that $D = X_D \cdot \boldsymbol{\theta}_D$, $L = X_L \cdot \boldsymbol{\theta}_L$ and $C_{sp} = X_{csp} \cdot \boldsymbol{\theta}_{csp}$ take part in a single equation each. Equation (7) is clearly the problematic one, because of the presence of a product between the unknowns C_{sp} and T. This means that we do not have linearity on the parameters, but also that this regression cannot be solved to determine both elements separately: only the product of them is *identifiable* here. These obstacles led us to tackle the regression problems together, in a multi-task framework [2]:

$$Y = f(X, \boldsymbol{\theta}) + \varepsilon, \tag{10}$$

where

$$Y = [Y_1, Y_2, Y_3]^\top, \quad \varepsilon = [\varepsilon_1, \varepsilon_2, \varepsilon_3]^\top \in \mathbb{R}^3, \tag{11}$$

$$X = [X_T, X_{csp}, X_L, X_D]^\top \in \mathbb{R}^m, \quad \boldsymbol{\theta} = [\boldsymbol{\theta}_T, \boldsymbol{\theta}_{csp}, \boldsymbol{\theta}_L, \boldsymbol{\theta}_D]^\top \in \mathbb{R}^p. \tag{12}$$

The main idea here is that multi-task learning allows us to share the same thrust function T between all tasks. By doing so, we expect that some information gathered while learning tasks (5) and (6) will be transferred to task (7) during the process. This should help to reduce the non-identifiability issue.

Three algorithms were used in this study to solve problem (10). They are all based on the Maximum Likelihood estimator, with different assumptions and different implementation choices.

3 Results

The feature selection algorithm has been used to build models of the aerodynamic forces D, L of several aircraft of the same type (B737). We expect these planes to have similar aerodynamic behaviors, which should translate into similar model structures. Comparing the selection supports confirms the robustness of the Bolasso algorithm for our application: we obtain similar selected features for all aircraft. Real data recorded from 10 471 flights from 25 different aircraft were used to get the results of this study, which corresponds to approximately 8 261 619 observations.

Concerning the parameters estimation part, the different approaches proposed were compared to each other. As a reference case, we used a most straightforward strategy consisting on assuming an off-the-shelf model for C_{sp} and identifying the three other functions using single-task linear least-squares. Taking into account the future use of the estimated models, an optimal control based criterion was designed for the assessment of these estimators. The results indicate that all of them approximate with good accuracy the dynamics of a given aircraft using its historic QAR data. The comparison between the four methods showed that the use of a multi-task scheme in three of them led to better accuracy and, in addition, allowed to estimate all unknown functions of the problem: T, L, D and C_{sp}. The next step will be to use the identified models for trajectory optimization.

References

1. F. Bach. Bolasso: model consistent Lasso estimation through the bootstrap. pages 33–40. ACM, 2008.
2. R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.
3. R. V. Jategaonkar. *Flight Vehicle System Identification: A Time Domain Methdology*. AIAA, 2006.
4. E. A. Morelli. Practical aspects of the equation-error method for aircraft parameter estimation. In *AIAA Atmospheric Flight Mechanics Conference*, Aug., 21-24 2006.
5. N. K. Peyada and A. K. Ghosh. Aircraft parameter estimation using a new filtering technique based upon a neural network and Gauss-Newton method. *The Aeronautical Journal*, 113(1142):243–252, 2009.

Finding Interesting Aggregates in RDF Graphs
[Poster demo submission]

Yanlei Diao, Ioana Manolescu, and Shu Shang

Ecole Polytechnique, Inria Saclay, and Université Paris Saclay

Abstract. RDF is the format of choice for representing Semantic Web data. RDF graphs may be large and their structure is heterogeneous and complex, making them very hard to explore and understand. To help users discover valuable insights from RDF graph, we present a method which automatically recommends and evaluates aggregation queries over the RDF graphs; the queries are proposed to the user ranked based on a measure of their interestingness. We present the motivation and main ideas of our method, which we have implemented in a prototype, as well as some preliminary results. We also propose to demonstrate our tool to the JDSE audience on a set of open-access RDF graphs.

Keywords: Semantic Web, Aggregate Queries, Data Exploration

1 Introduction

RDF (the Resource Description Format) describes interconnected *resources* by specifying the *values* of their *properties*. Resources and properties must be Uniform Resource Identifiers (URIs, in short), whereas property values must be either URIs or values such as strings, numbers, dates etc. An RDF dataset is a set of *triples*, typically denoted (s, p, o), where s denotes the resource described, p is the property, and o is the object, i.e. the value of the property p of resource s. The special property denoted *rdf:type* allows to describe the classes to which a resource belongs. For instance, the triple $(uri_1, rdf{:}type, uri_{Person})$ states that uri_1 has the type uri_{Person}. Importantly, in RDF, a resource may have one or several types, or it may have no type at all. For instance, an RDF graph of bibliographic information may comprise triples such as $(a_1, rdf{:}type, uri_{Author})$, $(a_1, authorName, n_1)$, $(a_1, institution, i_1)$, $(a_1, institution, i_2)$, $(a_1, email, e_1)$ where a_1, i_1 and i_2 are the URIs of the authors and institutions, while n_1, e_1 are strings (author's name and e-mail). An ontology is sometimes available to characterize the semantics of an RDF graph; we do not consider ontologies here, and focus just on RDF data graphs.

