Making Pattern Mining Useful

Jilles Vreeken*
Department of Information and Computing Sciences
Universiteit Utrecht
Utrecht, the Netherlands
http://www.cs.uu.nl/groups/ADA/people/vreeken/
jillesv@cs.uu.nl

1. INTRODUCTION

Generally speaking, finding a pattern is easy. Discovering interesting patterns, that's where things get complicated. My dissertation[1] is about finding interesting patterns, and, more boldly, about making pattern mining useful. It is about how to discover few, but highly interesting patterns. And, prominently, it is about how to put these patterns to good use, solving a number of data mining problems.

But, before we discuss the actual content, let us first informally discuss pattern mining and identify why it is not yet as useful as it could be.

2. PATTERN MINING IN ACTION

Let us consider an example how pattern mining could be used, e.g. in medicine, for gaining insight in the causes of a particular disease. Normally, following the scientific method, a doctor would build a hypothesis. In other words, a pattern. For this hypothesis not to be a shot in the dark, the doctor needs to be able to oversee the symptoms, behaviours, etc, that the patients exhibit. That is, he or she must be able to 'see' the pattern. The hypothesis can then be tested, and so shown to be correct or not.

This works very well, up till the point where the problem at hand becomes too complicated, when there exists such a gigantic number of possible combinations of causes, that it becomes impossible for the doctor to gain sufficient overview.

Yet, in those situations, we can apply pattern mining to discover important regularities. We simply mine for patterns in the gathered data, and return those that pass certain interestingness criteria. In this case, such a pattern could be a combination of factors with a strong relation to the disease. The doctor then selects the most promising patterns, and uses these to build a promising hypothesis.

So far, so good. However, not in practice, as the poor doctor will be swamped in patterns. From being unable to overview the data, the problem becomes that it is impossible to overview the potentially interesting patterns. Perhaps even worse, many of the discovered patterns are variations of the same theme, and convey the same information.

3. MAKING PATTERN MINING USEFUL

So, while pattern mining holds great promise, I dare say it collapses under its own weight: it finds patterns too easily. This is particularly due to the difficulty in using interestingness measures in practice: if we set the constraints too tight, only few but commonly known patterns are returned; and when we use more loose constraints the pattern explosion occurs, and we are overwhelmed by the number of results.

While patterns can clearly provide useful insight, finding just those patterns that are interesting is a question yet unanswered. The sheer amount of results makes it virtually impossible to let human experts, such as our doctor, interpret the results. Further, it prohibits pattern mining, and the detail provided by the discovered patterns, to be practically and more generally applied in data mining.

What these problems come down to, is that we are asking the wrong question. While we ask for *all* patterns that satisfy the conditions, at the same time we actually only want to have a small set of the best patterns.

This dissertation therefore proposes a different approach. We do not want to find all patterns in a database, or try to summarise those collections of patterns. Instead, we want small, non-redundant, sets of high-quality patterns that summarise the *data* well, i.e. patterns that describe the data. The resulting groups should be small enough to be analysed by an expert such as our doctor and provide a detailed overview of the data.

In my dissertation the problem of mining sets of patterns is approached through the Minimum Description Length principle (MDL) [1]. That is, by lossless compression. Intuitively, we can say that the better a set of patterns compress the data, the better it captures the regularities in the data. By MDL we define the best set of patterns as the set of patterns that compresses the data best.

One could ask, why would our doctor be interested in patterns that compress? Quite simply, because these are the patterns that matter. Because MDL takes the complexity of the selected patterns into account, we know that redundant and spurious patterns will be be eliminated. In other words, the doctor will find that the best compressing patterns to be a small group that together describes the data very well, without redundancy and noise. These two aspects make these patterns *useful*, i.e. they cannot only be evaluated by an expert such as the doctor, but also naturally be applied to solve various data mining problems.

The dissertation includes five such applications, including measuring and characterising differences between databases, finding blocks of data with similar characteristics, and esti-

*Currently, the author is a post-doctoral researcher in the ADReM group of University of Antwerp, Belgium.

[1] This dissertation was selected as the runner-up for the 2010 ACM SIGKDD Ph.D. Dissertation Award. The complete version can be downloaded from:
http://www.cs.uu.nl/groups/ADA/people/vreeken/

mating the missing values for data with incomplete records; all problems often faced by our doctor. These are all naturally approached through the MDL-principle. However, it is the level of detail captured in the pattern sets that makes the difference, allowing for both high performance *and* immediate characterisation of the why.

As such, the research objective of this dissertation was phrased by its title, **Making Pattern Mining Useful**. This goal included developing techniques for finding small groups of high-quality patterns, showing these provide insight, and can be used to solve open data mining problems.

