Interactive Medical Miner: Interactively exploring subpopulations in epidemiological datasets

Uli Niemann[1], Myra Spiliopoulou[1], Henry Völzke[2], and Jens-Peter Kühn[2]

[1]Knowledge Management and Discovery Lab, University Magdeburg, Germany
uliniemann@hotmail.com, myra@iti.cs.uni-magdeburg.de
[2]Institute for Community Medicine, University Greifswald, Germany
{voelzke,kuehn}@uni-greifswald.de

Abstract. We present our Interactive Medical Miner, a tool for classification and model drill-down, designed to study epidemiological data. Our tool encompasses supervised learning (with decision trees and classification rules), utilities for data selection, and a rich panel with options for inspecting individual classification rules, and for studying the distribution of variables in each of the target classes. Since some of the epidemiological data available to the medical researcher may be still unlabeled (e.g. because the medical recordings for some part of the cohort are still in progress), our Interactive Medical Miner also supports the juxtaposition of labeled and unlabeled data. The set of methods and scientific workflow supported with our tool have been published in [1].

1 Introduction

High quality decisions of personalized medicine involve identifying subgroups that share some risk factors or symptoms associated with a certain disease. In [1], we have presented mining methods for the discovery of risk factors in subgroups of an epidemiological study's cohort. In [1], we have shown a mining approach for splitting a cohort in subgroups and for discovering factors associated with the outcome for each subgroup. We have thereby demonstrated that different subgroups exhibit different factors. The top-classification rules found by our approach agree with research results in epidemiology publications.

We present here our Interactive Medical Miner[1], as well as utilities for learning and model inspection. The Interactive Medical Miner derives models from epidemiological data containing a nominal target variable, for instance a diagnosis report outcome, and allows the user to drill down to the data of distinct individuals, and to further explore detailed information about summary statistics or class distribution histograms. Contrary to medical research practice which is hypotheses-based, we use a data-driven approach, as practiced e.g. in [2,3]. Our Interactive Medical Miner offers, under a simple interface, several functionalities for medical researches who aim to interactively explore their datasets and inspect classification patterns derived on them. To this purpose, the Interactive Medical

[1] The tool can be downloaded at http://kmd.cs.ovgu.de/res/imm/.

Miner provides a tailored workflow for preparation and classification of epidemiological data, including model drill-down and summary statistics for each class and rule. The tool can also be used in a more general medical (clinical) context.

We use algorithms from the Weka[2] library: We leverage the HotSpot[3] algorithm for classification rule discovery and employ the J48 (equivalent to the C4.5 algorithm [4]) for decision tree induction.

2 Workflow of the Interactive Medical Miner

Our tool takes as input the data of an epidemiological cohort, where the target variable concerns a medical outcome, e.g. the presence of increased fat in the liver [1]. The tool allows the medical researcher to specify that either the complete dataset or a selection of cohort participants (a subpopulation) should be used for learning. On this dataset, the Interactive Medical Miner discovers classification rules and builds decision trees. The decision trees are presented to the expert, while classification rules can be further *explored*. In particular, the expert is shown histograms on the distribution of the participants supporting a rule with respect to each class, and histograms on the distribution of the rule's variables inside each class. Since some of the cohort participants in the dataset we study (SHIP [5]) are not yet labeled, our tool supports the inspection of the values in each classification rule's antecedent on the unlabeled data as well.

3 User Interface

The user interface of the Interactive Medical Miner consists of two areas; each one is comprised of six panels. Subsequently, we describe the layout of the Interactive Medical Miner while referring to Figure 1.

In the upper left "Settings" panel, the user controls the most important algorithmic parameters. For classification rules generation, the user has to specify:

- Minimum value count: the minimum percentage (or number) of instances supporting the rule AND belonging to specified target class,
- Maximum rule length: the number of variables in the antecedent,
- Maximum branching factor: the maximum number of variables that may be added to an existing classification rule,
- Minimum improvement: the minimum relative confidence improvement to be achieved by the addition of a further variable to the classification rule.

For decision tree induction, the following three parameters are required:

- Minimum number of data records in a leaf node,
- Pruning factor: the threshold that must be satisfied if the tree is pruned; if the value falls below this threshold, the tree is not pruned further,

[2] `http://www.cs.waikato.ac.nz/ml/weka/`
[3] `http://weka.sourceforge.net/packageMetaData/hotSpot/Latest.html`