Goal: identifying interesting aggregates RDF graphs may be very large and their structure complex and heterogeneous. In our bibliographic example, authors may lack an e-mail, have one, or have several; some may also have a *fundedBy* property etc. This leads to a lack of pre-defined schema, and makes it complicated for users to find out the interesting information hidden in an RDF graph. We propose acquainting the user with an RDF graph G by presenting her *interesting aggregate queries* evaluated over G, and whose results are automatically laid out in the form of a two-dimensional diagram or bar chart (as exemplified in Figure 1). For instance, sample aggregate queries based on the RDF graph representing the DBLP bibliographic dataset are:

α_0 "For each year, the total number of distinct authors having published a conference paper that year"

α_1 "For each year, the average number of authors of the conference papers published that year"

We consider α_0 or α_1 *interesting* if there is some strong trend or interesting spike in the average number of authors along the time axis. The problem we consider can be stated as: *given an RDF graph G and an integer k, find the k most interesting aggregate queries over G*. We discuss these notions in Section 2.

Our work shares the goal of finding interesting aggregates, with the SeeDB system [4] which, however, was designed for a single fact table T, in which the dimensions (group-by attributes $a_1, a_2, \ldots, a_k$) and measure m are known, and the goal is to find a subset $S \subseteq T$ and a dimension a_i, such that the distribution of the aggregate measure across dimension a_i strongly differs from the distribution of the same aggregate measure along a_i, for $T \setminus S$. In contrast, our input is a heterogeneous labeled RDF graph in which no facts, dimensions, and measures are known. Thus, most of our effort so far have been invested in identifying suitable candidates for these roles as well as aggregation functions, so as to lead to interesting aggregates.

2 Methodology

Given an RDF graph G, our approach is as follows.

An **RDF aggregate query** $q = \langle f, d, m, \oplus \rangle$ consists of (i) a *set of facts*, or resources over which the aggregate is computed (in α_0 and α_1, the conference papers); (ii) a *dimension* (in α_0 and α_1, the year). Note that due to the heterogeneity of RDF, some facts may lack the dimension (for instance, a paper may lack its publication year); in this case, that fact does not contribute to the aggregate query. Also due to RDF heterogeneity, some facts may have several values along the dimension, e.g., if the dimension is author affiliation and authors have multiple affiliations; (iii) a *measure* (the authors in α_0 and the *number of authors* in α_1; the latter is not present in the RDF graph, but we *derive* such properties); (iv) an aggregation function (*count(distinct(·))* in α_0 and *average* in α_1). The aggregation functions we consider are $\Omega = \{$*count*, *avg*, *sum*, *min*, *max*$\}$, possibly combined with a *distinct*. Again due to RDF heterogeneity, a fact may lack a measure, e.g., the authors may be unspecified for a paper. In such cases: if the aggregation function is *count*, the fact contributes with a count of 0; otherwise, that fact does not contribute to the aggregate query. A fact may also have several values for the measure, e.g. papers frequently have several authors; the aggregation function then applies over the complete set of measure values obtained for the facts having the same value for the dimension. Our queries are a subset of the RDF analytical queries introduced in [2,1] and also of those expressible in W3C's SPARQL 1.1.

Aggregate query interestingness is currently defined as the *variance* (second statistic moment) of the set of values $\{\oplus(\{f.m\}|f.d = x)\}_x$, where $\{f.m\}|f.d = x$ is the set of m values for all facts having the value x along dimension d.

To identify aggregate queries, we proceed as follows: **1.** we identify several *candidate fact sets*; **2.** for each candidate fact set, we explore a set of *candidate dimensions*; **3.** for each (f, d), we explore *candidate measures* m; **4.** for each (f, d, m) combination, we pick *applicable* aggregation functions from Ω; **5.** we evaluate the interestingness of all the aggregate queries resulting from steps **1.** to **4.** and we retain the top k, for a user-specified integer parameter k. Below, we outline each step.

1. Candidate fact set enumeration (i) If G contains some *rdf:type* triples, then we identify the set of class URIs in G (that is, the set of all URIs c such that a triple of

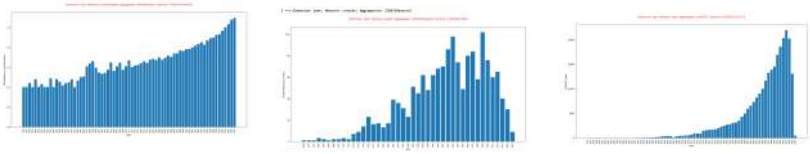


Fig. 1. Interesting aggregates found in DBLP RDF data (1936-2006). From left to right: the average number of article authors, per year; the total number of book authors, per book publication year; the number of conferences in DBLP, per year.

the form $(x, rdf{:}type, c)$ exists in G), and for each such c, the set f_C of all resources stated to be of type c in G is a candidate fact set; (ii) given a user-specified support threshold t_{supp} between 0 and 1, for any set of properties $P = \{p_1, p_2, \ldots, p_n\}$ such that at least t_{supp} of the triple subjects in G have (at least once) each property in P, the set f_P of all resources having all the properties in P is a candidate fact set.

2. Candidate dimension enumeration A dimension should be a property which all facts in f have (those lacking it cannot contribute to q anyway), or *count(p)* for some property p which some facts in f have. Further, d should have *relatively few distinct values*, as a property having almost a distinct value for each fact is not a meaningful agregation dimension. Formally, given a candidate fact set f and a property p which all facts from f have, let $S^0_{f,p}$ the set of distinct values of the property p on the facts from f. Property p is a candidate dimension if and only if $|S^0_{f,p}|/|f| \leq t_0$, where t_0 is a user-specified threshold t_0 between 0 and 1.

3. Candidate measure enumeration Given (f, d), a candidate measures m is a property of all facts in f, which is different from d, and such that d is not *count(m)*. For each m, we *automatically detect the majoritary data type* (string, integer, date, double) among all the values of property m for a fact in f. Again we use a threshold t_{type} and pick a type if at least t_{type} of the m values of facts f match that type.

4. Candidate aggregate functions Given (f, d, m), candidate aggregate functions $\oplus$ are chosen based on the type of the measure values. If this type is numeric, all Ω functions apply; *min*, *max*, *count* and *count(distinct)* apply to dates; *count* and *count(distinct)* apply to strings.

