Knowledge-Based Visualization to Support Spatial Data Mining

Gennady Andrienko and Natalia Andrienko

GMD - German National Research Center for Information Technology
Schloss Birlinghoven, Sankt-Augustin, D-53754 Germany
gennady.andrienko@gmd.de
http://allanon.gmd.de/and/

Abstract. Data mining methods are designed for revealing significant relationships and regularities in data collections. Regarding spatially referenced data, analysis by means of data mining can be aptly complemented by visual exploration of the data presented on maps as well as by cartographic visualization of results of data mining procedures. We propose an integrated environment for exploratory analysis of spatial data that equips an analyst with a variety of data mining tools and provides the service of automated mapping of source data and data mining results. The environment is built on the basis of two existing systems, Kepler for data mining and Descartes for automated knowledge-based visualization. It is important that the open architecture of Kepler allows to incorporate new data mining tools, and the knowledge-based architecture of Descartes allows to automatically select appropriate presentation methods according to characteristics of data mining results. The paper presents example scenarios of data analysis and describes the architecture of the integrated system.

1 Introduction

The notion of Knowledge Discovery in Databases (KDD) denotes the task of revealing significant relationships and regularities in data based on the use of algorithms collectively entitled "data mining". The KDD process is an iterative fulfillment of the following steps [6]:

1. Data selection and preprocessing, such as checking for errors, removing outliers, handling missing values, and transformation of formats.
2. Data transformations, for example, discretization of variables or production of derived variables.
3. Selection of a data mining method and adjustment of its parameters.
4. Data mining, i.e. application of the selected method.
5. Interpretation and evaluation of the results.

In this process the phase of data mining takes no more than 20 % of the total workload. However, this phase is much better supported methodologically

D.J. Hand, J.N. Kok, M.R. Berthold (Eds.): IDA'99, LNCS 1642, pp. 149–160, 1999.

and by software than all others [7]. This is not surprising because performing of these other steps is a matter of art rather than a routine allowing automation [8]. Lately some efforts in the KDD field have been directed towards intelligent support to the data mining process, in particular, assistance in the selection of an analysis method depending on data characteristics [2,4].

A particular case of KDD is knowledge extraction from spatially referenced data, i.e. data referring to geographic objects or locations or parts of a territory division. In analysis of such data it is very important to account for the spatial component (relative positions, adjacency, distances, directions etc.). However, information about spatial relationships is very difficult to represent in discrete, symbolic form required for the data mining methods. Known are works on spatial clustering [5] and use of spatial predicates [9], but a high complexity of data description and large computational expenses are characteristic for them.

2 Integrated Environment for Knowledge Discovery

For the case of analysis of spatially referenced data we propose to integrate traditional data mining instruments with automated cartographic visualization and tools for interactive manipulation of graphical displays. The essence of the idea is that an analyst can view both source data and results of data mining in the form of maps that convey spatial information to a human in a natural way. This offers at least a partial solution to the challenges caused by spatially referenced data: the analyst can easily see spatial relationships and patterns that are inaccessible for a computer, at least on the present stage of development. In addition, on the ground of such integration various KDD steps can be significantly supported.

The most evident use of cartographic visualization is in evaluation and interpretation of data mining results. However, maps can be helpful also in other activities. For example, visual analysis of spatial distributions of different data components can help in selection of representative variables for data mining and, possibly, suggest which derived variables would be useful to produce. On the stage of data preprocessing a map presentation can expose strange values that may be errors in the data or outliers. Discretization, i.e. transformation of a continuous numeric variable into one with a limited number of values by means of classification, can be aptly supported by a dynamic map display showing spatial distribution of the classes. With such a support the analyst can adjust the number of classes and class boundaries so that interpretable spatial patterns arise.

More specifically, we propose to build an integrated KDD environment on the basis of two existing systems, Kepler [11] for data mining and Descartes [1] for interactive visual analysis of spatially referenced data. Kepler includes a number of data mining methods and, what is very important, provides a universal plug-in interface for adding new methods. Besides, the system contains some tools for data and formats transformation, access to databases, querying, and is capable of graphical presentations of some kinds of data mining results (trees, rules, and groups).

