Summary Report of My Scientific Activities during ERCIM Postdoc fellowship at ISTI-CNR

Osman Abul
KDD Laboratory, Institute of Information Science and Technology,
Italian National Research Council, Pisa, Italy
osman.abul@isti.cnr.it

February 16, 2007

1 Introduction

This report is prepared to describe my scientific activities during the ERCIM postdoc fellowship at Institute of Information Science and Technology (ISTI), between June 5, 2006 and March 4, 2007. This is the second of my two period fellowship the first of which I have spent at Norwegian University of Science and Technology (NTNU).

I have been affiliated with Knowledge Discovery and Delivery (KDD) laboratory of ISTI. The laboratory is a joint initiative of ISTI and the computer science department of University of Pisa (UPisa). The principal researchers are my supervisor Prof. Fosca Giannotti (of ISTI), Prof. Dino Pedreschi (of UPisa) and Prof. Franco Turini (of UPisa). The laboratory consists of several Postdocs, PhD students and a few Master students and technical collaborators as well. The laboratory has many national and international collaborations mostly through collaborative projects/programme, most importantly an EU frame-6 project GeoPKDD and KDUbiq. The research interest of the group includes

- Data mining techniques: association rule mining, clustering, classification etc.,
- Efficient and constrained data mining,
- Data mining for structured data including text, graph and sequence,
- Privacy issues in data mining,
- Data models, querying and query languages, and
- Data mining theory and applications for special domains, e.g. bio-medical, geographic trajectories.

My research activities at ISTI can be broadly divided into two; computational motif discovery and spatio-temporal data mining. For the former, I have mainly collaborated with Dr. Francesco Bonchi (of ISTI), Prof. Ercan Kuruoglu (of ISTI), Mr. Geir Kjetil Sandve (of NTNU) and Prof. Finn Drabløs (of NTNU). For the latter, I have mainly collaborated with Prof. Fosca Giannotti (of ISTI), Dr. Francesco Bonchi (of ISTI), Dr. Mirco Nanni (of ISTI) and Dr. Maurizio Atzori (of UPisa).

The document is organized as follows. Section 2 gives a summary of research activities I have been involved. In Section 3, the scientific conferences and seminars to which I attended as audience and/or presenter are given. Section 4 gives abstracts and co-author list for accepted and submitted publications.

2 Summary of Research

2.1 Spatio-Temporal Data Mining

Data mining is by nature an application oriented science. That is, new techniques/algorithms should be developed for each non-standard application domain like spatio-temporal domain. Among other data mining tasks, clustering of spatio-temporal trajectories is more interest to us. Recognizing that in the hearth of the clustering is the measurement of similarity between objects, we have worked on identifying relevant similarity function for a given objective. What is also important is the clustering strategy or the paradigm. Thus, we have started a principled clustering approach (including ad-hoc ones) specific to spatio-temporal data. Some hard and soft constraints are also taken into account. This work is still ongoing.

2.2 Spatio-Temporal Knowledge Hiding

Institutions, companies and government agencies publish their datasets to third parties for do-itself computations. However, the dataset can have sensitive information or patterns that shouldn't be disclosed to third parties. In case the sensitive patterns are known, they can be suppressed (a.k.a. sanitization) in the disclosed data. But, too much suppression limits the utility of the data. Therefore, finding methods providing sanitization with less distortion is a challenge problem.

In this study, we investigate the knowledge hiding in spatio-temporal datasets, e.g. cellphone trajectories. The special kind of dataset poses further challenges such as continuity in the trajectory, underlying map conformance, extraction of sensitive patterns, effective finding of sensitive patterns and a number of constraints on sensitive patterns. We have provided the formal problem definition, proved some characteristics (e.g. NP-hardness) and experimented with real datasets. The published paper (Section 4.1) gives our results. We are currently working on possible extensions and elaborations.

2.3 Spatio-Temporal Anonymity

In case the published data is a set of microdata, then the privacy of individuals to whom the microdata refers to becomes an important issue. One of the solutions to this problem is suppressing some microdata which are not anonymous enough. Another solution is providing anonymization by attribute aggregation. In both case, the published data is considered to be privacy-preserving if there is at least k other individuals (so called k-anonymity) with the same quasi-identifiers. However, providing privacy comes with a cost of data utility loss. So, the objective is making a good balance and avoiding over distortions at the same privacy level.

We have tackled the k-anonymity problem for spatio-temporal datasets where each microdata is a trajectory (obtained through cell phone) of an individual. The formal problem definition is provided in which uncertainty of locations are allowed and k-anonymity is defined accordingly. We have explored a number of anonymity operators and shown their computational properties, most importantly we have proved the NP-hardness of these problems and therefore developed some heuristic algorithms. The algorithms are mostly using the clustering of trajectories to provide anonymization.

A manuscript describing the problem and algorithms is prepared and we are currently experimenting. After finishing experiments the manuscript is going to be submitted for publication.

2.4 Benchmark Datasets for Computational Motif Discovery

This is the work that we have started during my first period of ERCIM fellowship at NTNU.

There are over 100 algorithms for computational motif discovery using different search techniques, motif models and scoring functions. Unfortunately, the literature lacks the which one to choose and use. So, benchmarking them becomes an important issue to make an informed choice. However, there is no ground truth for selecting or creating the benchmark data. In this work, we explore the issue of creating a good benchmark data using supervised learning techniques. For some datasets, we show that even the perfect motif discovery algorithms can not discover motifs because of higher Bayes errors. Noting this, benchmarking datasets are constructed such that motif/non motif separation is possible theoretically. We have written a paper on the issue (see Section 4.2).

