The Data Problem in Data Mining

Albrecht Zimmermann

The Data Problem in Data Mining

Albrecht Zimmermann

INSA de Lyon
albrecht.zimmermann@insa-lyon.fr

Abstract. Computer science is essentially an applied or engineering science, creating tools. In Data Mining, those tools are supposed to help humans understand large amounts of data, and produce actionable insight. In this talk, I argue that for all the progress that has been made in Data Mining, in particular Pattern Mining, we are lacking understanding of key aspects of the performance and results of pattern mining algorithms. I will focus particularly on the difficulty of deriving actionable knowledge from patterns. I trace the lack of progress regarding those questions to a lack of data with varying, controlled properties, and argue that we will need to make a science of digital data generation, and use it to develop guidance to data practitioners.

1 Short-comings in evaluation

Data Mining, and in particular Pattern Mining, have been around for about two decades and the work in the field has led to a large number of techniques, which have been applied to pattern domains as diverse as itemsets, attribute-value data, sequences, trees, and graphs, and tasks ranging from finding associations to describing interesting subpopulations, to predicting unseen class labels.

In this talk, I will focus on the unsupervised pattern mining setting, i.e. finding unexpected, interesting and useful patterns that are not related to a variable of interest - nominal or otherwise. As I will argue, the *qualitative* evaluation of proposed techniques, i.e. how "good" the resulting patterns are, has been given short thrift in comparison to *quantitative* evaluation, i.e. how efficiently the output is found.

But also the latter has arguably not been given the attention it deserved. This case has been made convincingly early on by Zheng *et al.* [2], who showed that the evaluations performed in itemset mining up to that point in time had led to an over-fitting on the artificially generated data used. The reported performance did not transfer to real-life data, which showed different characteristics than the artificially generated data. Remarkably enough, the situation has barely improved since then, with quantitative evaluations focused on a small number of data sets, of which typically only few are used in a given evaluation.

The situation is worse for qualitative evaluations, which are rarely performed in the first place. This is understandable since the lack of a target variable corresponds to missing ground truth in the data. But at the same time, it means that even if we knew how to set parameters appropriately[1], we would not know

[1] Another area in which there is too little guidance.

how found patterns relate to the processes that generated the data. Since pattern mining is supposed to give us insight into those processes, and allow us to act based on found patterns, this is a serious short-coming.

2 Generating data (and understanding pattern mining)

When there is no ground truth available for real-life data (or when there is little real-life data available in the first place), generating artificial data is a promising alternative. This is not only the case in computer science, where, for instance, the SAT solving community has chosen this direction, but also in "hard sciences" like physics, see for instance [1].

Data generation allows us to both break the bottleneck of too few data sets (or data sets with a too narrow range of characteristics), and to understand how found patterns relate to the processes that generated the data. As Zheng *et al.* showed, however, and others have demonstrated since, approaching this task without forethought and an understanding of the data we aim to generate will lead to unrealistic data sets. Furthermore, limiting ourselves to a narrow selection of generative processes, e.g. generating itemset mining data only by combining itemsets, will restrict the lessons to be learned from matching patterns to processes, and carries the risk of biasing qualitative evaluations.

Fortunately, we do not have to start from scratch. More-or-less successful attempts at data generation have been made, and some infrastructure exists to support this task. Additionally, some researchers have attempted to relate patterns to different processes to evaluate their quality, especially in recent years. Finally, researchers and practitioners in other fields have developed theories of their own that, while necessarily taken with a grain of salt, can be built on to simulate real-life processes. By combining and building on this existing knowledge, we can fill in the current data gaps and start to understand those aspects of pattern mining that escape us so far.

References

1. Cern software development for experiments - Simulation. http://ph-dep-sft.web.cern.ch/project/simulation
2. Zheng, Z., Kohavi, R., Mason, L.: Real world performance of association rule algorithms. In: KDD. pp. 401–406 (2001)