3 Preliminary results, scenario and perspectives

We have implemented our approach using Postgres 9.6 to store the RDF data and evaluate agregate queries, and Java 1.8 for the other operations. Figure 1 depicts some of the most interesting agregate queries identified by our prototype, given an RDF graph of DBLP bibliographic data. We plan to demonstrate our tool at JDSE on open-data RDF graphs, with the help of a Jupyter notebook. Future work includes the optimization of our candidate query enumeration method, the early pruning of unpromising candidates as in [4], and supporting higher-order aggregates as in [3].

References

1. E. A. Azirani, F. Goasdoué, I. Manolescu, and A. Roatis. Efficient OLAP operations for RDF analytics. In *ICDE Workshops*, pages 71–76, 2015.
2. D. Colazzo, F. Goasdoué, I. Manolescu, and A. Roatis. RDF analytics: lenses over semantic graphs. In *WWW*, pages 467–478, 2014.
3. B. Tang, S. Han, M. L. Yiu, R. Ding, and D. Zhang. Extracting top-k insights from multi-dimensional data. In *SIGMOD*, pages 1509–1524, 2017.
4. M. Vartak, S. Rahman, S. Madden, A. G. Parameswaran, and N. Polyzotis. SEEDB: efficient data-driven visualization recommendations to support visual analytics. *PVLDB*, 8(13):2182–2193, 2015.

Alzheimer's disease diagnosis using synchrony and disorder measures
Poster

A. Aljane and N. Houmani

SAMOVAR, Télécom SudParis, CNRS, Université Paris-Saclay
9 rue Charles Fourier 91011 EVRY Cedex, France

Abstract: This study tackles the issue of Alzheimer's disease diagnosis with electroencephalography based on the quantification of the brain functional connectivity. In this work, we evaluated two measures of functional connectivity, namely Phase Synchrony largely used in the literature and Epoch-based entropy, to distinguish between Mild Alzheimer's disease patients (AD patients) and patients with Subjective Cognitive Impairment (SCI subjects). A Linear Discriminate Analysis was used to discriminate the two groups of subjects with a leave-one-out procedure. The obtained results indicated that the accuracy reached 98.33% with Epoch-based entropy and 52.33% with Phase synchrony in θ band.

Keywords: Alzheimer's disease, Subjective Cognitive Impairment subjects, Phase Synchrony, Epoch-based entropy, Linear Discriminate Analysis.

1 Introduction

Alzheimer's disease (AD) is characterized by progressive neurodegeneration of brain network associated with memory and cognition. Diagnosis and intervention at the early stage prevents the development of the disease and limits the severity of cognitive decline. Many studies have proposed functional connectivity changes as an electroencephalographic (EEG) marker to discriminate AD patients from healthy subjects and from other pathologies. It has been shown that EEG markers contributed to the discrimination of normal elderly subjects from AD patients with 75-86 % of success (Huang et *al*., 2000; Knott et *al*., 2001; Blinowska et *al*., 2017). Concerning the discrimination between Mild AD and healthy control subjects, it has been shown that a classification rate of 80.49% can be reached with Phase Synchrony, at the frequency range of 9-10Hz (Gallego-Jutglà et *al*., 2015). It is also reported in the literature that classification accuracy between Mild AD patients and MCI subjects is estimated at 88% and 78% using two synchrony measures, Granger causality and stochastic event synchrony respectively, computed in α and θ global field power (Huang et *al*., 2000; Dauwels et *al*., 2009). In this work, we present our preliminary study on AD diagnosis in real life conditions by confronting AD patients to subjective cognitive impairment subjects (SCI subjects). Indeed, SCI subjects complained of memory impairment but have not been diagnosed with brain disorder. To the best of our knowledge, there is no study in this direction in the literature. To discriminate AD patients from SCI subjects, we applied two functional connectivity measures, namely Phase Synchrony (Lachaux et *al*., 1999; Dauwels et *al*., 2010; Gallego-Jutglà et *al*., 2015) and Epoch-based entropy (Houmani et *al*., 2015).

2 Methods

The present study relates to AD diagnosis using EEG time series data of 28 Mild AD patients and 22 SCI subjects. The EEG recordings were obtained at rest and with eyes closed using 30 electrodes placed following the International 10-20 system with a sampling rate of 256 Hz. The data were preprocessed and filtered in δ (0.5-4Hz), θ (4-8Hz), α (8-12Hz) and β (12-30Hz) frequency bands.

To assess the functional connectivity between EEG time series, Phase synchrony and Epoch-based entropy measures were used. Phase Synchrony (PS) quantifies the independence between the spontaneous phases φ_X and φ_Y of tow signals X and Y. It ranges between 0 (no phase synchronization) and 1 (total synchronization)(Lachaux et *al*., 1999; Dauwels et *al*., 2010; Gallego-Jutglà et *al*., 2015). Epoch-based entropy (EpEn) quantifies disorder in EEG signals at the time and spatial level using local density estimation by a Hidden Markov Model on inter-channel stationary epochs (Houmani et *al*., 2015).

In this study, we evaluated the reliability of Epoch-based entropy measure in discriminating SCI subjects from Mild AD patients, and we compared the performance to those obtained with Phase Synchrony, which is largely used in the literature for AD diagnosis (Lachaux et *al*., 1999; Dauwels et *al*., 2010; Gallego-Jutglà et *al*., 2015; Blinowska et *al*., 2017). To this end, for each SCI subject and Mild AD patient, both PS and EpEn values were computed between all EEG channels for all frequency bands; thus two functional connectivity matrices are associated to each subject in each frequency band. The Linear Discriminate Analysis (LDA) with a leave-one-out approach was applied to discriminate between AD patients and SCI subjects, by taking into account as a feature the average functional connectivity matrix on AD patients and SCI patients.