2.5 Constrained Motif Discovery

Motif discovery from biosequences is a hot research topic because of its relevance to many bioinformatics tasks. There are three main ingredients of motif discovery tools; motif model, search technique and scoring function. With a particular instantiation a different algorithm is obtained. All the algorithms are accepting some essential parameters which can be interpreted as the preference of the user. The preference requires some meta-knowledge of the data set and some other experience. However, these parameters are usually algorithm specific and far beyond to capture all possible preferences from an experienced user. We have motivated with this and made a connection to constrained mining problem where the user can state all the preferences as constraints. As a side effect, some constraints also provides with efficient computation. In this work, we investigate the projection of constrained mining on motif discovery to get a principled constrained motif discovery problem. We are still working on the issue and no manuscript has been prepared yet.

3 Conferences and Seminars

3.1 Conferences

- The 21st International Symposium on Computer and Information Sciences (ISCIS'06), Istanbul, Turkey. Nov. 2006. Talk title "Asymptotical Lower Limits on the Required Number of Examples for Learning Boolean Networks".

- Computational Systems Bioinformatics Conference (CSB'06), Stanford University, CA, USA. Aug. 2006. Talk title "A Methodology for Motif Discovery Employing Iterated Cluster Re-assignment".

- The 3rd annual RECOMB Satellite Workshop on Regulatory Genomics (RECOMB-RG'06), Singapore. July 2006. Talk title "TSCAN: A two-step de novo Motif Discovery Method".

- Algorithmic Biology Workshop: Algorithmic Techniques in Computational Biology, Singapore. July 2006.

- 14th European Signal Processing Conference (EUSIPCO'06), Florence, Italy. Sep. 2006.

3.2 Seminars

- KDD Laboratory biweekly seminars. I have presented two talks with the titles 1) "Computational Biosequence Motif Discovery Problem and Related Issues", 2) "Anonymity of Location Trace Samples".
- GeoPKDD.it project meeting/seminar, Nov. 28, 2006. Pisa, Italy.
- GeoPKDD.eu project meeting/seminar, Dec. 12-13, 2006. Pisa, Italy.

3.3 Reviewer

External reviewer for International Conference on Data Mining (ICDM 2006), Hong Kong.

4 Publications

4.1 Paper 1

Title: Hiding Sequences.

Authors: Osman Abul, Maurizio Atzori, Francesco Bonchi and Fosca Giannotti.

Abstract: The process of discovering relevant patterns holding in a database, was first indicated as a threat to database security by O'Leary et al. Since then, many different approaches for knowledge hiding have emerged over the years, mainly in the context of association rules and frequent itemsets mining. Following many real-world data and applications demands, in this paper we shift the problem of knowledge hiding to contexts where both the data and the extracted knowledge have a sequential structure. We provide problem statement, some theoretical issues including NP-hardness of the problem, a polynomial sanitization algorithm and an experimental evaluation. Finally we discuss possible extensions that will allow to use this work as a basic building block for more complex kinds of patterns and applications.

Status: Accepted for publication in the 3rd Privacy Data Management workshop of International Conference on Data Engineering (ICDE'07) conference, Istanbul, Turkey. April

2006. Also appears as a technical report of ISTI (acc. number 2006-TR-40).

4.2 Paper 2

Title: Improved benchmarks for computational motif discovery.

Authors: Geir Kjetil Sandve, Osman Abul, Vegard Walseng and Finn Drabløs.

Abstract: *Background*: An important step in annotation of sequenced genomes is the identification of transcription factor binding sites. More than a hundred different computational methods have been proposed, and it is difficult to make an informed choice. Therefore, robust assessment of motif discovery methods becomes important, both for validation of existing tools and for identification of promising directions for future research. *Results*: We use a machine learning perspective to analyze collections of transcription factors with known binding sites. Algorithms are presented for finding position weight matrices (PWMs), IUPAC-type motifs and mismatch motifs with optimal discrimination of binding sites from remaining sequence. We show that for many data sets in a recently proposed benchmark suite for motif discovery, none of the common motif models can accurately discriminate the binding sites from remaining sequence. This may obscure the distinction between the potential performance of the motif discovery tool itself versus the intrinsic complexity of the problem we are trying to solve. Synthetic data sets may avoid this problem, but we show on some previously proposed benchmarks that there may be a strong bias towards a presupposed motif model. We also propose a new approach to benchmark data set construction. This approach is based on collections of binding site fragments from matrix alignments in TRANSFAC. Data sets are ranked according to the optimal level of discrimination achieved with our discrimination algorithms, allowing selection of subsets with specific properties. We present one benchmark suite with data sets that allow good discrimination between positive and negative instances with the common motif models. These data sets are suitable for evaluating algorithms for motif discovery. We present another benchmark suite where PWM, IUPAC and mismatch motif models are not able to discriminate reliably between positive and negative instances. This suite could be used for evaluating more powerful motif models. *Conclusion*: Supervised learning on motif collections may be a valuable approach for evaluation of benchmark data sets. Our improved benchmark suites have been designed to differentiate between the performance of motif discovery algorithms and the power of motif models.

Status: Submitted and under review for publication in BMC Bioinformatics journal.