VIKAMINE – An Overview

Martin Atzmueller
University of Würzburg,
Department of Computer Science VI,
Am Hubland, 97074 Würzburg, Germany
atzmueller@informatik.uni-wuerzburg.de

1 Introduction

VIKAMINE (Visual, Interactive and Knowledge-intensive Analysis and MINing Environment) is a rich client application implemented in Java. A JRE[1] 6 or better is required for the application.

The VIKAMINE system is an integrated environment and features editors and components for the automatic and interactive discovery of subgroups, for subgroup introspection and analysis, for including and editing background knowledge, and finally for visual inspection, analysis and comparison of subgroups.

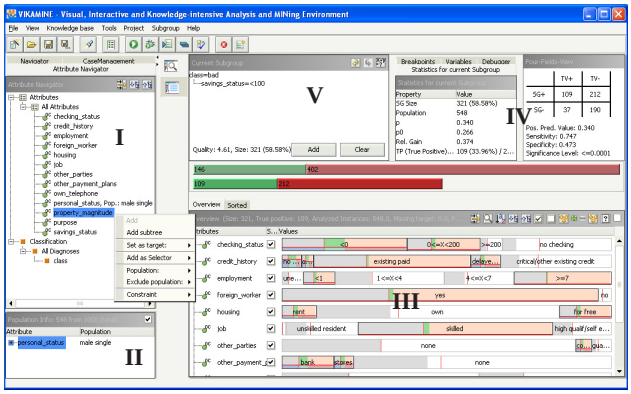

Figure 1: VIKAMINE: Main user interface.

2 Overview

The VIKAMINE tool can be applied for purely automatic, semi-automatic, and for interactive subgroup mining. Concerning interactive subgroup discovery and analysis, an intuitive tool is provided that enables simple navigation and easy interaction with respect to the data and the space of subgroup hypotheses.

For more automatic methods, the navigation options become less important while the selection and definition capabilities are essential: Since purely automatic methods cannot be guided directly by the user, appropriate interestingness and quality criteria need to be specified. In addition, providing high-quality background knowledge for constraining the search space and for controlling the search process is often an important prerequisite. VIKAMINE implements techniques for both highly interactive and automatic mining that can also be suitably combined.

Figure 1 shows a screenshot of the main user-interface of VIKAMINE. In general, the user interface is split into two main panes: The navigation and selection pane is shown on the left and includes the *attribute navigator* (Annotation I) and the population panel (Annotation II) defining the used instance population on the left. On the right, the interaction pane is shown that includes the zoomtable (Annotation III), the current subgroup view (Annotation V) and the subgroup statistics panes (Annotation IV). The navigation and selection pane contains the available attributes. These can be selected and inserted into the zoomtable, which then shows the corresponding value distributions of the selected attributes.

[1] Java Runtime Environment

3 Import/Load A Data Set

As a first step when using VIKAMINE, you need to import or load a dataset. VIKAMINE can import several standard data formats, e.g., CSV, ARFF, XSL format. This can be done using the 'Import' options in the file menu.

4 VIKAMINE for Data Analysis

Besides subgroup discovery, the zoomtable also provides for basic data analysis capabilities. Since the value distributions of the selected attributes are visualized in the zoomtable using portions (cells) of a bar contained in the rows of the table, the user can get an easy and intuitive overview on the data. Then, including additional information in the cells of the table can often help the user in performing the analysis.

Figure 2 shows a screenshot of a simple configuration: In this figure, the frequencies of the individual values are depicted by the widths of the sub-bars contained in the zoomtable. Then, a first overview of the value distributions can be obtained. The zoomtable is highly configurable and can include additional information, e.g., it can display the individual frequencies below the individual value strings.

Additionally, the zoomtable can be used for simple correlation analysis similar to basic *OLAP* (Online Analytical Processing) techniques. The *sorted* mode of the zoomtable enables the analysis of the different rows with respect to other rows, i.e., given a sorting attribute the values of the other attributes are grouped by the values of the respective attribute. An example is shown in Figure 3.

