In: Proceedings of the LWA 2011 - Learning, Knowledge, Adaptation.
University of Magdeburg, Germany. 2011

VIKAMINE – A Rich-Client Environment for Intelligent Pattern Mining and Subgroup Discovery

(Demonstration Paper)

Martin Atzmueller
Knowledge and Data Engineering Group
University of Kassel

Florian Lemmerich
Artificial Intelligence Group
University of Wuerzburg

atzmueller@cs.uni-kassel.de, lemmerich@informatik.uni-wuerzburg.de

Abstract

This paper presents an overview on the VIKAMINE[1] system, an open-source Eclipse-based[2] mining environment. VIKAMINE focuses on efficient and effective techniques for pattern mining and subgroup discovery; as of VIKAMINE version 2 it is implemented as rich-client platform (RCP) application, providing a solid framework for high extensibility. We present the system and sketch applications in medical and technical domains.

1 Introduction

Pattern mining and subgroup discovery are important tools for descriptive data mining, in order to obtain an overview on the relations in the data, for automatic hypotheses generation, and for a number of knowledge discovery applications. We present the VIKAMINE system for such applications. As of version 2, VIKAMINE is an open environment for intelligent pattern mining and subgroup discovery, which features a variety of state-of-the-art automatic algorithms, visualizations, broad extensibility, and powerful customization capabilities enabled by the Eclipse RCP environment.

VIKAMINE is targeted at a broad range of users, from industrial practitioners to ML/KDD researchers, students, and users interested in knowledge discovery and data analysis in general. Especially the visual mining methods enable the direct integration of the user to overcome major problems of automatic data mining methods, e.g., the presentation of uninteresting results, lack of acceptance of the discovered findings, or limited confidence in these. Thus, the automatic methods complement the interactive tools by providing efficient and effective mining solutions.

We briefly outline the VIKAMINE system and sketch two exemplary applications, considering medical knowledge discovery, and for technical fault analysis and optimization in an industrial setting.

2 VIKAMINE

VIKAMINE 2.x is implemented as an Eclipse-based rich-client application. In contrast to general purpose data mining systems, it is specialized for the task of subgroup discovery and pattern mining. It focuses on visual, interactive and knowledge-intensive methods and aims to integrate a comprehensive set of features with an easy-to-use interface. The main features of VIKAMINE include:

- **State-of-the-Art Algorithms:** VIKAMINE comes with a variety of established and state-of-the-art algorithms for automatic subgroup discovery, e.g., Beam-Search [Lavrac *et al.*, 2004], BSD [Lemmerich *et al.*, 2010], and SD-Map* [Atzmueller and Lemmerich, 2009]. As target concepts binary, nominal, and numeric targets can be used, with a number of popular interestingness measures. The results can be optimized by applying discretization or filtering techniques utilizing relevancy or significance criteria.
- **Visualizations:** For successful interactive mining, powerful visualizations are essential to achieve a quick understanding of the data and mined patterns. Visualizations implemented in VIKAMINE include, for example, the *zoomtable* [Atzmueller and Puppe, 2005] for visual and semi-automatic subgroup discovery shown in Figure 1, pattern specialization graphs, or visualizations of patterns in the ROC-space.

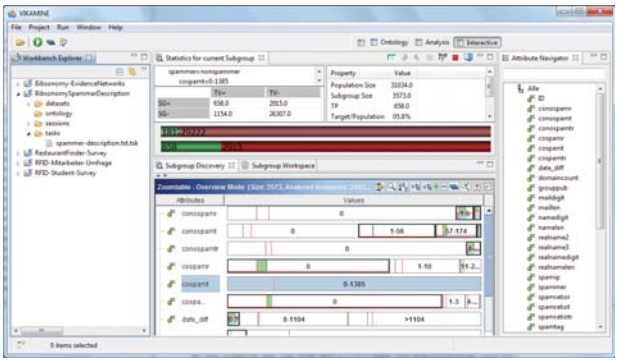

Figure 1: Screenshot of the VIKAMINE workbench: Projects (left), the zoomtable (middle, bottom), pattern statistics (middle, top) and the attribute view (right).

- **Import/Export:** Datasets can be imported in different file-formats, e.g., CSV or ARFF files. Modified/filtered datasets can then be exported in the same file formats. Connecting to and from standard SQL databases is enabled by plugins, e.g., utilizing the Eclipse Data Tools Platform. The results can then be exported as plain text, to XML, or can be directly reused in standard Office Tools, e.g., MS-Excel.
- **Prior Knowledge:** VIKAMINE supports various methods to integrate prior knowledge into the mining process, e.g., considering expected/known dependencies, causal analysis, and pattern filtering options. Background knowledge can be acquired using form-based approaches or integrated using text documents, that can be edited in the integrated Eclipse text editor.

