Short-term mobility program for scientists/researchers from Italian and Foreign Countries (2008)

Toon Calders
t.calders@tue.nl
tel. : (+31)(0)40 247 4568
Information Systems Group
Faculty of Mathematics and Computer Science
Eindhoven University of Technology
PO Box 513
MB 5600 Eindhoven
The Netherlands

Administrative Information

Arrival at the research site: 22/9
Departure from the research site: 02/10
Time actually spent at the research site:
W1: Mon 22/9 till Fri 26/9, Sun 28/9,
W2: Mon 29/9 till Thu 02/10;
Total time spent at CNR approximately 100h.

Declaration

I hereby declare that the information provided in this report is correct and accurate. I commit myself to mention the support of CNR in any scientific report or publication that may result from my visit to CNR.

Toon Calders

(Scientific report of the activities during the visit are attached.)

Research Activities During Visit of 22/9—3/10 at CNR, Pisa

Toon Calders
Technische Universiteit Eindhoven
The Netherlands

Introduction

During the visit there have been numerous meetings with the different team members of the KDD lab of CNR/University of Pisa, group meetings as well as face-to-face meetings with smaller groups. An overview of the the most important meetings, is given in the following tables:

Mon	Making agenda for the visit; practical arrangements.
Tue	"Classification without Discrimination" (Faisal Kamiran) "Discrimination aware Data Mining" (Franco Turini)
Wed	"Mining Conjunctive Sequential Patterns" (Toon Calders) "Basics on Trajectory patterns" (Mirco Nanni) "Anonymization of Sequential patterns" (Ruggero Pensa)
Thu	Trajectory patterns: ideas to condense the collection of T-patterns Workflow mining
Fri	Workflow mining, continued
Sun	No meeting — organizing and structuring information

Mon	"Anti-Monotonic Graph Support Measures" (Toon Calders) "Ideas from SNA applied to Digital bibliography" (Michele Coscia)
Tue	"Focused Rules Classification" (Laura Spinsanti) "Location prediction in the mobility data analysis environment Daedalus" (Roberto Trasarti)
Wed	"Data Mining query language"; Roberto Trasarti, Chaira Renso, Fosca Giannotti
Thu	"Discrimination-aware data mining" "Anonymization of sequential data via segmentation" "Tax dataset" Discussion of idea for research proposal "Social networks"

During these meetings the different recent research directions in the KDD lab and in the data mining group at the TU/e have been discussed and different areas for future collaboration have been identified. Three topics were selected for further collaboration:

- Discrimination-aware Data mining;
- Anonymization of sequential data;
- Mining suspicious tax filings with high accuracy;

These three areas are shortly described and future research directions are highlighted in the remainder of this document.

Discrimination-Aware Data Mining

Collaborators

Franco Turini	CNR and University of Pisa
Salvatore Rugieri	CNR and University of Pisa
Dino Pedreschi	University of Pisa
Toon Calders	Technische Universiteit Eindhoven
Faisal Kamran	Technische Universiteit Eindhoven

Problem Description

Classification models predict the class labels of unknown data samples. Often, however, the training data is biased towards certain groups or classes of objects. For example, throughout the years, in a certain organization black people might systematically have been denied from jobs. As such, the historical employment information of this company concerning job applications will be biased towards giving jobs to white people while denying jobs from black people.

In order to reduce this type of racial discrimination, new laws requiring equal job opportunity have been enacted by the government. As such, the organization receives instructions in the form of, e.g., minimum quota for black employees. Suppose now that the company wants to partially automate its recruitment strategy by learning a classifier that predicts the most likely candidates for a job. As the historical recruitment data of the company is biased, the learned model may show unlawfully prejudiced behavior. This partial attitude of the learned model leads to discriminatory outcomes for future unlabeled data objects.

