VIKAMINE – Open-Source Subgroup Discovery, Pattern Mining, and Analytics

Martin Atzmueller[1] and Florian Lemmerich[2]

[1] University of Kassel,
Knowledge and Data Engineering Group
Wilhelmshöher Allee 73, 34121 Kassel, Germany
[2] University of Wuerzburg,
Artificial Intelligence and Applied Informatics Group,
Am Hubland, 97074 Wuerzburg, Germany
atzmueller@cs.uni-kassel.de,
lemmerich@informatik.uni-wuerzburg.de

Abstract. This paper presents an overview on the VIKAMINE[1] system for subgroup discovery, pattern mining and analytics. As of VIKAMINE version 2, it is implemented as rich-client platform (RCP) application, based on the Eclipse[2] framework. This provides for a highly-configurable environment, and allows modular extensions using plugins. We present the system, briefly discuss exemplary plugins, and provide a sketch of successful applications.

Keywords: Pattern Mining, Subgroup Discovery, Analytics, Open-Source.

1 VIKAMINE

Subgroup discovery and pattern mining are important descriptive data mining tasks. They can be applied, for example, 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.

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. It features a variety of state-of-the-art automatic algorithms, visualizations, broad extensibility, and rich customization capabilities enabled by the Eclipse RCP environment. 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 distinctive set of features with an easy-to-use interface:

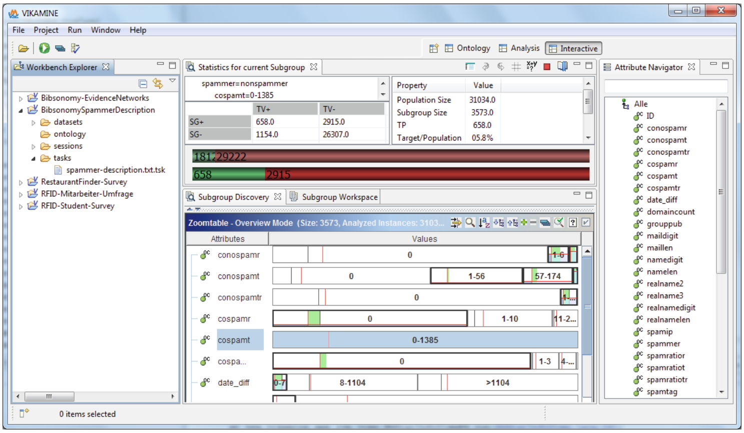
- **State-of-the-Art Algorithms:** VIKAMINE comes with a variety of established and state-of-the-art algorithms for automatic subgroup discovery, e.g., Beam-Search [7], BSD [9], and SD-Map* [2]. A wide variety of popular interestingness measures can be used for binary, nominal, and numeric target concepts.

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

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

P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 842–845, 2012.

- **Visualizations:** For successful interactive mining, specialized visualizations are essential to achieve a quick and intuitive understanding of the data and mined patterns [6]. Visualizations implemented in VIKAMINE include, for example, the *zoomtable* [5] for visual and semi-automatic subgroup discovery shown in Figure 1, pattern specialization graphs, or visualizations of patterns in the ROC-space.



spammer=nonspammer
cospamt=0-1385

	TV+	TV-
SG+	658.0	2915.0
SG-	1154.0	26307.0

Property	Value
Population Size	31034.0
Subgroup Size	3573.0
TP	658.0
Target/Population	05.8%

Zoomtable - Overview Mode (Size: 3573, Analyzed Instances: 3103...

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

- **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 text documents.
- **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 interestingness measures, search algorithms, visualizations, types of background knowledge and domain specific views on the data. 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 declaratively be stored in 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] `http://www.r-project.org/`

2 Selected Plugins

- **VIKAMINE R-Plugin**: In order to integrate external data mining and analysis methods, the VIKAMINE R-Plugin features the ability to connect to the R[4] environment for statistical computing. Using the plugin, for example, external methods and visualizations can easily be integrated using R scripts.
- **VIKAMINE Geo-Plugin**: For mining spatial data, the VIKAMINE Geo-Plugin provides for specialized mining and visualization options taking geo-locations into account. For example, patterns characterizing specific locations can be mined, including tagging data for the description [8]. The plugin provides suitable visualization options for further inspection, browsing, and refinement.
- **VIKAMINE Community Mining Plugin**: Descriptive community mining solves one of the major problems of standard approaches, i.e., that the discovered communities have no inherent *meaning*. The community mining plugin implements such an approach; using a graph structure and descriptive information, e.g., friendship networks and tags applied by the users [4], descriptive patterns with high commnity qualities according to standard measures are obtained.