3 Results

Figure 1 shows the box plots of average functional connectivity values for SCI subjects and AD patients in different frequency bands. First, we observe a better discrimination between SCI subjects and AD patients with Epoch-based Entropy (Figure1.B) compared to Phase Synchrony (Figure 1.A), in all frequency bands. Figure 1 also shows that AD patients have higher values of EpEn than SCI subjects in δ and θ bands. We observe an inversion of this behavior for higher frequencies (α and β bands). Moreover, we can see in δ band that the variability is greater for SCI subjects than for AD patients; this reflects the high correct classification of Mild AD patient (sensitivity=82%).

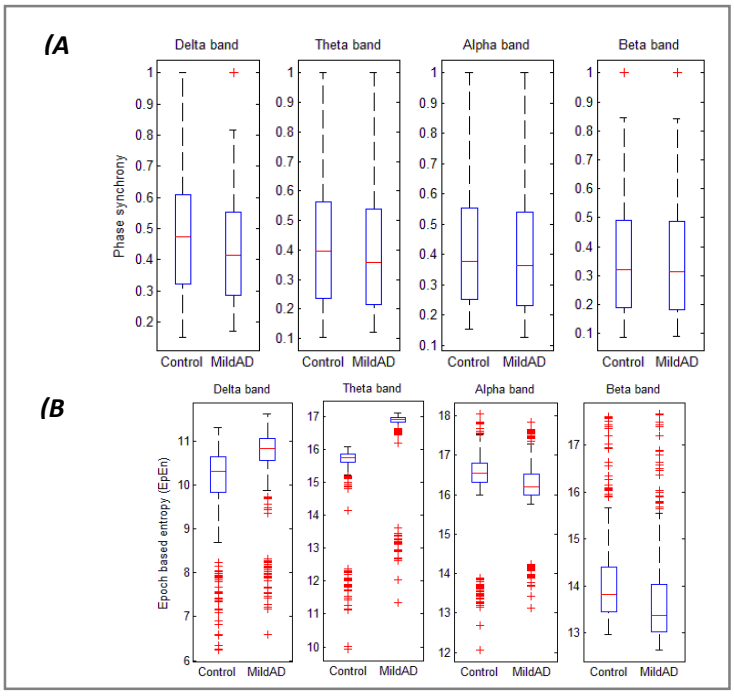


Figure 1 : The box plots of average functional connectivity matrix for the controls (SCI) and patients (Mild AD) in the different frequency bands: ***(A)*** phase synchrony (PS) measure; ***(B)*** Epoch-based entropy (EpEn) measure.

Results of LDA classification showed a higher correct classification rate with Epoch-based entropy measure than with Phase Synchrony, and that for all frequency bands. Indeed, the classification accuracy with Epoch-based entropy measure is greater than 70 %, while it is around 50% with Phase Synchrony. More precisely, with Epoch-based entropy measure in θ band, we reached an accuracy of 98.33% with a specificity of 100% (percentage of SCI subjects correctly classified) and a sensitivity of 96.66% (percentage of AD patients correctly classified).

4 Conclusion

Overall, the results of this preliminary study suggest that, on one hand, we have a better discrimination between SCI subjects and AD patients with Epoch-based entropy measure compared to Phase Synchrony measure. On the other hand, the classification result with Epoch-based entropy measure (a classification accuracy of 96.33% in θ band) is superior to those obtained in the literature when confronting AD patients to healthy subjects (Blinowska et *al*., 2017; Gallego-Jutglà et *al*., 2015). Therefore, this study confirms the effectiveness of Epoch-based entropy measure for AD diagnosis.

References

Blinowska, K.J., Rakowski, F., Kaminski, M., Fallani, F.D.V., Percio, C.D., Lizio, R., Babiloni, C., 2017. Functional and effective brain connectivity for discrimination between Alzheimer's patients and healthy individuals: A study on resting state EEG rhythms. Clin. Neurophysiol. 128, 667–680. doi:10.1016/j.clinph.2016.10.002.

Dauwels, J., Vialatte, F., Latchoumane, C., Jeong, J., Cichocki, A., 2009. EEG synchrony analysis for early diagnosis of Alzheimer's disease: A study with several synchrony measures and EEG data sets. in: 2009 Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society. pp. 2224–2227. doi:10.1109/IEMBS.2009.5334862

Dauwels, J., Vialatte, F., Musha, T., Cichocki A., 2010. A comparative study of synchrony measures for the early diagnosis of Alzheimer's disease based on EEG. Neuroimage. 49(1), 668–693.

Gallego-Jutglà, E., Solé-Casals, J., Vialatte, F.-B., Elgendi, M., Cichocki, A., Dauwels, J., 2015. A hybrid feature selection approach for the early diagnosis of Alzheimer's disease. J. Neural Eng. 12, 016018. doi:10.1088/1741-2560/12/1/016018

Houmani, N., Dreyfus, G., Vialatte, F.B., 2015. Epoch-based Entropy for Early Screening of Alzheimer's Disease. Int. J. Neural Syst. 25, 1550032. doi:10.1142/S012906571550032X

Huang, C., Wahlund, L., Dierks, T., Julin, P., Winblad, B., Jelic, V., 2000. Discrimination of Alzheimer's disease and mild cognitive impairment by equivalent EEG sources: a cross-sectional and longitudinal study. Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol. 111, 1961–1967.

Knott, V., Mohr, E., Mahoney, C., Ilivitsky, V., 2001. Quantitative electroencephalography in Alzheimer's disease: comparison with a control group, population norms and mental status. J. Psychiatry Neurosci. 26, 106–116.

Lachaux, J.P., Rodriguez, E., Martinerie, J., Varela, F.J., 1999. Measuring phase synchrony in brain signals. Hum. Brain Mapp. 8, 194–208.

Efficient Learning of Functional Outputs using Operator Random Fourier Features

Poster submission

Romain Brault, Alex Lambert, Florence d'Alché-Buc

LTCI, Télécom ParisTech

Abstract. In various application domains, one has to deal with complex outputs, such as functional output. Learning functional output can be tackled through the use of operator-valued kernels, at the cost of great computational load. In this work we propose a model based on Operator Random Fourier Features that scales up while still giving the statistical guarantees needed for learning.