Descartes [1] automates generation of maps presenting user-selected data and supports various interactive manipulations of map displays that can help to visually reveal important features of the spatial distribution of data. Descartes also supports some data transformations productive for visual analysis, and has a convenient graphical interface for outlier removal and an easy-to-use tool for generation of derived variables by means of logical queries and arithmetic operations over existing variables. It is essential that both systems are designed to serve the same goal: help to get knowledge about data. They propose different instruments that can complement each other and together produce a synergistic effect.

Currently, Kepler contains data mining methods for classification, clustering, association, rule induction, and subgroup discovery. Most of the methods require selection of a target variable and try to reveal relationships between this variable and other variables selected for the analysis. The target variable most often should be discrete. Descartes can be effectively used for finding ”promising” discrete variables including, implicitly or explicitly, a spatial component. The following ways of doing this are available:

1. Classification by segmentation of a value range of a numeric variable into subintervals.
2. Cross-classification of a pair of numeric attributes. In both cases the process of classification is highly interactive and supported by a map presentation of the spatial distribution of the classes that reflects in real time all changes in the definition of classes.
3. Spatial aggregation of objects performed by the user through the map interface. Results of such an aggregation can be represented by a discrete variable. For example, the user can divide city districts into center and periphery or encircle several regions, and the system will generate a variable indicating to which aggregate each object belongs.

Results of most of the data mining methods are naturally presentable on maps. The most evident is the presentation of subgroups or clusters: painting or an icon can designate belonging of a geographical object to a subgroup or a cluster. The same technique can be applied for tree nodes and rules: visual features of an object indicate whether it is included in the class corresponding to a selected tree node, or whether a given rule applies to the object and, if so, whether it is correctly classified.

Since Kepler contains its own facilities for non-geographical presentation of data mining results, it would be productive to make a dynamic link between displays of Kepler and Descartes. This means that, when a cursor is positioned on an icon symbolizing a subgroup, a tree node, or a rule in Kepler, the corresponding objects are highlighted in a map in Descartes. And vice versa, selection of a geographical object in a map results in highlighting the subgroups or tree nodes including this object or marking rules applicable to it.

[1] See on-line demos in the Internet at the URL http://allanon.gmd.de/and/java/iris/

The above presented consideration can be summarized in the form of three kinds of links between data mining and cartographic visualization:

– From "geography" to "mathematics": using dynamic maps, the user arrives at some geographically interpretable results or hypotheses and then tries to find an explanation of the results or checks the hypotheses by means of data mining methods.
– From "mathematics" to "geography": data mining methods produce results that are then visually analyzed after being presented on maps.
– Linked displays: graphics representing results of data mining in the usual (non- cartographic) form are viewed in parallel with maps, and dynamic highlighting visually connects corresponding elements in both types of displays.

3 Scenarios of Integration

In this section we consider several examples of data exploration sessions where interactive cartographic visualization and different traditional methods of data mining were productively used together in data exploration.

3.1 Analysis with Classification Trees

In this session the user works with economic and demographic data about European countries [2]. He selects the attribute "National product per capita" and makes a classification of its values that produced interesting semantic and geographic clustering (Fig. 1). Then he asks the system to investigate how the classes are related to values of other attributes. The system starts the C4.5 algorithm and after about 15 seconds of computations produces the classification tree (Fig. 2).

It is important that displays of the map and the tree are linked:

– pointing to a class in the interactive classification window highlights the tree nodes relevant to this class (i.e. where this class dominates other classes);
– pointing to a geographical object on the map results in highlighting of the tree nodes representing groups including the object;
– pointing to a tree node highlights contours of objects on the map that form the group represented by this node (generally, in colors of classes)

3.2 Analysis with Classification Rules

In this session the user works with a database about countries of the world [3]. He selects the attribute "Trade balance" with an ordered set of values: Import much bigger than export, "import bigger than export", "import and export are

[2] The data have been taken from CIA World Book
[3] The data originate from ESRI world database

