Preprint, final version is available at http://dx.doi.org/10.1007/978-3-642-33486-3_52

Association Rule Mining Following the Web Search Paradigm

Radek Škrabal, Milan Šimůnek, Stanislav Vojíř, Andrej Hazucha, Tomáš Marek, David Chudán, Tomáš Kliegr

Department of Information and Knowledge Engineering, University of Economics, Nám. Winstona Churchilla 4, Prague 3, 130 67, Czech Republic
{xskrr06|simunek|stanislav.vojir|andrej.hazucha|xmart17|david.chudan|tomas.kliegr}@vse.cz

Abstract. I:ZI Miner (sewebar.vse.cz/izi-miner) is an association rule mining system with a user interface resembling a search engine. It brings to the web the notion of interactive pattern mining introduced by the MIME framework at ECML'11 and KDD'11. In comparison with MIME, I:ZI Miner discovers multi-valued attributes, supports the full range of logical connectives and 19 interest measures. A relevance feedback module is used to filter the rules based on previous user interactions.

1 Introduction

The goal of the association rule mining task is to discover patterns interesting for the user in a given collection of objects. The user interest is defined through a set of features that can appear on the left and right side of the discovered rules and by thresholds on selected interest measures. A typical task produces many rules that formally match these criteria, but only few are interesting to the user [1]. The uninteresting rules can be filtered using domain knowledge; the challenge is to balance the investment of user's time to provide the required input with the utility gained from the filtered mining result.

To address this challenge, we draw inspiration from the information retrieval task, which tackles a similar problem – select documents interesting to the user from the many that match the user's query. Web search engines are successful in retrieving subjectively interesting documents from their index; with I:ZI Miner[1] we try to apply the underlying principles to the task of discovering subjectively interesting rules from the dataset. We consider these principles to be interactivity, simplicity of user interface, immediate response and relevance feedback.

I:ZI Miner is based on similar ideas as the MIME framework [2], a desktop application introduced at ECML 2011 and KDD 2011. The main differences are that I:ZI Miner works with *multi-valued attributes* (see examples in Fig. 1) and supports the *full range of logical connectives*. Additional features that distinguish I:ZI Miner from MIME include *web interface*, *rule filtering* based on relevance feedback and a *preprocessing module*.

[1] The name I:ZI Miner is pronounced as 'easy miner'.

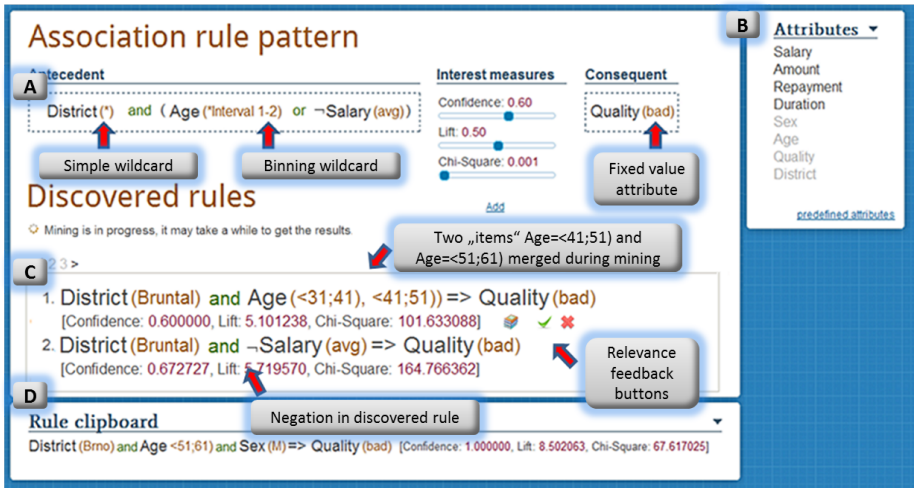

Fig. 1. I:ZI Miner screenshot

2 User Interface

The mining task is defined in the *Pattern Pane* (Fig. 1A) by selecting interest measures and placing attributes from the *Attribute Palette* (Fig. 1B). Real-time results are shown in the *Result Pane* (Fig. 1C), which also serves for relevance feedback. Interesting rules are saved to the *Rule Clipboard* (Fig. 1D).

Pattern Pane. By dragging attributes from the Attribute palette to the antecedent and consequent of the rule, the user creates an intuitively understandable 'rule pattern' that discovered rules must match. Attributes are by default connected by conjunction, but it can be changed to disjunction. For a specific attribute, the user can either select a value or a wildcard (ref. to Sec. 3). There is also the option to put negation on an attribute. The user adds at least one interest measure and its threshold. A unique feature is that any combination of the available 19 interest measures can be used.

Result Pane. It displays the rules as soon as they have been discovered, i.e. the user does not have to wait for the mining process to finish to get the first results. If the discovered rule is only a confirmation [4] of a known rule, it is visually suppressed by gray font. In contrast, exception to a known rule is highlighted in red. Uninteresting rules that pass the domain knowledge check can be discarded by clicking on the red cross; this stores negative relevance feedback. Green tick moves the rule to the Rule clipboard and stores positive relevance feedback.

Attribute Palette. For an attribute to be used during mining, the user needs to drag it to the Pattern pane. If the 'best pattern' feature is on, the attributes are ordered according to their estimated value for predicting the consequent.