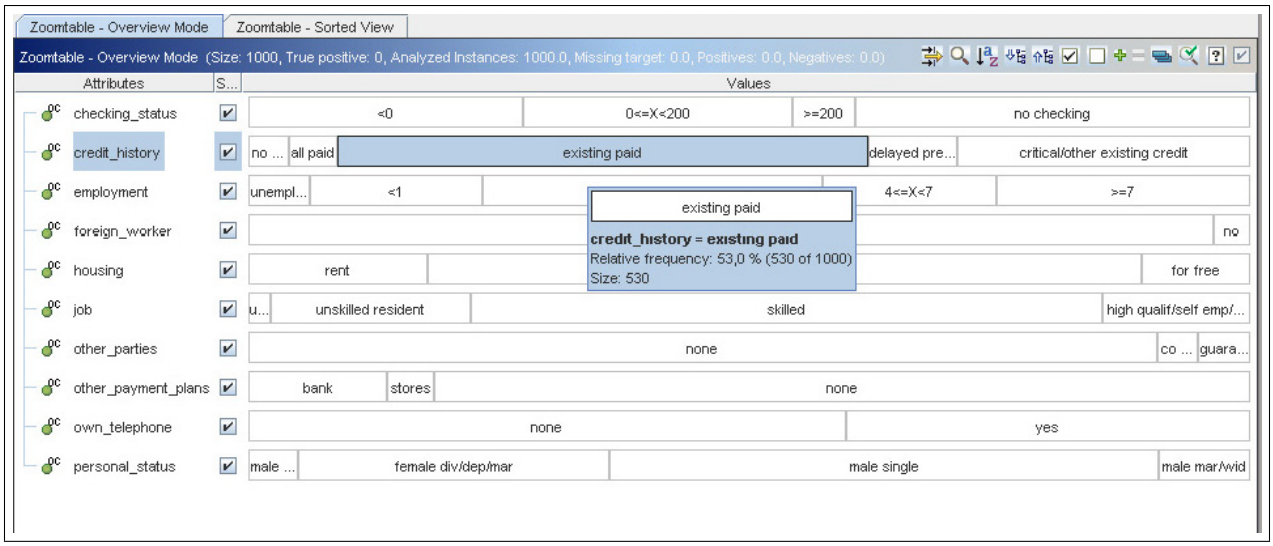

Figure 2: Zoomtable: Overview

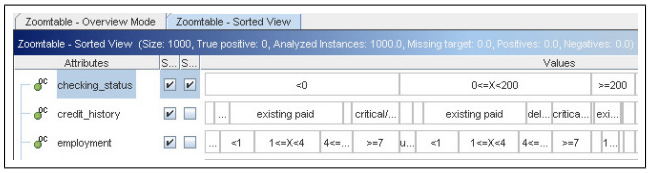

Figure 3: Zoomtable: Sorted view

5 Performing Subgroup Discovery

The interaction pane for subgroup discovery includes the zoomtable view (Annotation III in Figure 1), in addition to various views that display statistical information (Annotation IV) above the main zoomtable, and the current subgroup view (Annotation V).

5.1 Interactive Subgroup Discovery

After the attributes have been added to the zoomtable they are available for subgroup discovery and the current subgroup view can be (interactively) modified.

The zoomtable visualization shows the value distributions of selected analysis attributes/variables in the rows of the table corresponding to the attributes in the first column. Usually the distribution is scaled such that the widths of the bars depicting the attribute values correspond to the respective frequencies of the attribute values. Optionally, the bars can also be evenly scaled in order to show infrequent attribute values.

In addition to this basic statistical information, the zoomtable can contain a number of visual markers for guiding the discovery process that are configurable on the fly. For example, interesting values can be highlighted, or the general trend of the quality measure can be indicated. An example is shown in Figure 1.

The zoomtable always shows the distribution of the data restricte to the currently selected subgroup: Each row of the zoomtable shows the value distribution of a specific attribute limited to the cases covered by the current subgroup; the width of each cell relates to the frequency of the respective attribute value.