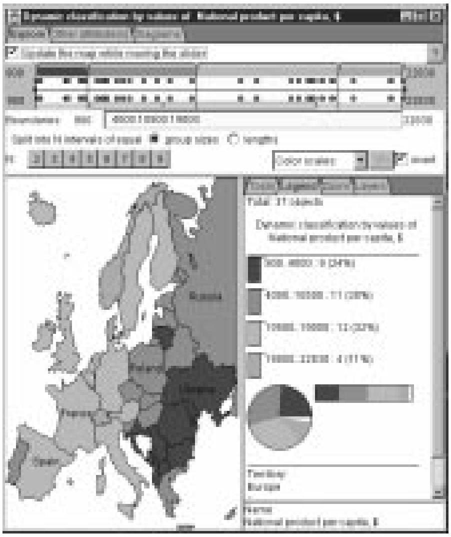
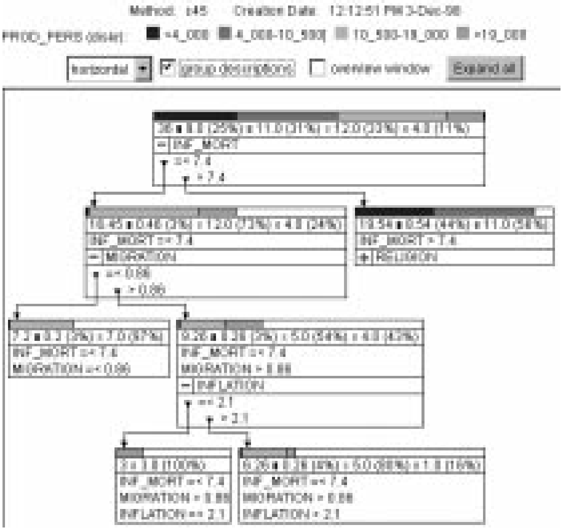

Dynamic classification by values of National product per capita, $
Update the map while moving the slider
Split into N intervals of equal ● group sizes ○ lengths
Color scales
Total: 31 objects
Dynamic classification by values of National product per capita, $
Russia
Poland
Ukraine
France
Spain
Territory: Europe
Name: National product per capita, $

Fig. 1. Interactive classification of values of the target attribute

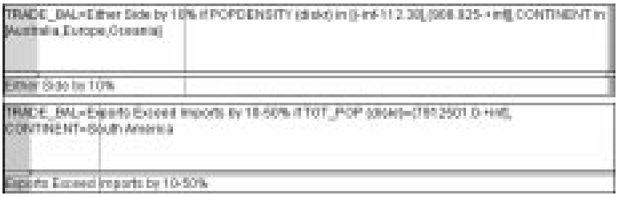

Method: c45 Creation Date: 12:12:51 PM 3-Dec-98

PROD_PERS (dollar): ■ <4_008 ■ 4_008-10_590] ■ 10_590-18_000 ■ >19_008

horizontal ☑ group descriptions □ overview window Expand all

Fig. 2. The classification tree produced by the C4.5 algorithm

approximately equal", "export bigger than import", and "export much bigger than import". He looks on the distribution of values over the World and does not find any regularity. Therefore, he asks the system to produce classification rules explaining distribution of values on the basis of other attributes. After short computation by the C4.5 method the user receives a set of rules. Two examples of the rules are shown in Fig. 3.



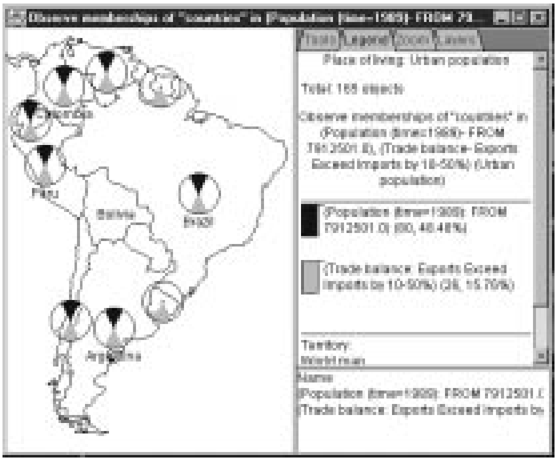
TRADE_BAL=Either Side by 10% if POPDENSITY (disk) in [-inf-112.30],[908.825-+inf], CONTINENT in [Australia,Europe,Oceania]

Either Side by 10%

TRADE_BAL=Exports Exceed Imports by 10-50% if TOT_POP (disk)=[7912501.0-+inf], CONTINENT=South America

Exports Exceed Imports by 10-50%

Fig. 3. Classification rules



Observe memberships of "countries" in [Population (time=1989)- FROM 79...
Tools
Legend
Zoom
Layers
Colombia
Peru
Bolivia
Brazil
Argentina