Keywords: operator-valued kernels, efficient learning, functional output, random features

1 Introduction

Statistical learning of structured outputs is a fundamental and challenging subject of computer science and applied mathematics [10, 1]. There are several important application domains where the output exhibits such a structured nature; examples include words, sentences, graphs, images, time series, or even probability distributions.

Key to the success of prediction is how the dependencies of the output values are encoded, and how they are exploited. A mathematically sound way of encoding prior information about the relation of the outputs can be realized by operator-valued kernels and the associated vector-valued RKHS-s (reproducing kernel Hilbert spaces). Vector-valued RKHS-s have already found a few promising initial applications for example in multitask regression [4], vector field learning, vector autoregression, link prediction, joint quantile regression [8], and functional output prediction [5], which is the subject of this work. In all these applications, operator-valued kernels allow to cope with various surrogate losses adequate to structured outputs. The flexibility and expressive power of kernel methods, however, have a price: they are expensive in terms of memory and computational load.

In order to leverage the expressiveness of kernel techniques, numerous approximate schemes have been proposed in the literature for the case of scalar-valued kernels such as low-rank matrix approximations including the Nyström method [11], explicit feature maps designed for additive kernels, hashing [9, 6], and random Fourier features (RFF) [7] constructed for continuous shift-invariant kernels, the focus of the current work.

RFFs implement an extremely simple, yet efficient idea: by the Bochner's theorem one can construct an explicit low-dimensional random Fourier feature map, and a plug-in Monte-Carlo estimator to the kernel. RFFs now serve as a baseline for approximate

kernel machines allowing to cope with millions of data points and possess sound theoretical guarantees both on the quality of kernel approximation and on generalization performance.

Similarly to the scalar-valued case, operator-valued kernels suffer from serious computational and memory issues. The number of available approximate solutions are, however, quite limited in this domain: the RFF approach has just recently been extended [2], with the first concentration results on the quality of the operator-valued kernel approximation. This extension allows to deal with bigger amounts of data.

2 Methods

We consider the problem of learning functional output. Let $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{Y}$ be a RKHS of functions. Given i.i.d observations $(x_i, y_i)_{i\in\{1,n\}} \in \mathcal{X} \times \mathcal{Y}$, we are interested in estimating an unknown function F which minimize some criterion. For example, as introduced in [5], we may be interested in solving the following problem

$$\tilde{F} \in \underset{F\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} ||y_i - F(x_i)||_{\mathcal{Y}}^2 + \lambda||F||_{\mathcal{F}}^2 \tag{1}$$

where $\lambda \in \mathbb{R}_+$ is a regularization parameter.
Depending on the structure of $\mathcal{F}$, this problem can be hard to solve. One way to make it solvable is to consider $\mathcal{F}$ to be a vector-valued RKHS induced by a non-negative operator-valued kernel K [3]. The choice of K is of great importance since it will have a great influence over the family of functions $\mathcal{F}$ it defines. The knowledge of K allows us to formally resolve the above problem, since we have an equivalent of the representer theorem in the vector-valued case.

Theorem 1. *Representer theorem for vector-valued functions. Let K be an non-negative operator-valued kernel defining a vector-valued RKHS of functions $\mathcal{F}$.*
There exist $u_1, \ldots, u_n \in \mathcal{Y}$ such that the solution $\tilde{F}$ of the above problem can be written as

$$\tilde{F} = \sum_{i=1}^{n} K(x_i, .)u_i$$

The learning problem is now reduced to finding those elements. By computing the directional derivative and setting it to zero, one can derive an analytic solution for this problem. As in [5], the vector of functions $\mathbf{u} \in \mathcal{Y}^n$ satisfies the system of linear operator equations

$$(\mathbf{K} + \lambda I)\mathbf{u} = \mathbf{y}$$

where $\mathbf{K} = [K(x_i, x_j)]_{i\in\{1,n\}}$ is a $n \times n$ block operator matrix and $\mathbf{y}$ is the vector of functions y_i. The major drawback of this method is the high computational load associated to inverting the matrix K. To tackle this problem, we will approximate the kernel using Operator Random Fourier Features (ORFF) as introduced in [2]. This can be done thanks to an extension of the Bochner theorem for operator valued kernel, which basically says that any shift invariant non-negative operator valued kernel is the Fourier transform of some operator valued measure.

Theorem 2. *Spectral decomposition for shift invariant kernels Let μ be a positive measure on $\mathbb{R}^d$, and $A\colon \mathbb{R}^d \to \mathcal{L}(\mathcal{Y})$ such that $A(\omega) \geq 0$ for μ-almost all $\omega \in \mathbb{R}^d$,*

and A is bochner integrable with respect to μ. Then

$$K_e(\delta) = \int_{\mathbb{R}^d} e^{-i\langle \delta, \omega \rangle} A(\omega) \mathrm{d}\mu(\omega)$$

is the kernel signature of some shift-invariant non-negative operator valued kernel K. Moreover, any shift invariant kernel K is of the form above for some pair (A, μ).

Finding a decomposition is hard in general, but one can design its kernel so that we know such a pair (A, ω). The interest of this is to construct an approximated feature map thanks to a decomposition of the positive operator $A = BB^*$. Indeed, by drawing D random vectors $\omega_j, j = 1, \ldots, D$, one has that

$$\tilde{\Phi}(x) = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} e^{-i\langle x, \omega_j \rangle} B(\omega_j)^*$$

is an approximated feature map of K, which means that $\tilde{\Phi}(x)^* \tilde{\Phi}(z) \underset{D \to \infty}{\to} K(x, z)$ in the weak operator sense. Using this approximation, the learning problem becomes scalable, since we now model our function via this approximated feature map : $\tilde{F}(x) = \tilde{\Phi}(x)^* \theta$, and learning $\tilde{F}$ is now reduced to learning a parameter θ lying in some low dimensional space.