In this context, two research questions are quite obvious:

1. *how can we find specific regions where discrimination is particularly high?*
2. *how can we train an unbiased classifier when the training data is biased?*

Kamiran and Calders [3] introduced a classification model which is learnt on biased training data but works impartially for future data. First, the discriminatory data is changed in a minimal way as to remove the existing discrimination. To this end we use a ranking function learned on the biased data. Then, based on the sanitized data, a non-discriminatory model can be learned. The fact that this model is learned on non-discriminatory data reduces the prejudicial behavior for future classification to a minimum level. This solution provides us with an opportunity to keep the discriminatory behavior of a classification model at a minimum level on the basis of any sensitive attribute. We refer this model as *Classification with No Discrimination (CND)*. Obviously, changing the training data might result in lower accuracy scores. Nevertheless, as we try to keep the changes as minimal and least intrusive as possible, the trade-off between accuracy and non-discrimination will be minimal.

Orthogonally to this work, Pedreschi et al. [5] concentrated on identifying discriminatory rules that are present in a dataset. From a given dataset, they learn potential discriminatory guidelines that have been followed in the decision procedure. A central notion in their work is that of the *context* of the discrimination. That is, specific regions in the data are identified in which the discrimination is particularly high. One of the main strengths of the work by Pedreschi et al. is the realistic assumption that often the dataset itself does not contain the attribute against which there is discrimination (e.g., race). Nevertheless, they show an elegant way to instead use background knowledge in order to identify discriminatory rules. This observation is particularly important as it is to be assumed that often the discriminatory attribute will be missing, because, e.g., it might not be legally allowed to ask the race from the clients.

Future Collaborations: Discrimination-Aware Classifying with Background Information

An obvious area for collaboration lies at the intersection of the research conducted in [5] and in [3], namely the learning of a classifier without discrimination when the training data is biased and at the same time the attribute against which there is discrimination is not present in the dataset. This situation is extremely realistic and highly relevant; often, e.g., banks or credit scoring companies have large databases with highly valuable historic information, but with a large bias towards certain classes of people. Due to legal restrictions, in these datasets attributes such as ethnicity are not present. This extension of the work of [3] poses interesting problems that both groups jointly studied during the research visit:

1. What is the best way to represent the background knowledge? What is convenient, what is realistic to assume? Some ideas that came forward during the discussion was to either assume some tables relating few of the attributes in the dataset with the sensitive attribute. Such tables can be expected to come from, e.g., freely available census data.

2. How should we deal with inconsistent or contradicting background information? This will clearly be an issue; some attributes, e.g. postal code, might suggest that a certain person is black, while, e.g. his income and housing status might be highly correlated with white people. From the discussions the consensus arose that this is a well-researched area and that we should not go here; rather it is much more useful to take one of the off-the-shelf knowledge representation and reasoning frameworks, such as, e.g., Bayesian nets or probabilistic logics.

3. How can the identification of discrimination in certain contexts as identified in the work of [5] be incorporated in the classification without discrimination framework? A possible solution that emerged from the discussion was to enrich the dataset with an extra attribute, the sensitive attribute, as it can be derived from the background information. The uncertainty of the attribute will have to e taken into consideration. Some very concrete proposals based on a two-step reweighing approach were proposed.

Clearly, from the discussions, a wealth of ideas and future research directions have emerged. Joint collaboration on this topic has been started and will continue in the future, leading to joint publications on this topic.

Anonymity in Sequence Mining

Collaborators

Fosca Giannotti	CNR
Ruggero Pensa	CNR
Anna Monreale	CNR
Dino Pedreschi	University of Pisa
Toon Calders	Technische Universiteit Eindhoven

Problem Statement

A lot of data is available in the form of sequences. Examples of such data is trajectory data of cars. This data is often highly sensitive and revealing the complete data would harm the privacy of the people whose trajectories are in the data. Nevertheless, the data also has high economical value, e.g., for transportation firms optimizing their routes, decision makers planning traffic construction works, retail stores deciding on the location of a new retail point, etc. As such there is a tension between the privacy of the individual contributors to the data and the potential valuable public or commercial use of the data.