Place of living: Urban population

Total: 165 objects

Observe memberships of "countries" in (Population (time=1989)- FROM 7912501.0), (Trade balance- Exports Exceed Imports by 10-50%) (Urban population)

(Population (time=1989): FROM 7912501.0) (80, 48.48%)

(Trade balance: Exports Exceed Imports by 10-50%) (26, 15.76%)

Territory:
World map

Name
(Population (time=1989): FROM 7912501.(
(Trade balance: Exports Exceed Imports by

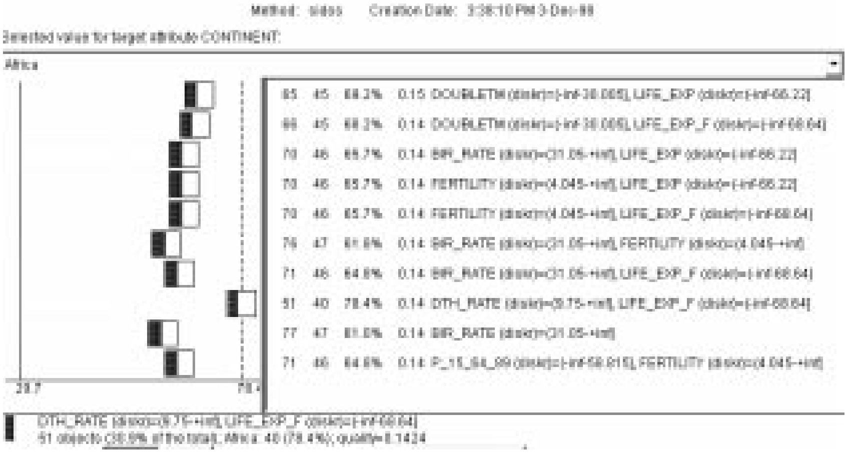
Fig. 4. Visualization of the rule for South America

For each rule, upon user's selection, Descartes can automatically produce a map that visualizes the truth values of left and right parts of the rule for each country. In this map it is possible to see which countries are correctly classified by the rule (both parts are true), which are misclassified (the premise is true while the consequence is false), and which cases remain uncovered (the consequence is true while the premise is false). Thus, in the example map in Fig. 4 (representing the second rule from Fig. 3) darker circle sectors indicate truth of the premise and lighter ones - truth of the concequence. One can see here seven cases of correct classification marked by signs with both sectors present and two cases of non-coverage where the signs have only the lighter sectors.

The user can continue his analysis with interactive manipulation facilities of maps to check the stability of relationships found. Thus, he can try to change boundaries of intervals found by the data mining procedure and see whether the correspondence between conditions in the left and the right parts of the rule will be preserved.

3.3 Selection of Interesting Subgroups

In this session the user wants to analyze the distribution of demographic attributes over continents. He selects a subset of these attributes and ran the SIDOS method to discover interesting subgroups (see some of the results in Fig. 5). For example, the group with "Death rate" less than 9.75 and "Life expectancy for female" greater than 68.64 includes 51 countries (31 % of the World countries), and 40 of them are African countries (78 % of African countries). To support the consideration of this group, Descartes builds a map (Fig. 6). The map shows all countries satisfying the description of the group. On the map the user can see specifically which countries form the group, which of them are in Africa and which are in other continents.



```
Method: sidos     Creation Date: 3:38:10 PM 3-Dec-98
Selected value for target attribute CONTINENT:
Africa

65  45  69.2%  0.15 DOUBLETM (disk)=(-inf-38.005], LIFE_EXP (disk)=(-inf-66.22]
66  45  68.2%  0.14 DOUBLETM (disk)=(-inf-38.005], LIFE_EXP_F (disk)=(-inf-68.64]
70  46  65.7%  0.14 BIR_RATE (disk)=(31.05-+inf], LIFE_EXP (disk)=(-inf-66.22]
70  46  65.7%  0.14 FERTILITY (disk)=(4.045-+inf], LIFE_EXP (disk)=(-inf-66.22]
70  46  65.7%  0.14 FERTILITY (disk)=(4.045-+inf], LIFE_EXP_F (disk)=(-inf-68.64]
76  47  61.8%  0.14 BIR_RATE (disk)=(31.05-+inf], FERTILITY (disk)=(4.045-+inf]
71  46  64.8%  0.14 BIR_RATE (disk)=(31.05-+inf], LIFE_EXP_F (disk)=(-inf-68.64]
51  40  78.4%  0.14 DTH_RATE (disk)=(9.75-+inf], LIFE_EXP_F (disk)=(-inf-68.64]
77  47  61.0%  0.14 BIR_RATE (disk)=(31.05-+inf]
71  46  64.8%  0.14 P_15_64_99 (disk)=(-inf-58.815], FERTILITY (disk)=(4.045-+inf]

DTH_RATE (disk)=(9.75-+inf], LIFE_EXP_F (disk)=(-inf-68.64]
51 objects (30.9% of the total); Africa: 40 (78.4%); quality=0.1424
```