3 Conclusion

We propose a model for learning functional output based on the ORFF framework. By approximating the kernel through these feature maps, we can reduce the computational load and deal with big amounts of data. This modeling is valid in the optimization problem presented above, but can be adapted to any optimization problem in vector-valued RKHS defined by an operator-valued kernel (e.g quantile regression).

References

1. Álvarez, M.A., Rosasco, L., Lawrence, N.D.: Kernels for vector-valued functions: a review. Foundations and Trends in Machine Learning 4(3), 195–266 (2012)
2. Brault, R., Heinonen, M., d'Alché-Buc, F.: Random Fourier features for operator-valued kernels. Asian Conference in Machine Learning (ACML); PMLR 63, 110–125 (2016)
3. Carmeli, C., Vito, E.D., Toigo, A., Umanitá, V.: Vector valued reproducing kernel Hilbert spaces and universality. Analysis and Applications 8, 19–61 (2010)
4. Ciliberto, C., Mroueh, Y., Poggio, T., Rosasco, L.: Convex learning of multiple tasks and their structure. pp. 1548–1557 (2015)
5. Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., Audiffren, J.: Operator-valued kernels for learning from functional response data. Journal of Machine Learning Research 17, 1–54 (2016)
6. Kulis, B., Grauman, K.: Kernelized locality-sensitive hashing. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 1092–1104 (2012)
7. Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: Neural Information Processing Systems (NIPS). pp. 1177–1184 (2007)

8. Sangnier, M., Fercoq, O., d'Alché-Buc, F.: Joint quantile regression in vector-valued RKHSs. In: Neural Information Processing Systems (NIPS). pp. 3693–3701 (2016)
9. Shi, Q., Petterson, J., Dror, G., Langford, J., Smola, A., Strehl, A., Vishwanathan, V.: Hash kernels. International Conference on Artificial Intelligence and Statistics (AISTATS) 5, 496–503 (2009)
10. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research 6, 1453–1484 (2005)
11. Williams, C.K.I., Seeger, M.: Using the Nyström method to speed up kernel machines. In: Neural Information Processing Systems (NIPS). pp. 682–688 (2001)

Modeling Spatio-Temporal Data with Operator-Valued Kernel Methods: an Application to Epidemics
[Poster submission]

Camille Jandot, Florence d'Alché-Buc

LTCI, Télécom ParisTech

Abstract. In various application domains, such as climatology or epidemiology, data are collected through space and time, and it can be critical to be able to detect outbreaks in data. In this work, we propose an extension of Operator-valued Kernel-based Vector Autoregression (OKVAR) for autoregressive models of order p, that also allows to input exogenous data. We also propose a scaled-up version of this model using Operator Random Fourier Features. The proposed models will be tested on epidemics data.

Keywords: space-time series, operator-valued kernels, epidemics modeling

1 Motivation

Space-time series are collected in many fields. For instance, French public health institute *Sentinelles* provides weekly geographical reports of the incidence rates observed by general practitioners for a list of diseases under surveillance (Flahault et al. 2006).

Lots of research has focused on the problem of detection of epidemics recently. Google Flu Trends (Ginsberg et al., 2009) relies on linear models where the target in the influenza incidence rate and the features are the number of Google queries identified as most predictive of the incidence rate. Held et al. (2006) propose a model with an endemic component that describes seasonal patterns and an epidemic component that enables to capture outbreaks through autoregression. Meyer et al. (2014) propose an extension that takes the space dimension into account.

We aim at building a non-linear non-parametric model that takes the relationship between regions into account. We propose to model space-time series using operator-valued kernels (Micchelli and Pontil, 2004). Operator-valued kernels are an extension of scalar-valued kernels and enable to learn vector-valued functions. The model will be applied to the modeling of influenza epidemics, using *Sentinelles* data.

2 Operator-valued Kernel-based Vector Autoregressive Model of Order p

We consider the problem of epidemics modeling (e.g. influenza). We write $\boldsymbol{x_t}$ the d-dimensional vector whose i^{th} coordinate denotes the incidence rate of the considered

disease in location i, and $\boldsymbol{x_t^p} \in \mathbb{R}^{dp}$ the vector that results in the concatenation of $\boldsymbol{x_t}, \boldsymbol{x_{t-1}}, \ldots, \boldsymbol{x_{t-p+1}}$. The dataset consists of $\mathcal{D}_n = (\boldsymbol{x_i^p}, \boldsymbol{x_{i+1}})_{i=1}^n$.

In Lim et al. (2014), the authors introduce Operator-valued Kernel-based Vector Autoregression (OKVAR) for non-linear autoregression of order 1. We extend OKVAR to autoregression of order p: $\boldsymbol{x_{t+1}} = f(\boldsymbol{x_t^p}) + \boldsymbol{\varepsilon}_{t+1}$, where $\boldsymbol{\varepsilon}_t$ is a noise term. The optimization problem $\mathcal{P}$ writes