In the context of frequent itemset mining and also in classifier induction, this problem is well-studied. One of the approaches taken there is to slightly change the data in such a way that the resulting dataset does no longer allow for identifying individuals [7], but at the same time still contains the useful patterns. For sequence mining, however, the privacy-preserving data mining problem has not yet been studied. Due to the specific temporal nature of the sequence data, the techniques for itemsets, which are mainly based on randomly disturbing the data, cannot no longer be used.

Future Collaboration: Segmentation for Anonymization

Consider a database containing the following trajectory:

$$A \to B \to C \to D \to E \to F$$

It is very well possible that a part of the path is unique; consider, e.g., A and B are the locations of schools, C of a bakery and D of a factory; it might be the case that the trajectory belongs to a father, bringing his kids to their schools in the morning and passing by his favorite bakery for some bread for lunch lunch on the way to his job. It is not unlikely that such sequences are unique in the database. As such, when an adversary knows that this particular person followed the unique combination $A \to B \to C \to D$, he or she can derive what other places were visited by this person during that trajectory; namely E and F. As such, the privacy of the person is violated. Particular worrisome in this context is the fact that many retail stores are organized into large chains and at the same time use a system of loyalty cards tracking individuals visits. Combining such information with non-anonymized trajectory data might allow such retail stores to get very detailed information about the whereabouts of their customers. This information can on its turn be abused for direct marketing; e.g., people visiting certain sports clubs would get special promotions on sports equipement, and customers doing grocery shopping at a competitors supermarket would be especially targeted.

One very rough solution applied in practice to ensure privacy is removing all the connections between the different segments in the dataset. In the example above, this would result in splitting up the trajectory into the separate segments

$$A \to B, B \to C, C \to D, D \to E, E \to F \quad .$$

In this way, however, the usefulness of the data is seriously affected. Consider, e.g., a situation where the sub-trajectory $C \rightarrow D \rightarrow E \rightarrow F$ is very frequent. When cutting the trajectories into their constituent segments, questions like "If road constructions in the segment $D \rightarrow E$ take place, which routes will be affected?" cannot be answered anymore, whereas from the original data it is clear that the frequent path from C to F is affected. For this reason we do not consider such a rough division as a viable option.

One approach that was identified during the visit in Pisa is the so-called *segmentation approach*. The total sequence will be decomposed into smaller segments that are still large enough for identifying interesting frequent sequences. For this purpose, the trajectories need to be split in an optimal way. Optimal here means that, on the one hand, the privacy cannot be compromised. This is achieved by requiring that every of the subsequences has a minimal frequency k; in that way none of the segments is unique; there are always k people that followed the trajectory. On the other hand, the interesting patterns should be maintained, in the sense that their frequency should be affected as little as possible. These two conditions seem contradictory: the optimal situation for the first one is by cutting every sequence into segments of length one, whereas the optimum for the second one is keeping the database as is. The optimal point in between giving strong privacy guarantees while still allowing for frequent patterns to be found needs to be identified.

The splitting algorithm has two sets as parameter: one with the sequences that need to be broken because they violate privacy, and, on the other hand, those that need to be preserved because they are highly frequent. Some problems that are identified and need to be dealt with in future research:

1. How can a split into segments be evaluated? It seems plausible to use the number of frequent sequences that see their support decreased, as this represents the error made by splitting. Infrequent sequences should *never* be retained; as such, the final segments should all be frequent. Some additional remarks to take into consideration:

   (a) For illustrating that the problem is nontrivial: in many cases sequences can be split in multiple ways; e.g., consider the sequence $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ and let both $A \rightarrow B \rightarrow C$ and $C \rightarrow D \rightarrow E$ be frequent (and hence also their subsequences). Where should we split the sequence? After B or after C?