Fig. 5. Descriptions of interesting subgroups

It is necessary to stress once again that Descartes does the map design automatically on the basis of the knowledge base on thematic data mapping. The subgroups found give the user some hints for further analysis: which countries to select for closer look; collection of attributes that best characterizes the continents; groups of attributes with interesting spatial co-distribution. Thus, if the user selects the pair of attributes cited in the definition of the considered group

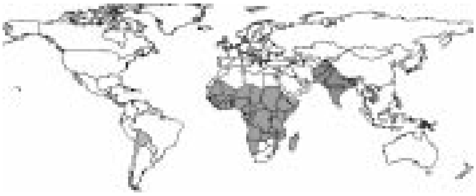
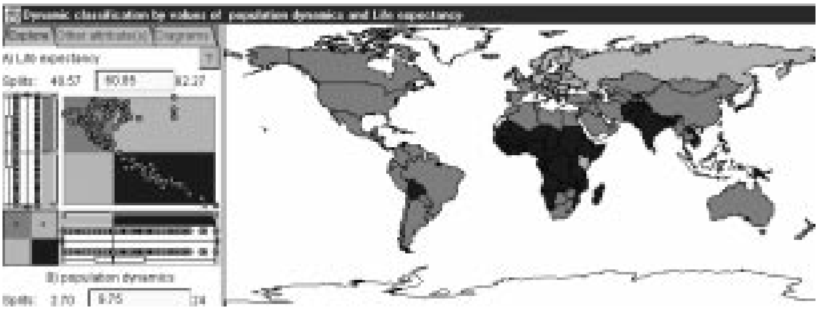


Fig. 6. Visualization of the subgroup

for further analysis, the system automatically creates a map for dynamic cross-classification on the basis of these attributes. The user may find other interesting threshold value(s) that leads to clear spatial patterns (Fig. 7).



Dynamic classification by values of population dynamics and Life expectancy

A) Life expectancy

Splits: 48.57 60.85 82.27

B) population dynamics

Splits: 2.70 9.75 24

Fig. 7. Co-distribution of 2 attributes: "Death rate" and "Life expectancy, female". Red (the darkest) countries are characterized by high death rate and low life expectancy. Green (lighter) countries have small death rates and high life expectancy. Yellow (light) countries are characterized by high death rate and high life expectancy.

3.4 Association Rules

In this session the user studies co-occurrence of memberships in various international organization. Some of them have similar spatial distributions. To find a numeric estimation of the similarity the user selected "association rules" method. The method produced a set of rules concerning simultaneous membership in different organizations. Thus, it was found that 136 countries are members of UNESCO, and 128 of them (94 %) are also members of IBRD. This rule was supported by visualization of membership on automatically created maps. One of them demonstrates members of UNESCO not participating in IBRD (Fig. 8).

Generally, this method is applicable to binary (logical) variables. It is important that Descartes allows to produce various logical variables as results of data analysis. Thus, they can be produced by: marking table rows as satisfying