For a detailed view, Figure 4 shows the abstract structure of a row of the zoomtable including the type of the attribute, its current ranking, the attribute name, and its value distribution annotated with several visual markers. In general, two of the most important parameters of a subgroup are the *target share* (p) and the *size* (n) of the respective subgroup. There is always a trade-off between these parameters that is usually formalized by the applied quality function. So, for the interactive part of the semi-automatic process for subgroup mining, we want to visualize possible future changes or improvements regarding these parameters. The *subgroup size* with respect to a future subgroup is given by the width of a specific selector cell. The current target share is visualized in the individual cells by visual markers: (a) indicates the

Figure 4: The zoomtable – detail view

positive and (b) the negative instances of the current subgroup SG_c; (c) shows the positive instances for the subgroup SG_n, i.e., the subgroup that is constructed by including the particular attribute value. If (c) is larger than (a), then the target share increases adding this selector. Furthermore, (d) shows the relative gain in the target share p, comparing the subgroups SG_c and SG_n, i.e., $d \sim \frac{c-a}{b}$, $c \geq a$; the marker (d) can then be used for an easier assessment of small cells. If the height of (d) is zero, then the target share does not increase. If it fills the entire bar, then the target share reaches 100%.

By interpreting these visual markers of each cell which are shown using different colors the user can immediately identify promising improvements of the currently active subgroup. If the target share increases, then the horizontal marker (c) is indicated in green, otherwise a gray bar is shown. For an improving selector the remaining area $b - c$ is shown in red color.

Usually the user is supported by the visualization of the distribution of the variables and by the visual markers contained in the zoomtable. Then, the spe-

cific selectors of the current subgroup can be selected in the zoomtable directly. Alternatively, the current subgroup can be modified using the elements of the navigation tree. However, then there is no visual feedback concerning the variables that are not contained in the zoomtable. This general workflow is summarized in Figure 5.

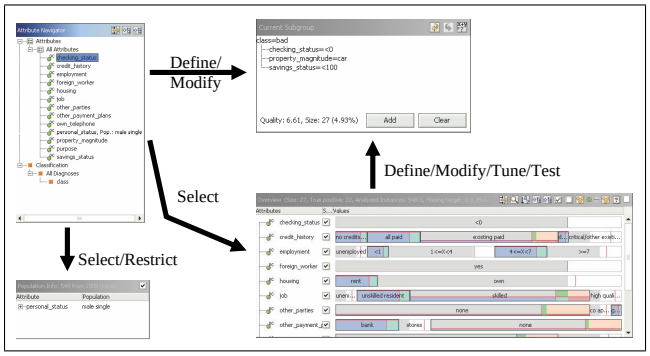

Figure 5: VIKAMINE: Interactions for performing interactive subgroup discovery.

5.2 Automatic Subgroup Discovery

After the the relevant attributes and values have been selected, the automatic subgroup discovery step can be started as shown in Figure 7.

The discovery process can be launched by the user either selecting a purely automatic discovery option, or a mode that is similar to a debugger for conventional programming languages. Then, the discovery process can be interrupted at any time by the user in order to inspect and/or change the current state of the search task. Thereafter, the search process can be continued. This provides for a 'supervised' automatic subgroup mining step that is more transparent for the user since the intermediate results of the search process can be inspected.

Additionally, a comprehensive subgroup search can be performed such that each possible (binary) target variable is considered that can be constructed given the settings of the zoomtable. Such a subgroup discovery task can be used for initial exploration of the hypothesis space. If no target variable was selected, a dialog window (shown in Figure 6) appears. The user can then define and extend the search space choosing the target variable(s) of interest.

