VIPER – Visual Pattern Explorer

Matthijs van Leeuwen (✉) and Lara Cardinaels

Department of Computer Science, KU Leuven, Belgium
matthijs.vanleeuwen@cs.kuleuven.be

Abstract. We present VIPER, for *Visual Pattern Explorer*, an innovative, browser-based application for interactive pattern exploration, assisted by visualisation, recommendation, and algorithmic search. The target audience consists of domain experts who have access to data but not to –potentially expensive– data mining experts. The goal of the system is to enable the target audience to perform true *exploratory* data mining. That is, to discover interesting patterns from data, with a focus on subgroup discovery but also facilitating frequent itemset mining.

1 Introduction

Pattern mining concerns the discovery of patterns in data, where a pattern is a succinct description of some structure that occurs locally in the data. Two typical pattern mining tasks are frequent itemset mining (FIM) and subgroup discovery (SD). The task of FIM is to find combinations of items, such as products in retail data, that frequently occur together. The task of SD is to find subsets of the data for which a certain target of interest has a deviating distribution. Given a dataset containing movie ratings, for example, one might find that women tend to give higher ratings to movies with genre 'drama' than men.

Many interestingness measures and even more algorithms have been proposed for both tasks. See [1] and [2] for recent overviews. One of the key challenges of pattern mining is the infamous *pattern explosion*, i.e., the huge numbers of patterns that are commonly found by pattern mining algorithms. A recent trend to alleviate this issue is to mine pattern sets [5] rather than individual patterns.

Unfortunately, existing systems are hard to use due to the many parameters and other choices that have to be made. As a result, pattern mining as a technology is only accessible to data mining experts, strongly limiting its potential impact. The average person who has access to data and wants to analyse it, whether it is in academia or industry, is a domain expert, *not* a data mining expert. She knows what patterns she finds interesting and doesn't need someone to determine this, but she does need help in finding these interesting patterns.

For exactly that purpose we present VIPER, for *Visual Pattern Explorer*, an innovative, browser-based application for exploring the space of all possible patterns, assisted by visualisation, recommendation, and algorithmic search. The target audience are domain experts who have access to data but not to –potentially expensive– data mining experts. The goal of the system is to enable the target audience to discover interesting patterns from data, with a focus on subgroup discovery but also facilitating frequent itemset mining.

2 Related work

Throughout the years several pattern mining systems with a graphical user interface have been developed, such as MIME [8] for FIM and Cortana[1] for SD. These systems, however, first mine a (large) number of patterns and then give the user the opportunity to browse this collection; search and interaction are completely decoupled. These tools are generally inaccessible to domain experts, our target audience, due to the large number of algorithms, measures, and parameters.

More recently, interestingness measures have been investigated that can adapt to the background knowledge and/or feedback of a user. Bhuiyan et al. [3] proposed to use user feedback to adapt the sampling distribution of itemsets. Dzyuba et al. [7] proposed to learn pattern rankings using techniques from preference learning. Orthogonally, De Bie [6] focused on a theoretic framework for iterative data mining and a formalisation of subjective interestingness. None of these works presented a working system though. Boley et al. [4] did present a system for 'one-click-mining', in which the preferences of the user for certain algorithms and patterns are learned. Still, objective interestingness measures are used to mine patterns, which are then presented to the user.

3 VIPER – Visual Pattern Explorer

We present VIPER, a browser-based application that allows the user to explore the pattern space, assisted by visualisation, recommendation, and search. Given that the target audience are domain experts rather than data mining experts, one of the primary design goals is to keep it as simple as possible.

Implementation The web application is publicly available[2] and has been implemented in JavaScript and runs locally in the browser. It has been developed and tested using Chrome, making it available on many platforms.

Features The features of our initial implementation are focused on the application's core functionality, i.e., visually assisted pattern exploration.

Measures To avoid overwhelming the user with a plethora of interestingness measures, VIPER uses just two: 1) coverage (a.k.a. support or frequency) for itemsets and 2) weighted relative accuracy (WRAcc) for subgroups.

Data VIPER can be used for the analysis of categorical datasets (in a simple text format). This allows to use bitvector computations for many operations, making the application respond nearly instantly for datasets of moderate size.

Pattern exploration Once a dataset has been loaded, the user can start exploring the pattern space; see Figure 1. A key principle is that the user is always in control. At any time, the current pattern can be specialised (extended) by clicking an attribute-value combination in one of the charts, or be generalised by removing an attribute-value combination by clicking on it.

The charts indicate all possible specialisations: each chart represents an attribute, and each bar in each chart a possible value for that attribute. The bar