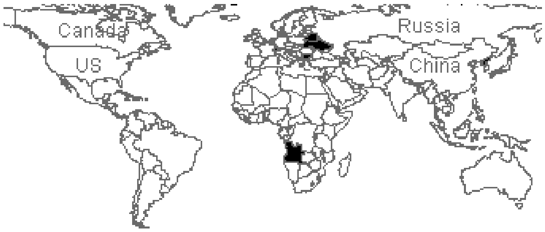
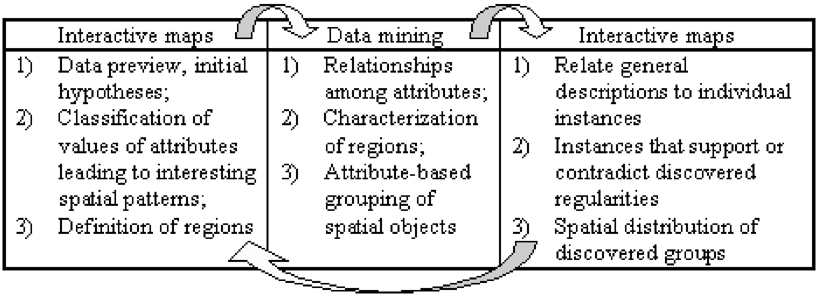


Fig. 8. Countries - members of UNESCO and non-members of IBRD

or contradicting some logical or territorial query, classifying numeric variables into two classes, etc. Association rules method is a convenient tool for analysis of such attributes.

3.5 Analysis of Sessions

It is clear that in all the sessions described above interactive visualization and data mining act as complementary instruments for data analysis. Their integration supported the iterative process of data analysis:



Interactive maps	Data mining	Interactive maps
1) Data preview, initial hypotheses;	1) Relationships among attributes;	1) Relate general descriptions to individual instances
2) Classification of values of attributes leading to interesting spatial patterns;	2) Characterization of regions;	2) Instances that support or contradict discovered regularities
3) Definition of regions	3) Attribute-based grouping of spatial objects	3) Spatial distribution of discovered groups

We should stress the importance of knowledge-based map design in all stages of the analysis. The ability of Descartes to automatically select presentation methods makes it possible for the user to concentrate on problem solving.

Generally, for the first prototype we selected only high-speed data mining methods to avoid long waiting time. However, currently there is a strategy in the development of data mining algorithms to create so called any time methods that can provide rough results after short computations and improve them with longer calculations. The open architecture of Kepler allows to add such methods later and to link them with map visualizations of Descartes.

One can note that we applied the system to already aggregated relatively small data sets. However, even with these data the integrated approach shows its advantages. Later we plan to extend the approach to large sets of raw data. The main problem is that maps are typically used for visualization of data aggregated over territories. A solution may be through automated or interactive aggregation of raw data and of results of data mining methods.

4 Software Implementation

The software implementation of the project is supported by the circumstance that both systems have client-server architecture and use socket connections and TCP/IP protocol for the client-server communication. The client components of both systems are realized in the Java language and provide the user interface. To couple the two systems, we implemented an additional link between the two servers. The Descartes server activates the Kepler server, establishes a socket connection, and commands Kepler to load the same application (workspace).

In the current implementation, the link between the two systems can be activated only in one direction: working with Descartes, the user can make Kepler apply some data mining method to selected data. A list of applicable methods is available to the user depending on the context (how many attributes are selected, what are their types, etc.). The selection of appropriate data analysis methods is done on the basis of an extension to the current visualization knowledge base existing in Descartes.

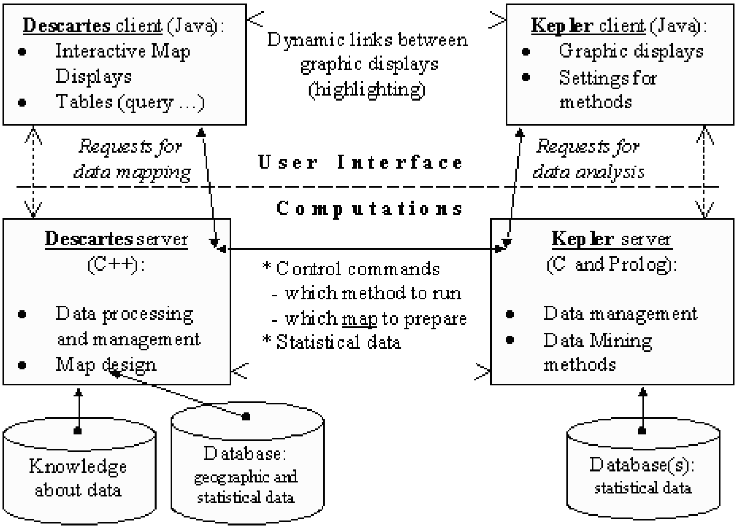


Fig. 9. The architecture of the integrated system

The link to data mining is available both from a table window and from some types of maps. Thus, classification methods (classification trees and rules) as well as subgroup discovery methods are available both from a table containing qualitative attribute(s) and from maps for interactive classification or cross-classification. The association rules method is available from a table with several logical attributes or from a map presenting such attributes.

