Interactive Data Exploration Using Pattern Mining

Matthijs van Leeuwen

Machine Learning group
KU Leuven, Leuven, Belgium
matthijs.vanleeuwen@cs.kuleuven.be

Abstract. We live in the era of data and need tools to discover valuable information in large amounts of data. The goal of exploratory data mining is to provide as much insight in given data as possible. Within this field, pattern set mining aims at revealing structure in the form of sets of patterns. Although pattern set mining has shown to be an effective solution to the infamous pattern explosion, important challenges remain.

One of the key challenges is to develop principled methods that allow user- and task-specific information to be taken into account, by directly involving the user in the discovery process. This way, the resulting patterns will be more relevant and interesting to the user. To achieve this, pattern mining algorithms will need to be combined with techniques from both visualisation and human-computer interaction. Another challenge is to establish techniques that perform well under constrained resources, as existing methods are usually computationally intensive. Consequently, they are only applied to relatively small datasets and on fast computers.

The ultimate goal is to make pattern mining practically more useful, by enabling the user to interactively explore the data and identify interesting structure. In this paper we describe the state-of-the-art, discuss open problems, and outline promising future directions.

Keywords: Interactive Data Exploration, Pattern Mining, Data Mining.

1 Introduction

We live in the era of data. Last year it was estimated that 297 exabytes of data had been stored, and this amount increases every year. Making sense of this data is one of the fundamental challenges that we are currently facing, with applications in virtually any discipline. Manually sifting through large amounts of data is infeasible, in particular because it is often unknown what one is looking for exactly. Therefore, appropriate tools are required to digest data and reveal the valuable information it contains.

Although data is everywhere, it is not unusual that the domain experts who have access to the data have no idea what information is contained in it. KDD, which stands for *Knowledge Discovery in Data*, aims to extract knowledge from data. In particular, the goal of the field of *exploratory data mining* is to provide

A. Holzinger, I. Jurisica (Eds.): Knowledge Discovery and Data Mining, LNCS 8401, pp. 169–182, 2014.

a domain expert as much insight in given data as possible. Although inherently vague and ill-defined, it aims to provide a positive answer to the question: *Can you tell me something interesting about my data?*

As such, its high-level aim is similar to that of *visual analytics*, but the approach is rather different. Whereas visual analytics focuses on visualization in combination with human-computer interaction to improve a user's understanding of the data, exploratory data mining focuses on finding models and patterns that explain the data. This results in (typically hard) combinatorial search problems for which efficient algorithms need to be developed. Depending on the problem and the size of the data, exact or heuristic search is used.

Pattern Mining. Within exploratory data mining, *pattern mining* aims to enable the discovery of patterns from data. A *pattern* is a description of some structure that occurs locally in the data, i.e., it describes part of the data. The best-known instance is probably frequent itemset mining [1], which discovers combinations of 'items' that frequently occur together in the data. For example, a bioinformatician could use frequent itemset mining to discover treatments and symptoms that often co-occur in a dataset containing patient information.

A pattern-based approach to data mining has clear advantages, in particular in an exploratory setting. One advantage is that patterns are interpretable representations and can thus provide explanations. This is a very desirable property, and is in stark contrast to 'black-box' approaches with which it is often unclear why certain outcomes are obtained. A second large advantage is that patterns can be used for many well-known data mining tasks.

Unfortunately, obtaining interesting results with traditional pattern mining methods can be a tough and time-consuming job. The two main problems are that: 1) humongous amounts of patterns are found, of which many are redundant, and 2) background knowledge of the domain expert is not taken into account. To remedy these issues, careful tuning of the algorithm parameters and manual filtering of the results is necessary. This requires considerable effort and expertise from the data analyst. That is, the data analyst needs be both a domain expert and a data mining expert, which makes the job extremely challenging.

Pattern Set Mining. As a solution to the redundancy problem in pattern mining, a recent trend is to mine pattern *sets* instead of individual patterns. The difference is that apart from constraints on individual patterns, additional constraints and/or an optimisation criterion are imposed on the complete set of patterns. Although *pattern set mining* [2] is a promising and expanding line of research, it is not yet widely adopted in practice because, like pattern mining, directly applying it to real-world applications is often not trivial.

One of the main issues is that the second problem of pattern mining has not yet been addressed: background knowledge of the domain expert is not taken into account. Because of this, algorithms still need to be tuned by running the algorithm, waiting for the final results, changing the parameters, re-running, waiting for the new results, etc. Most existing methods can only deal with interestingness measures that are completely *objective*, i.e., interestingness of a pattern or pattern set is computed from the data only.