[1] http://www.vikamine.org
[2] http://www.eclipse.org

- **Extensibility:** Using the Rich Client Platform of Eclipse, VIKAMINE can easily be extended by specialized plug-ins for the target application area. Customized extension points allow, for example, for a quick integration of new interesting measures, search algorithms, new visualizations, new types of background knowledge and specialized views on the data into the graphical user interface. Specialized plugins using such extension points also allow for the integration of other data mining and statistic libraries, e.g., for a connection to the statistic environment R[3].
- **Modularity:** VIKAMINE utilizes a strict separation between kernel components, i.e., data representations and algorithms, and the graphical interface components. Thus, the algorithmic core functionalities of VIKAMINE can be easily integrated in other systems and applications, e.g., for the integration in production environments, or for evaluations of algorithms by researchers.
- **Organization:** Automatic discovery tasks can be stored as a declarative XML. By utilizing Eclipse workspace and project concepts, VIKAMINE supports the user in keeping track of all the data, performed tasks and results of a data mining project.

3 Exemplary Applications

Below, we briefly summarize two successful real-world applications of VIKAMINE. First, we discuss a knowledge discovery and quality control setting in the medical domain. After that, we sketch an industrial application, in which VIKAMINE was applied for identifying patterns indicating faults in production processes.

3.1 Knowledge Discovery and Quality Control in the Medical Application Domain

VIKAMINE has been applied for large-scale knowledge discovery and quality control in a clinical application, cf., [Puppe *et al.*, 2008]. For this, several data sources including structured clinical documentation and unstructured documents, e.g., [Atzmueller *et al.*, 2011], were integrated. The main data source was given by the SONOCONSULT system, which has been in routine use since 2002 as the only documentation system for ultrasound examinations in the DRK-hospital of Berlin-Köpenick; since 2005, it is in routine use at the university hospital of Würzburg. The physicians considered statistical analysis as one of the desirable features.

According to the physicians, subgroup discovery is quite suitable for examining common medical questions, e.g. whether a certain pathological state is significantly more frequent if combinations of other pathological states exist or if there are diagnoses, which one physician documents significantly more or less frequently than the average. Furthermore, VIKAMINE also provides an intuitive overview on the data, in addition to the knowledge discovery and quality control functions.

3.2 Technical Fault Analysis in the Industrial Application Domain

The second application example concerns large-scale applications in the industrial domain concerning technical fault analysis and optimization in production and service support scenarios.

[3]http://www.r-project.org/

In the application, one important goal was the identification of subgroups (as combination of certain factors) that cause a significant increase/decrease in certain parameters. This concerns, for example, the number of service requests for a certain technical component, the fault/repair rate of a certain manufactured product, or the number of calls of customers to service support.

Such industrial applications of subgroup discovery often require the utilization of continuous parameters, for example, certain measurements of machines or production conditions. Then, the target concepts can often not be analyzed sufficiently using the standard discretization techniques, since the discretization of the variables causes a loss of information. As a consequence, the interpretation of the results is often difficult using standard data mining tools. In this context, VIKAMINE provides state-of-the-art algorithmic implementations [Atzmueller and Lemmerich, 2009; Lemmerich *et al.*, 2010] for supporting the knowledge discovery and analysis, and enables a semi-automatic involvement of the domain experts for effectively contributing in a discovery session.

4 Conclusions

In this paper, we presented an overview on the open-source data mining environment VIKAMINE: As of version 2, it is implemented as an Eclipse-based rich-client platform (RCP) application. This provides for a solid framework that is highly modular and broadly extensible. The system focuses on efficient and effective pattern mining and subgroup discovery. We briefly summarized the unique features of VIKAMINE and sketched two exemplary applications in medical and technical domains.

References

[Atzmueller and Lemmerich, 2009] Martin Atzmueller and Florian Lemmerich. Fast Subgroup Discovery for Continuous Target Concepts. In *Proc. 18th Intl. Symp. Method. Intelligent Systems*. Springer Verlag, 2009.

[Atzmueller and Puppe, 2005] Martin Atzmueller and Frank Puppe. Semi-Automatic Visual Subgroup Mining using VIKAMINE. *Journal of Universal Computer Science (JUCS), Special Issue on Visual Data Mining*, 11(11):1752–1765, 2005.

[Atzmueller *et al.*, 2011] Martin Atzmueller, Stephanie Beer, and Frank Puppe. *Collaboration and the Semantic Web: Social Networks, Knowledge Networks and Knowledge Resources*, chapter Data Mining, Validation and Collaborative Knowledge Capture. IGI Global; forthcoming, 2011.

[Lavrac *et al.*, 2004] Nada Lavrac, Branko Kavsek, Peter Flach, and Ljupco Todorovski. Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[Lemmerich *et al.*, 2010] Florian Lemmerich, Mathias Rohlfs, and Martin Atzmueller. Fast Discovery of Relevant Subgroup Patterns. In *Proc. 21st Intl. FLAIRS Conference*, pages 428–433. AAAI Press, 2010.

[Puppe *et al.*, 2008] Frank Puppe, Martin Atzmueller, Georg Buscher, Matthias Huettig, Hardi Lührs, and Hans-Peter Buscher. Application and Evaluation of a Medical Knowledge-System in Sonography (SonoConsult). In *Proc. 18th European Conference on Artificial Intelligence (ECAI 2008)*, pages 683–687, 2008.