When the user decides to apply a data mining method, the Descartes client allows him to specify the scope of interest (choose a target variable or value when necessary, select independent attributes, specify method-specific parameters, etc.) and then sends this information to the Descartes server. The server creates a temporary table with selected data and commands the Kepler server to import this table and to start the specified method. After finishing the computations, the Kepler server passes the results to the Kepler client, and this component visualizes the results. At this point a new socket connection between the Descartes client and the Kepler client is established for linking of graphics components. This link provides simultaneous highlighting of active objects on map displays in Descartes and graphic displays in Kepler.

Results of most data mining methods can be presented by maps created in Descartes. For this purpose the Kepler server sends commands to the Descartes server to activate map design, and the Descartes client displays the created maps on the screen.

5 Conclusions

To compare our work with others, we may note that exploratory data analysis has been traditionally supported by visualizations. Some work was done on linking of statistical graphics built in xGobi package with maps displayed in ArcView GIS [3] and on connecting clustering dendrograms with maps [10]. However, all previous works we are aware of utilize only a restricted set of predefined visualizations.

In our work we extend this approach by integrating data mining methods with knowledge-based map design. This allows us to create a general mapping interface for data mining algorithms. This feature together with the open architecture of Kepler gives an opportunity to add new data mining methods without system reengineering.

Acknowledgements

We are grateful to all members of GMD knowledge discovery group for numerous discussions about our work. Dr. H.Voss, D.Schmidt and C.Rinner made useful comments on the early version of the paper. We express our special thanks to Dr. D.Wettschereck (Dialogis GmbH) for the implementation of the link on Kepler's side.

References

1. Andrienko, G., and Andrienko, N.: Intelligent Visualization and Dynamic Manipulation: Two Complementary Instruments to Support Data Exploration with GIS. In: Proceedings of AVI'98: Advanced Visual Interfaces Int. Working Conference (L'Aquila Italy, May 24-27, 1998), ACM Press (1998) 66-75
2. Brodley, C.: Addressing the Selective Superiority Problem: Automatic Algorithm / Model Class Selection. In: Machine Learning: Proceedings of the 10th International Conference, University of Massachusetts, Amherst, June 27-29, 1993. San Mateo, Calif.: Morgan Kaufmann (1993) 17-24
3. Cook, D., Symanzik, J., Majure, J.J., and Cressie, N.: Dynamic Graphics in a GIS: More Examples Using Linked Software. Computers and Geosciences, **23** (1997) 371-385
4. Gama, J. and Brazdil, P.: Characterization of Classification Algorithms. In: Progress in Artificial Intelligence, Lecture Notes in Artificial Intelligence, Vol.990. Springer-Verlag: Berlin (1995) 189-200
5. Gebhardt, F.: Finding Spatial Clusters. In: Principles of Data Mining and Knowledge Discovery PKDD97, Lecture Notes in Computer Science, Vol.1263. Springer-Verlag: Berlin (1997) 277-287
6. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, **39** (1996), 27-34
7. John, G.H.: Enhancements to the Data Mining Process. PhD dissertation, Stanford University. Available at the URL http://robotics.stanford.edu/∼gjohn/ (1997)
8. Kodratoff, Y.: From the art of KDD to the science of KDD. Research report 1096, Universite de Paris-sud (1997)
9. Koperski, K., Han, J., and Stefanovic, N.: An Efficient Two-Step Method for Classification of Spatial Data. In: Proceedings SDH98, Vancouver, Canada: International Geographical Union (1998) 45-54
10. MacDougall, E.B.: Exploratory Analysis, Dynamic Statistical Visualization, and Geographic Information Systems. Cartography and Geographic Information Systems, **19** (1992) 237-246
11. Wrobel, S., Wettschereck, D., Sommer, E., and Emde, W.: Extensibility in Data Mining Systems. In Proceedings of KDD96 2nd International Conference on Knowledge Discovery and Data Mining. AAAI Press (1996) 214-219