3 Architecture, Performance and Expressiveness

The software has a web service architecture; an extension of the PMML format (`www.dmg.org`) is used for communication between individual modules. The mining module is reused from the LISp-Miner system (`lispminer.vse.cz`). It is written in C++, uses a proprietary bitstring-based mining algorithm derived from the GUHA method [3], and offers a range of unique features including:

- Arbitrary combination of interest measures, including confidence, support, lift, Chi-square, Fisher and eight other interest measures developed within the GUHA method.
- Full range of logical connectives (conjunction, disjunction and negation).
- *Simple wildcard* on an attribute tells the miner to generate as many 'items' as there are values of the attribute. This is similar to other association rule mining systems that support multiple attributes.
- Adding a *Fixed value attribute* to the mining setting allows the user to limit the search space only to rules containing a selected attribute-value pair.
- *Binning wildcards* can be used to instruct the miner to dynamically merge multiple values into one 'item' during mining. If value order was specified in the preprocessing stage, only adjacent values can be binned.
- Support for distributed computing on top of the Techila platform [6].

The relevance feedback module [4] is a Java application running on top of the XML Berkeley database. I:ZI Miner is part of the SEWEBAR-CMS project [5], which provides data preprocessing and reporting capabilities.

4 Comparison with the MIME Framework

Our system shares many features with the MIME framework introduced in last year's ECML and KDD conferences. In this section we will list the features the two system share, the additional features of I:ZI Miner in comparison with MIME, and the features of MIME not implemented in I:ZI Miner, in turn.

Both systems have a 'best pattern' extension. While MIME orders *items* according to their impact on the existing mined pattern, I:ZI Miner implements a heuristic algorithm which orders *attributes* according to their estimated impact.[2] Another common feature is the ability to define groups of items which are considered as one value during mining. While in MIME these groups are defined manually, I:ZI Miner features multiple types of binning wildcards.

I:ZI Miner unique features include mining over multi-valued attributes, negation and disjunction in rules, wider choice of interest measures and filtering of discovered rules with relevance feedback.[2] Technologically, I:ZI Miner is a web application. Through integration with SEWEBAR-CMS it offers a preprocessing and reporting capability. Concerning scalability, mining runs as a web service with the underlying grid platform giving an option to upscale to Microsoft Azure.

Rule post-processing and visualization algorithms, on the other hand, are only implemented in MIME.

[2] Its technical description cannot be included for space reasons.

5 Screencasts and Demo

Screencasts and a live demonstration of I:ZI Miner on the ECML PKDD'99 Financial dataset are available at `sewebar.vse.cz/izi-miner`.

Screencast 1: Explorative Task featuring Binning Wildcards. In the exploration mode the user investigates the relationships in the dataset without being bound to a specific outcome. Unlike other association rule mining systems that operate on binary items, in I:ZI Miner the task is defined directly on multi-valued attributes (Fig. 1A). For attributes having many values with low support. I:ZI Miner offers a unique feature – binning wildcards, which allow to group fine-grained values on the fly, thus producing 'items' with higher support. Fig. 1C shows a rule with two values of the Age attribute grouped by a binning wildcard.

Screencast 2: Predictive Task featuring the Best Pattern Extension. In the prediction mode the user has a specific outcome in mind and wants to explore novel combinations of attribute values that are associated with this outcome. The user is aided by the best pattern extension, which orders the attributes in the Attribute Pane (Fig. 1B) according to their estimated impact on the results if added to the definition of the task in the Rule Pattern Pane.

Screencast 3: Data Preprocessing featuring Automatic Binning. Despite the availability of binning wildcards, it is more efficient to decrease the dimensionality of the attribute space in the preprocessing phase. Every manually specified binning is saved as a transformation scenario. In the automatic mode attributes in the dataset are compared with the saved scenarios using algorithms from the schema matching domain. If there is a sufficient match, the new attribute is binned according to a scenario defined earlier for a similar attribute.

Acknowledgements The work described here was supported by grants IGA 26/2011, GACR 201/08/0802 of the Czech Science Foundation and the EU FP7 grant no. 287911, LinkedTV project. We thank Vojtěch Svátek for his feedback.

References

1. Tijl De Bie, Kleanthis-Nikolaos Kontonasios, and Eirini Spyropoulou. A framework for mining interesting pattern sets. In *Useful Patterns*, UP '10, pages 27–35, New York, NY, USA, 2010. ACM.
2. Bart Goethals, Sandy Moens, and Jilles Vreeken. MIME: a framework for interactive visual pattern mining. ECML PKDD'11, pages 634–637, Berlin, 2011. Springer.
3. Petr Hájek, Martin Holeňa, and Jan Rauch. The GUHA method and its meaning for data mining. *Journal of Computer and System Sciences*, 76:34–48, 2010.
4. Tomáš Kliegr, Andrej Hazucha, and Tomáš Marek. Instant feedback on discovered association rules with PMML-based query-by-example. In *Web Reasoning and Rule Systems*. Springer, 2011.
5. Tomáš Kliegr, Vojtěch Svátek, Milan Šimůnek, and Martin Ralbovský. Semantic analytical reports: A framework for post-processing of data mining results. *Journal of Intelligent Information Systems*, 37(3):371–395, 2011.
6. Milan Šimůnek and Teppo Tammisto. Distributed data-mining in the LISp-Miner system using Techila grid. In *Networked Digital Technologies'10*, pages 15–21, Berlin, 2010. Springer.