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{t=1}^{n} \|\boldsymbol{x_{t+1}} - f(\boldsymbol{x_t^p})\|_2^2 + \lambda \|f\|_{\mathcal{H}}^2, \lambda > 0$$

where $\mathcal{H}$ denotes the Reproducing Kernel Hilbert Space associated to the operator-valued kernel K. We have a representer theorem (Micchelli and Pontil, 2005) and $f(.) = \sum_{i=1}^n K(., \boldsymbol{x_i})\boldsymbol{c_i}$, where the vectors $c_i \in \mathbb{R}^d,\ i = 1, \ldots n$ need to be learned, and K is an operator-valued kernel. In case $p = 1$, we can simply use the decomposable Gaussian kernel $\boldsymbol{x_i}, \boldsymbol{x_j} \rightarrow K(\boldsymbol{x_i}, \boldsymbol{x_j}) = Ak(\boldsymbol{x_i}, \boldsymbol{x_j})$ where A is a $(d \times d)$-symmetric positive-definite matrix that should also be learned and k denotes the scalar Gaussian kernel. In case $p > 1$, we need kernels that can input $\boldsymbol{x_t}, \cdots, \boldsymbol{x_{t-p+1}}$. We can build them combining kernels acting on d-dimensional vector either multiplicatively,

$$K(\boldsymbol{x_i^p}, \boldsymbol{x_j^p}) = A \prod_{\ell=1}^{n} k(\boldsymbol{x_{i-\ell+1}}, \boldsymbol{x_{j-\ell+1}}),$$

or additively,

$$K(\boldsymbol{x_i^p}, \boldsymbol{x_j^p}) = A \sum_{\ell=1}^{p} \alpha_\ell k(\boldsymbol{x_{i-\ell+1}}, \boldsymbol{x_{j-\ell+1}}),\ \alpha_\ell \geq 0\ \forall \ell = 1 \ldots p.$$

We will solve $\mathcal{P}$ by alternately optimizing in A, C and $\boldsymbol{\alpha}$ in the case kernels are combined additively.

3 Scaling up the Model with Operator Random Fourier Features

In this section, we propose a scaled-up version of the model introduced in the previous section based on Operator Random Fourier Features (ORFF). ORFF were introduced in Brault et al. (2016) to extend Random Fourier Features (Rahimi and Recht, 2008) and propose an approximation of operator-valued kernels. The authors propose the following approximation for the decomposable Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{x'}) = A \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x'}\|^2}{2\sigma^2}\right)$:

$$\tilde{K}(\boldsymbol{x}, \boldsymbol{x'}) = \tilde{\phi}(\boldsymbol{x})^\top \tilde{\phi}(\boldsymbol{x'}),$$

where

$$\tilde{\phi}(\boldsymbol{x}) = \frac{1}{\sqrt{V}} \begin{pmatrix} \cos\langle \boldsymbol{x}, w_1 \rangle \\ \sin\langle \boldsymbol{x}, w_1 \rangle \\ \vdots \\ \cos\langle \boldsymbol{x}, w_V \rangle \\ \sin\langle \boldsymbol{x}, w_V \rangle \end{pmatrix} \otimes B^\top,$$

with $w_j \sim \mathcal{N}(\mathbf{0}, \sigma^{-2} I)$ and B a $(d \times m)$-matrix such that $A = BB^\top$.

Using this approximation, we propose the following optimization problem:

$$\min_{\boldsymbol{\theta}, \boldsymbol{B}} \frac{1}{n} \sum_{t=1}^{n} \|\boldsymbol{x_{t+1}} - \tilde{\Phi}_B(\boldsymbol{x_t^p})^\top \boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_2^2, \lambda > 0$$

with

$$\tilde{\Phi}_B(\boldsymbol{x_t^p}) = \begin{pmatrix} \tilde{\phi}(\boldsymbol{x_t}) \\ \vdots \\ \widetilde{\phi}(\boldsymbol{x_{t-p+1}}) \end{pmatrix} \in \mathbb{R}^{2Vmp \times d}$$

if the kernels are combined through positively-weighted addition, and $\tilde{\Phi}_B(\boldsymbol{x_t^p}) = \widetilde{\phi}(\boldsymbol{x_t^p}) \in \mathbb{R}^{2Vm \times d}$ if they are combined multiplicatively.

4 Conclusion

We propose a model for non-linear non-parametric autoregression of order p extending OKVAR models. We will use the same combinations of kernels as introduced in Section 2 to input exogenous data in the model (e.g. Google Trends data, seasonality, ...). We also provide a scaled-up version of the learning problems based on Operator Random Fourier Features.

References

Réseau Sentinelles, INSERM/UPMC. `http://www.sentiweb.fr`.

Romain Brault, Markus Heinonen, and Florence Buc. Random fourier features for operator-valued kernels. In *Asian Conference on Machine Learning*, pages 110–125, 2016.

Antoine Flahault, Thierry Blanchon, Yves Dorleans, Laurent Toubiana, Jean-François Vibert, and Alain-Jacques Valleron. Virtual surveillance of communicable diseases: a 20-year experience in france. *Statistical methods in medical research*, 15(5):413–421, 2006.

Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

Leonhard Held, Mathias Hofmann, Michael Höhle, and Volker Schmid. A two-component model for counts of infectious diseases. *Biostatistics*, 7(3):422–437, 2006.

Néhémy Lim, Florence D'Alché-Buc, Cédric Auliac, and George Michailidis. Operator-valued kernel-based vector autoregressive models for network inference. 2014.

Sebastian Meyer, Leonhard Held, and Michael Höhle. Spatio-temporal analysis of epidemic phenomena using the r package surveillance. *arXiv preprint arXiv:1411.0416*, 2014.

Charles A Micchelli and Massimiliano Pontil. Kernels for multi–task learning. In *NIPS*, volume 86, page 89, 2004.

Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

Smart ease : reducing energy costs by storage and consumption forcasts

Pierre-Louis Guhur et Charles Lorenzo[1]

[1]Départment EEA, ENS Paris-Saclay

July 31, 2017

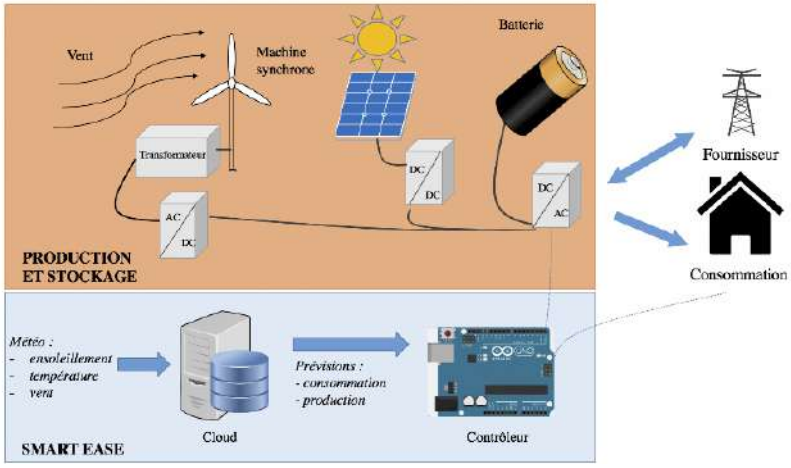


Figure 1: Prediction and optimisation of the energy flows

Abstract

Generation and storage of renewable energie are fastly improving. As a consequence, the energy landscape overcomes a complete overturning. Nowadays, *smart grids* quickly spread up in European countries in order to enable self-consumption. Nevertheless, the existing solutions use complicated methods, and are not adapted to the computing capacity of the connected objects.

With this in mind, we offer a new solution called *smart ease* optimizing energy storage for a home which can produce and store energy.

Our method relies on a *gradient boosting* algorithm for consumption forecast. In this study, we compare several ways to charge the battery. The realized simulations are based on real datas of consumption and production. The results show significant reductions on the energy bill.

keywords

renewable energy, smart grids, connected objects, machine learning, self consumption

1 Motivations

This study has been realized during our second year of Master at the Ecole Normale Supérieur de Paris-Saclay. It started for a student competition organised by CIGRE about *smart grids*. Then, it enables a collaboration with the start-up *Elum Energy*, which is currently working in similar perspectives, but in a larger geographic scale. But our research has also been driven by our strong enthusiasm in machine learning and renewable energy.

2 Introduction

Information technologies have enabled the developpement of the *smart grids*, using sensors and devices of information processing in order to optimize production, distribution and consumption of electricity.
Our study uses a Rasberry Pi to calculate and order the charge or the discharge of a battery. The home possesses a wind turbin or solar panels. The micro controller must optimize the storage and reduce the electrical bill. The figure 1 illustrates the purpose of the connected object. It acquires meteological datas (in terms of temperature, enlightment and wind speed) for the 24 next hours. These datas are used for the prediction of production, in real time.
The cost of the batteries is still high. These simulations help to determine the optimal capacity of the battery for the *smart home*.

3 Minimization of the electric bill

We tried several algorithms in order to predict the electrical consumption. The input datas are the historic of the consumption in a home, and the meteorological datas. The most reliable algorithme used is called *gradient boosting*, because the execution time is the shorter, for the same performance. The difference between real consumption and estimated consumption is calculated and reaches 2%

Moreover, in order to compare the savings made, we have tried five different laws to charge and discharge the battery, and to sell or to purchase to the supplier.

For this two first laws, the home does not own a battery.

- First law : *sell*, the home sells all the production to the supplier.
- the law *consumption*, the home consumes the production and sells the surpluses to the supplier.

  The cost of takeover of the produced energy is cheaper than the cost of energy sold by the supplier. That is why the 3 others laws use a battery to avoid buying from the supplier.
- The law *normal* stores the production surpluses until the maximum capacity of the battery.
- the law *purchase* takes advantage of the volatility of the price of energy. The home buys electricity to store during the off-peak electricity tariffs hours. During the last hour of off-peak electricity tariffs, the Raspberry decides to fill the battery until a pourcentage of filling (called *alpha*). The main purpose of this law is to find the optimal ratio *alpha*.
- The law *optimal* : the ratio *alpha* has to be adapted for every period of off-peak period, estimating the production and the filling of the battery for the hours when the price of electricity is higher. We developed a simple method which simulates the expenses during the peak time.

4 Conclusion and first results

This study presents *smart ease*, an optimized solution for a home that can produce and store energy. The tools used to reduce the electric bill are very cheap, and the speed of the calculation is really high. First, our method forecasts the consumption with a *gradient boosting* algorithm. The production forecast is predicted with meteorological datas. Then, the law called *optimal* estimates the amount of electricity to buy from the supplier during the period when the price is the cheapest. The simulations have been written in Python, and are publicly available in *https://github.com/plguhur/prod-electricite*
The results shows substantial benefits: the bill is reduced of 19% between the laws *sell* and *optimal* for a solar production. The table below summarizes the first results and compares the annual bill according to the employed law. If the price is positive, it means that the consumer have to pay the supplier. If the price is negative, it means that the consumer earns money with his/her equipment.

Bill	Wind turbine	Solar panels
sell	500	1250
consumption	-10	920
normal	-150	890
purshase	-155	885
optimal	-185	850

5 Evolution of the research

In order to improve the accuracy of our modelisation, we would like to integrate the ageing of the battery. The ageing of the battery depends on the loading rate, the temperature, and the depth of discharge. A model may optimize the ageing of the battery and improve the return on investment of the equipments. This model of forecasts can also be used as a failure detector.

6 Bibliography

Security and privacy challenges in the smart grid, *McDaniel, Patrick and McLaughlin, Stephen*

Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid, *Mohsenian-Rad, Amir-Hamed and Wong, Vincent WS and Jatskevich, Juri and Schober, Robert and Leon-Garcia, Alberto*

Données de consommation dans le canton de Bâle *https://www.swissgrid.ch/*

Electricity estimation using genetic algorithm approach: a case study of Turkey*Ozturk, Harun Kemal and Ceylan, Halim and Canyurt, Olcay Ersel and Hepbasli, Arif*

Electrical energy consumption estimation by genetic algorithm *Azadeh, A and Ghaderi, SF and Tarverdian, S*

SmartCharge: cutting the electricity bill in smart homes with energy storage*Mishra, Aditya and Irwin, David and Shenoy, Prashant and Kurose, Jim and Zhu, Ting*