4. OUTLINE OF THE DISSERTATION

My dissertation is divided into nine chapters. Chapters 2 to 8 are edited versions of work published during the course of my studies. The topics treated are summarised as follows.

First, we propose to use the Minimum Description Length principle to select small groups of frequent itemsets that describe the data well. To this end, we introduce Krimp; a heuristic parameter-free algorithm for finding the optimal set of frequent itemsets [4]. Through extensive evaluation, amongst which through the Krimp-classifier [2], we show the high quality of these *code tables*.

Next, we show how one can measure and characterise the difference between transaction databases. Difference is measured by calculating the relative Krimp compressibility of the data; code tables allow detailed characterisation of the difference, with three levels of increasing detail [7].

Following, we give two algorithms, by which one can identify and characterise the components of a database. Data is split into homogeneous blocks, such that the compression is optimised—without requiring a distance measure. The methods are orthogonal in approach: one is data-driven, while the second extracts components from Krimp models [3] (ECMLPKDD'09 best student paper).

In Chapter 5 we discuss how code tables, while mined as descriptive models, can also be used as generative models [8]. We introduce an algorithm that generates data that is virtually indistinguishable from the original. We show the use for this in privacy-preserving data mining as an alternative to data perturbation, as our method provides anonymised data with all important patterns intact.

We further onto the generative path in Chapter 6, and introduce three algorithms for completion of data with missing values. All three follow the MDL principle, i.e. the completed database that can be compressed best, is the best completion. As an objective test we propose (ϵ, δ)-correctness to measure the difference between two databases in terms of count statistics. Experimentation shows our pattern-based approach to be superior to the current state of the art, both in terms of accuracy and count statistics [6].

Chapter 7 extends the concept of selecting patterns by MDL to low-entropy sets. The algorithm we introduce, LESS, selects very small collections: typically only tens of patterns. These small numbers, and the interpretability of the patterns, facilitate thorough analysis. By using entropy instead of frequency, LESS is particularly suited for mining dense data. Further, by regarding data 0/1 symmetric, all major interactions are captured, not just co-occurrences.

Last, but not least, we introduce Pack [5], an algorithm for selecting itemsets through refined MDL. It employs decision trees to compress data 0/1 symmetrically and attains high compression-ratios. Besides selection of itemsets from large collections, it can also discover models directly from data.

5. CONCLUDING REMARKS

The conclusions the dissertation draws are straightforward. To the end of making pattern mining useful, the best *set of patterns* should be mined, as opposed to *all* patterns that satisfy certain criteria. The MDL principle is particularly well-suited for mining useful patterns; by using this principle to select the set of patterns that describe the data best, we are returned very few, but high-quality, patterns.

Moreover, MDL is a natural approach to many data mining problems. We have shown, by stating a variety of problems in terms of MDL, that very high performance can be attained. In particular, we have shown that the patterns returned by Krimp are generally applicable, provide high performance and insightful characterisation.

Many of the addressed problems, such as incomplete records and unknown dissimilarity, are often faced by our doctor: it is therefore safe to say that the sets of patterns MDL identifies are indeed useful.

Summarising, the results presented in this dissertation will make our doctor very happy. However, the ulterior peak of usefulness is not yet reached: pattern mining can be made *even more useful*. Using MDL to mine sets of patterns comes with many new challenges and opportunities. Addressing these issues will make for interesting future research that can further increase the usefulness of pattern mining.

Ph.D. Dissertation Committee

Arno Siebes (advisor), Heikki Mannila, Johannes Fürnkranz, Jean-François Boulicaut, Peter Grünwald, Linda van der Gaag, Toon Calders, and Ad Feelders.

6. REFERENCES

[1] P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[2] M. van Leeuwen, J. Vreeken, and A. Siebes. Compression picks the item sets that matter. In *Proceedings of ECML PKDD'06*, pages 585–592, 2006.

[3] M. van Leeuwen, J. Vreeken, and A. Siebes. Identifying the components. *Data Min. Knowl. Discov.*, 19(2):173–292, 2009.

[4] A. Siebes, J. Vreeken, and M. van Leeuwen. Item sets that compress. In *Proceedings of SIAM SDM'06*, pages 393–404, 2006.

[5] N. Tatti and J. Vreeken. Finding good itemsets by packing data. In *Proceedings of ICDM'08*, pages 588–597, 2008.

[6] J. Vreeken and A. Siebes. Filling in the blanks – Krimp minimisation for missing data. In *Proceedings of ICDM'08*, pages 1067–1072, 2008.

[7] J. Vreeken, M. van Leeuwen, and A. Siebes. Characterising the difference. In *Proceedings of KDD'07*, pages 765–774, 2007.

[8] J. Vreeken, M. van Leeuwen, and A. Siebes. Preserving privacy through data generation. In *Proceedings of ICDM'07*, pages 685–690, 2007.