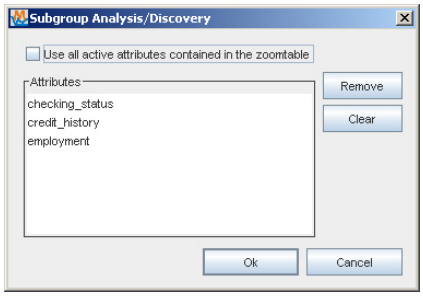

Figure 6: VIKAMINE: Target Selection

Automatic subgroup discovery is configured in the subgroup settings dialog, in the zoomtable, and in

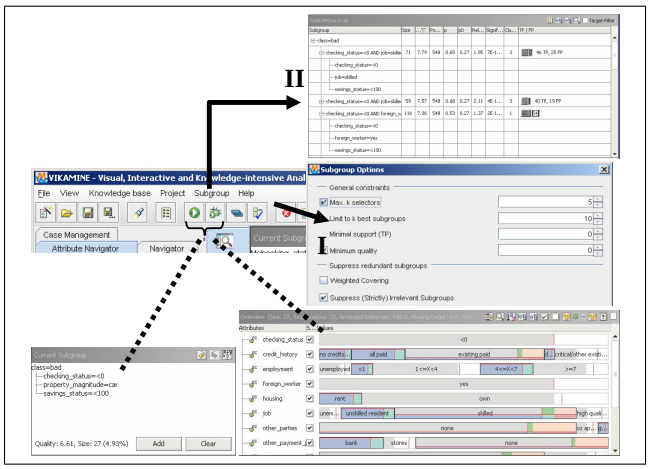

Figure 7: VIKAMINE: Automatic subgroup discovery.

the current subgroup view. There are several options that can be configured in the *SettingsDialog* (Annotation I) shown in Figure 7. These are mainly language constraints, e.g., specifying the maximum length of the subgroup description, and the minimum subgroup quality. Additionally, a quality function needs to be selected that can also be defined manually using a formula-editor to include the quality parameters.

Additionally, both the current subgroup view and the zoomtable state affect the automatic discovery method. The currently active subgroup, i.e., the one that is contained in the current subgroup view is used as the initial hypothesis for subgroup search. Furthermore, only the variables that are enabled in the zoomtable are considered for automatic subgroup discovery, i.e., the attributes for which the checkbox to the right of the attribute column is selected. Further constraint knowledge can be provided to the system, e.g., using the context menus of the attributes contained in the zoomtable. Then, the discovery process can be started.

The result applying the automatic subgroup discovery is a set of subgroups that is shown in a tree-table view (Annotation II) in Figure 7. A detailed view is given in Figure 8. The interesting subgroups can be analyzed, inspected, and compared using visualizations and that are enabled using the context menu of the subgroups results.

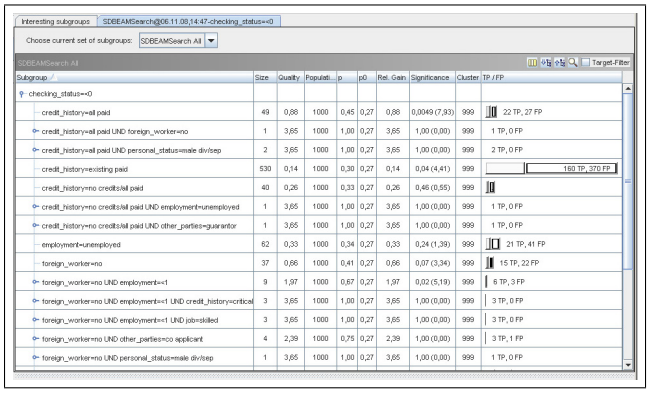

Figure 8: VIKAMINE: Results of subgroup discovery.

6 Interactive Subgroup Refinement and Tuning

A visualization that is orthogonal to the mechanisms provided in the interaction pane, i.e., the zoomtable and the current subgroup view, is provided by the subgroup tuning table.

The subgroup tuning table can be used for purely interactive subgroup discovery by the user, and for detailed analysis of automatically discovered subgroups. Given a set of user-determined attributes, the user can select each attribute value as a selector for specialization by a single click in a value cell. Furthermore, a specific subgroup can be analyzed in the subgroup tuning table: Then, all its attributes and valid attribute values are included in the table. In this manner, the subgroup description can be fine-tuned by the user, e.g., by extending a selector into a disjunctive selections expression. Thus, subgroup specialization and generalization operations can be performed very intuitively.