   (b) Can we use a dynamic programming approach to find the most optimal splitting, in the same spirit as in [2]? It seems plausible that if the optimal solution for a sequence

   $$A_1 \rightarrow \ldots \rightarrow A_i \rightarrow \ldots \rightarrow A_n$$

   is to split at $A_{i_1}, \ldots, A_{i_k}$, then the most optimal way to split

   $$A_1 \rightarrow \ldots \rightarrow A_i \rightarrow \ldots \rightarrow A_{i_\ell}$$

   is to split at $A_{i_1}, \ldots, A_{i_{\ell-1}}$, although this will depend on the definition of "optimal".

   (c) It might be good to take into account the relative error made on the support; a sequence with support 100 that becomes 99 is less severe than a sequence of support 10 that becomes 9; especially if, e.g., the minimal support threshold is 10. In such a case a split lowering the support to 9 should be avoided anyhow, probably. (Can we find examples where this is not possible?)

   (d) Sometimes a split might actually *increase* the support of a sequence; for example: splitting $A \rightarrow B \rightarrow C \rightarrow A \rightarrow B$ at C will increase the number of sequences supporting $A \rightarrow B$ with 1. Although this does not seem to be all too unreasonable, it is problematic from a definition point of view.

2. What about non-consecutive subsequences? In some settings (e.g., weblogs), it might make sense to split into segments that are not contiguous; e.g.,

$$A \to B \to C \to D$$

can be split into $A \to C$ and $B \to D$.

Learning to Rank Highly Unbalanced Tax Data

Collaborators

Laura Spinsanti	CNR and University of Pisa
Fosca Giannotti	CNR
Dino Pedreschi	University of Pisa
Toon Calders	Technische Universiteit Eindhoven
Faisal Kamran	Technische Universiteit Eindhoven

Problem Description

The motivation for this topic is the study performed at CNR on mining tax evaders based on their tax files. This dataset poses some highly original challenges:

1. Only a small part of the data is labeled; for most of the tax filings, it is not known if there is tax fraud present or not.

2. The part of the data that is labeled was carefully selected, so the training data is to be expected to have a different distribution than the overall data that needs to be classified. This bias in combination with the fact that many tax filings are unlabeled implies that it is difficult to predict the performance of a learned model in a real-life setting.

3. The class of "really bad" tax-evaders that needs to be distinguished is very small.

4. The predictions should be extremely accurate; the ratio of false positives should be extremely low.

5. The labeled data comes from different sources (regional tax offices). It is to be expected that different criteria were applied in different regions. A learned model should not discriminate between the

For this particular dataset, a method was developed by Spinsanti, Giannotti and Pedreschi, which is currently still under evaluation.

Potential Collaborations

We would like to see if it is possible to apply the methods that were developed in the context of the discrimination-aware data mining for this type of problems. In short, we would consider all unlabeled examples as being "fraude-negative" and add them with this label to the set of training examples. Subsequently, we would try to learn a classifier that does not discriminate between tuples that were in the dataset from the start, and the ones with an artificial label. In this way, there should be an equal (or any fixed ratio by the user) portion of fraud-predictions in the originally unclassified data as in the originally classified data. Doing the data massaging a couple of times, may lead to a better classifier as the overall distribution is taken into account. There are also interesting links with transfer learning [6], co-training [1] and similar approaches in text mining [4], to name some, that need to be explored as well.

References

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM New York, NY, USA, 1998.

[2] Niina Haiminen, Aristides Gionis, and Kari Laasonen. Algorithms for unimodal segmentation with applications to unimodality detection. *Knowl. Inf. Syst.*, 14(1):39–57, 2008.

[3] F. Kamiran and T. Calders. Classification without discrmination. manuscript, Eindhoven University of Technology, 2008.

[4] K. Nigam, A.K. Mccallum, S. Thrun, and T. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2):103–134, 2000.

[5] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[6] R. Raina, A.Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 713–720. ACM New York, NY, USA, 2006.

[7] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, pages 1–19, 1998.