3 Exemplary Applications

- **Medical Knowledge Discovery and Quality Control**: VIKAMINE has been applied for large-scale knowledge discovery and quality control in the clinical application SONOCONSULT , cf., [10]. According to the physicians, subgroup discovery and analysis 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 interface for providing an overview on the data, in addition to the knowledge discovery and quality control functions.
- **Industrial Fault Analysis**: Another application concerned large-scale technical fault analysis. The task required the identification of subgroups (as combination of certain factors) that cause a significant increase/decrease in the fault/repair rates of certain products. In the application, one important goal was the identification of subgroups (as combination of certain factors) that cause a significant increase/ 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 applications often require the utilization of continuous parameters. 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. In this context, VIKAMINE provides state-of-the-art algorithmic implementations, cf. [2], for supporting the knowledge discovery and analysis, and enables an effective involvement of the domain experts.

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

- **Mining Social Media**: In addition to the approaches sketched above VIKAMINE features a number of successful applications in the social media domain, see [8,4]. It was applied, for example, for obtaining descriptive profiles of spammers, i.e., for their characterization [3]. The mined patterns capturing certain spammer subgroups provide explanations and justifications for marking or resolving spammer candidates in a social bookmarking systems. In such contexts, it is also useful to identify high-quality tags, i.e., tags with a certain maturity, cf. [1]. VIKAMINE was applied for obtaining maturity profiles of tags based on graph centrality features on the tag—tag cooccurrance graph. Then, the obtained information can be utilized for tag recommendations, faceted browsing, or for improving search.

4 Conclusions

In this paper, we presented an overview on VIKAMINE, focusing on efficient and effective pattern mining and subgroup discovery. As of version 2, VIKAMINE is implemented as an Eclipse-based rich-client platform (RCP) application. This provides for an integrated system that is highly modular and broadly extensible using plugins. VIKAMINE can be freely downloaded from `http://www.vikamine.org` under an *LGPL* open-source license.

References

1. Atzmueller, M., Benz, D., Hotho, A., Stumme, G.: Towards Mining Semantic Maturity in Social Bookmarking Systems. In: Proc. Workshop Social Data on the Web, International Semantic Web Conference (2011)
2. Atzmueller, M., Lemmerich, F.: Fast Subgroup Discovery for Continuous Target Concepts. In: Proc. 18th Intl. Symp. Method. Intelligent Systems. Springer (2009)
3. Atzmueller, M., Lemmerich, F., Krause, B., Hotho, A.: Who are the Spammers? Understandable Local Patterns for Concept Description. In: Proc. 7th Conference on Computer Methods and Systems (2009)
4. Atzmueller, M., Mitzlaff, F.: Efficient descriptive community mining. In: Proc. 24th Intl. FLAIRS Conference. AAAI Press (2011)
5. Atzmueller, M., Puppe, F.: Semi-Automatic Visual Subgroup Mining using VIKAMINE. Journal of Universal Computer Science, Visual Data Mining 11(11), 1752–1765 (2005)
6. Gamberger, D., Lavrac, N., Wettschereck, D.: Subgroup Visualization: A Method and Application in Population Screening. In: Proc. 7th Intl. Workshop on Intelligent Data Analysis in Medicine and Pharmacology, IDAMAP 2002 (2002)
7. Lavrac, N., Kavsek, B., Flach, P., Todorovski, L.: Subgroup Discovery with CN2-SD. Journal of Machine Learning Research 5, 153–188 (2004)
8. Lemmerich, F., Atzmueller, M.: Modeling Location-based Profiles of Social Image Media using Explorative Pattern Mining. In: Proc. IEEE SocialCom 2011, Workshop on Modeling Social Media (MSM 2011). IEEE Computer Society, Boston (2011)
9. Lemmerich, F., Rohlfs, M., Atzmueller, M.: Fast Discovery of Relevant Subgroup Patterns. In: Proc. 21st Intl. FLAIRS Conference, pp. 428–433. AAAI Press (2010)
10. Puppe, F., Atzmueller, M., Buscher, G., Huettig, M., Lührs, H., Buscher, H.-P.: Application and Evaluation of a Medical Knowledge-System in Sonography (SonoConsult). In: Proc. 18th European Conference on Artificial Intelligence, pp. 683–687 (2008)