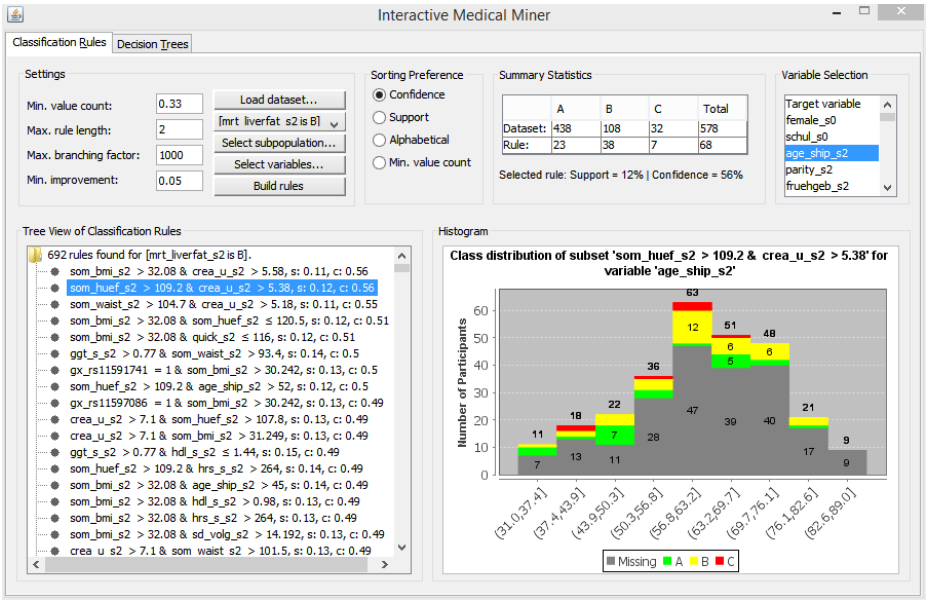

Fig. 1. Screenshot of the Interactive Medical Miner with its six panels.

– Only binary splits: a flag indicating, whether nominal variables with a value range of n values should be subjected to binary splits or to n-ary splits.

Further, the tool allows the user to specify a subpopulation of the dataset. For instance, one might filter out the male participants and study only female participants. To this purpose, a new frame pops up where filtering queries in the form of `Variable Operator Value` can be specified. The defined restrictions are shown in a table where the user can optionally select and remove single entries. Next, the Interactive Medical Miner also offers a button "Filter Variables..." which opens a pop-up frame where the user can (de-)select one or more variables for model generation by shifting them from one list to the other. For example, the user might exclude a variable that is known to be highly correlated with another variable that is already considered for model learning.

By clicking on "Build Rules"/ "Generate Tree", the resulting model is depicted in the "Tree View" panel. For classification rules, the retrieved rules can be sorted via the radio buttons in the panel "Sorting Preference" (area right to "Settings") according to several criteria, including confidence (default), support, alphabetical and minimum value count. For decision trees, the output tree structure can be visualized with Weka's TreeVisualizer.

When a rule/ node is selected, the top middle area "Summary Statistics" is refreshed. The first row shows the class distribution of the dataset, while the second row shows how the instances supporting the antecedent are distributed among the existing classes. Hence, the user gets insights about the class distribution of a single rule/ node and thus can control the mining process by adapting

parameters or selecting specific variables or value ranges. For instance, the tool allows the user to trade of high confidence against high support rules.

The summary statistics table contains information about the labeled instances, while the histogram in the bottom right panel covers the unlabeled instances. The user can choose a further variable from the panel "Variable Selection" (cf. Figure 1, upper right corner) to see how the values of these variables are distributed. Unlabeled data are marked as "Missing". To plot the histograms, we employ the open source chart library JFreeChart[4].

4 Conclusion

The Interactive Medical Miner generates classification models on epidemiological datasets and supports interactive exploration of individual classification rules and decision tree nodes. The tool provides options for filtering cohort participants and selecting a subset of variables. It allows the user to tune algorithm parameters and thus guide the mining process. The visual representation of class distributions and the juxtaposition of labeled and unlabeled cohort participants improves data understanding and might reveal idiosyncrasies of the labeled data. In future work, we intend to extend the tool by adding more classifiers, a more elaborate visualization as well as model and graphic export possibilities.

Acknowledgements

Part of this work was supported by the German Research Foundation project SP 572/11-1 "IMPRINT: Incremental Mining for Perennial Objects".
The data used in this work were made available through the cooperation SHIP/ 2012/06/D "Predictors of Steatosis Hepatis".

References

1. U. Niemann, H. Völzke, J.-P. Kühn, and M. Spiliopoulou, "Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5405–5415, 2014.
2. C. Zhanga and R. L. Kodell, "Subpopulation-specific confidence designation for more informative biomedical classification," *Artificial Intelligence in Medicine*, vol. 58, no. 3, pp. 155–163, 2013.
3. F. Pinheiro, M.-H. Kuo, A. Thomo, and J. Barnett, "Extracting association rules from liver cancer data using the FP-growth algorithm," in *3rd International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, 2013.
4. J. R. Quinlan, "Learning with continuous classes," in *5th Australian Joint Conference on Artificial Intelligence*, vol. 92, pp. 343–348, 1992.
5. H. Völzke, D. Alte, C. O. Schmidt, D. Radke, R. Lorbeer, N. Friedrich, *et al.*, "Cohort Profile: The Study of Health in Pomerania," *International Journal of Epidemiology*, vol. 40, no. 2, pp. 294–307, 2011.

[4] `http://www.jfree.org/jfreechart/`