[1] `datamining.liacs.nl/cortana.html`

[2] Demo and source code at `www.patternsthatmatter.org/viper`

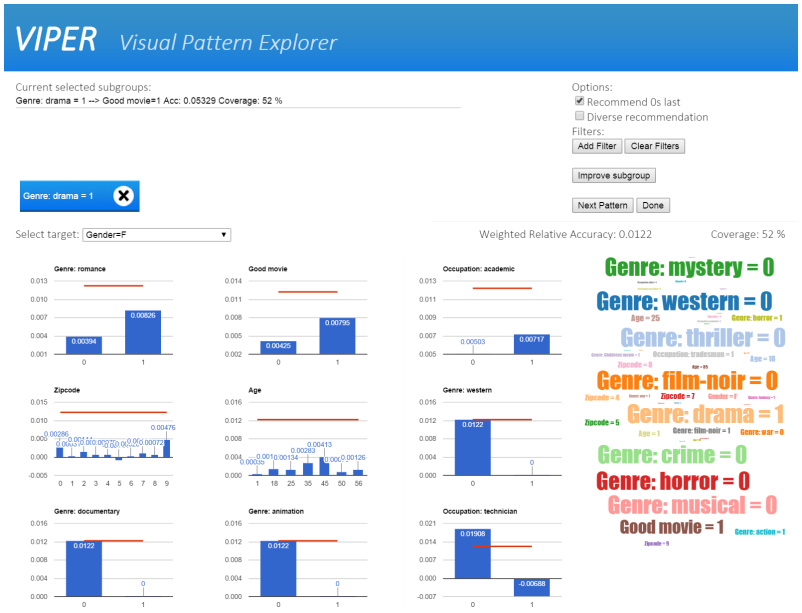

Fig. 1. Screenshot of VIPER. Example using movie rating data.

heights represent the interestingness that is obtained by the respective specialisations. For reference the horizontal red lines show the current interestingness. For SD, a target can be chosen (and changed!) at any time; WRAcc is used as interestingness measure when a target is chosen, otherwise coverage is used. To assist the user to assess patterns, a 'word cloud' shows those attribute-values that occur more frequently within the pattern than in the remainder of the data.

Specialisation recommendation To enable the user to efficiently explore the pattern space, by default all specialisations are ordered by attainable interestingness descending. Moreover, several filtering and specialisation recommendation options are offered. For example, text-based filters can be applied to exclude specialisations from the view. Finally, datasets tend to contain strongly correlated attributes, which may result in very similar specialisations ranked first. To remedy this, VIPER offers a diverse recommendation option, which is based on the cover-based beam selection procedure from the DSSD algorithm [9].

Greedy search A user may be interested in quickly finding the best possible specialisation of the current pattern. To facilitate this, we implemented a greedy hillclimbing search strategy ('Improve subgroup') that iteratively specialises a given pattern until weighted relative accuracy cannot be improved.

Iterative mining Once the user has found a pattern she likes, she can move to the next pattern or end the session. In the former case, the pattern is stored and the user can opt to apply sequential covering, meaning that the data selected by the pattern is removed and exploration continues on the remainder of the data. After finishing, the final set of patterns can be inspected and exported.

3.1 Preliminary user study: movie ratings

We conducted a preliminary user study to evaluate whether domain experts with little to no knowledge of pattern mining can successfully use VIPER. To this end, asked ten test users (CS students) to answer six questions about a movie rating dataset[3], e.g., "*Who like movies of the genre comedy?*". Half of the users got to use VIPER, the other half used Cortana. Answers were manually checked.

On average, VIPER users answered 5.4 out of 6 questions correctly (in 16.4 mins), Cortana users only 2.25 (in 26.5 mins). The responses to a usability questionnaire (USE) were also clearly in favour of VIPER. Of course, no definitive conclusions can be drawn from these preliminary results. However, it does demonstrate that a tool that gives full control to the user can be powerful for answering simple data analysis questions.

4 Conclusions

We presented VIPER, a web-based application for interactive data exploration. Contrary to existing pattern mining systems, VIPER does not mine a huge number of patterns and then lets the user sift through them, but instead offers the user the tools needed to explore and discover those patterns that the user finds interesting. This initial version of VIPER is only a first step towards integrating visually assisted data exploration with pattern mining and we hope to extend it in the future. But before that, we hope to get feedback from the community.

Acknowledgments Matthijs van Leeuwen is supported by a Postdoctoral Fellowship of the Research Foundation Flanders (FWO).

References

1. Aggarwal, C., Han, J. (eds.): Frequent Pattern Mining. Springer (2014)
2. Atzmueller, M.: Subgroup discovery. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery 5(1), 35–49 (2015)
3. Bhuiyan, M., Mukhopadhyay, S., Hasan, M.A.: Interactive pattern mining on hidden data: a sampling-based solution. In: Proc. of CIKM'12. pp. 95–104 (2012)
4. Boley, M., Mampaey, M., Kang, B., Tokmakov, P., Wrobel, S.: One click mining: Interactive local pattern discovery through implicit preference and performance learning. In: Proceedings of IDEA'13. pp. 27–35. ACM, New York, NY, USA (2013)
5. Bringmann, B., Nijssen, S., Tatti, N., Vreeken, J., Zimmermann, A.: Mining sets of patterns: Next generation pattern mining. In: Tutorial at ICDM'11 (2011)
6. De Bie, T.: An information theoretic framework for data mining. In: Proceedings of KDD'11. pp. 564–572 (2011)
7. Dzyuba, V., van Leeuwen, M., Nijssen, S., Raedt, L.D.: Interactive learning of pattern rankings. International Journal on Artificial Intelligence Tools 23(6) (2014)
8. Goethals, B., Moens, S., Vreeken, J.: MIME: a framework for interactive visual pattern mining. In: Proceedings of KDD'11. pp. 757–760 (2011)
9. van Leeuwen, M., Knobbe, A.: Diverse subgroup set discovery. Data Mining and Knowledge Discovery 25, 208–242 (2012)

[3] `grouplens.org/datasets/movielens/`