Similar to the zoomtable, attributes can be transfered from the navigation tree to the subgroup tuning table. Then, a purely interactive subgroup discovery session is possible, using a restricted search space. Only selectors corresponding to attributes included in the tuning table can be examined with respect to the target variable: a subgroup is shown in the rows of the table, and the selectors making up the subgroup description correspond to the marked columns. Figure 9 shows an exemplary screenshot. In this visualization single factors can be evaluated very easily. Furthermore disjunctive subgroup descriptions can also generated in a very simple way.

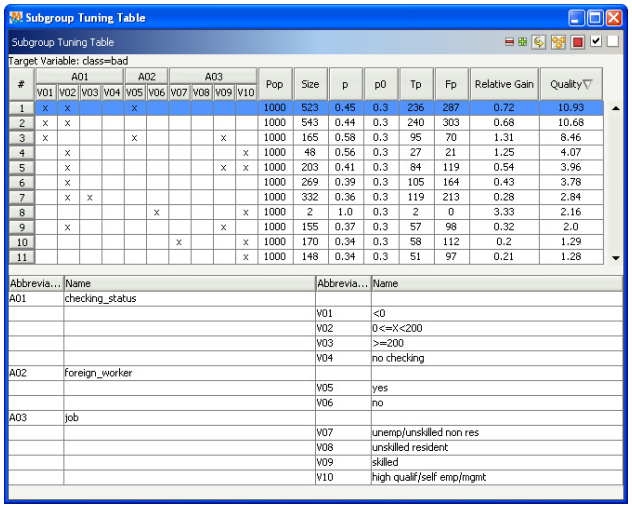

Figure 9: VIKAMINE: The subgroup tuning table.

The tuning table displays the usual subgroup parameters, the relative gain of the subgroups, and their quality, and can be sorted according to these criteria. In the example, the subgroups have been sorted according to the subgroup quality. The lower pane shows the abbreviated attribute and value descriptions that correspond to the columns of the tuning table. After interesting subgroup hypotheses have been identified, these can also be included in the subgroup workspace by a drag-and-drop operation.

7 Data Analysis and Subgroup Discovery using VIKAMINE in 7 Simple Steps

In the following we propose an approach for a general setup for data analysis and subgroup discovery using VIKAMINE. This process model summarizes the necessary steps for subgroup analysis and discovery.

1. First, load/import a dataset using the import filters in the 'File'-Menu. Currently, ARFF, CSV, Excel^{TM} and D3WEB XML files are supported (c.f., Section 3).
2. Include the relevant/interesting attributes in the zoomtable by selecting them from the navigation tree. The attributes can be added using the context-menu of each attribute, or they can be dropped directly into the zoomtable.
3. Define the population of interest in the population panel: This can be performed by selecting the respective attributes in the navigation tree, and configuring the population restriction using the contet menu.
4. Obtain an overview of the data using the overview mode of the zoomtable (c.f., Section 4).
5. Subgroup discovery (c.f., Section 5):
   (a) Automatic subgroup discovery: select a target variable via the context menu, or select a set of target variables or attributes using the discovery configuration dialog.
   (b) Interactive subgroup discovery: Select a specific target variable, and adapt the subgroup description, taking into account the information in the cells of the zoomtable.
   (c) Optional: Add the interesting subgroups to the global subgroup results panel. This action can be performed using the 'Add' button of the current subgroup. Furthermore, subgroups can be transfered from the local subgroup results tables to the global results table using the context menu of each subgroup.
6. For a set of selected interesting subgroups, perform subgroup comparison, introspection and analysis using the visualizations and options provided by the context menus in the subgroup results panel (c.f., Section 6).
7. Finally, select a discriminative set of subgroups: This can be obtained manually by selecting a subset of subgroups based on the tests performed in the previous step. Additionally, several post-processing techniques for sets of subgroups can be enabled in the subgroup options dialog.

8 Conclusion

In this tutorial we have outlined the basic features of the VIKAMINE system, and we have shown a simple process model how VIKAMINE can be applied for simple data analysis and especially for both automatic and interactive subgroup discovery. The system can be downloaded at `http://www.vikamine.de`, or see the Sourceforge project site: `http://www.sourceforge.net/projects/vikamine`.