GIS Visualization Support to the C4.5 Classification Algorithm of KDD

Gennady L. Andrienko and Natalia V. Andrienko
GMD - German National Research Center for Information Technology
Schloss Birlinghoven, Sankt-Augustin, D-53754 Germany
WWW: http://allanon.gmd.de/and/
Tel: +49-2241-142329, -142486 Fax: +49-2241-142072
E-mail gennady.andrienko@gmd.de

Abstract

Methods of Knowledge Discovery in Databases (KDD) are used for finding regularities and dependencies in collections of attribute data. These methods can be applied, in particular, to data referring to spatial objects. However, the general KDD methods do not make any respect of the spatial aspect of the data, and results of application of these methods are completely aspatial. Hence, to enable effective analysis of spatially referenced data, KDD methods should be supplemented by visualization of source data and obtained results in maps. An interactive map with special manipulation facilities allows the user to form input for a KDD algorithm that reflects certain aspects of spatial distribution of data in the appropriate for this algorithm form. A visualization module capable of automatic map generation enables viewing results of KDD procedures in the spatial context. The paper demonstrates a synergy of KDD and cartographic visualization achieved by integration of a KDD system KEPLER and an intelligent tool for visual exploration of spatially referenced data DESCARTES.

1 Introduction

The **C4.5** classification learning algorithm [Quinlan 1993] is one of the best known, efficient, and widely used algorithms of KDD (Knowledge Discovery in Databases). The main goal of the algorithm is to discover relationships between values of a target, or dependent attribute (which should be a qualitative attribute with 2 or more values) and those of some set of independent attributes (numeric or qualitative). Quite often the target attribute is derived from a numeric attribute by means of discretization, or division of the whole value range into subintervals and treating these subintervals as values. The output of the algorithm is a classification tree that can be transformed to a set of classification rules.

An example classification tree is shown in Figure 1. The tree has been generated from a table with economic and demographic data about the countries of Europe. The target attribute is PROD_PERS (gross national product per capita), a discretized numeric attribute. Its value range has been divided into 4 subintervals, which produced the values '<4000', '4000-10500', '10500-19000', '>19000'. Accordingly, the countries are divided into 4 classes, depending on to what interval the associated value of the target attribute fits. The tree shows us how these classes can be differentiated on the basis of values of other attributes (infant mortality, migration rate, inflation, dominant religion, etc.). In this way relationships between attributes are revealed.

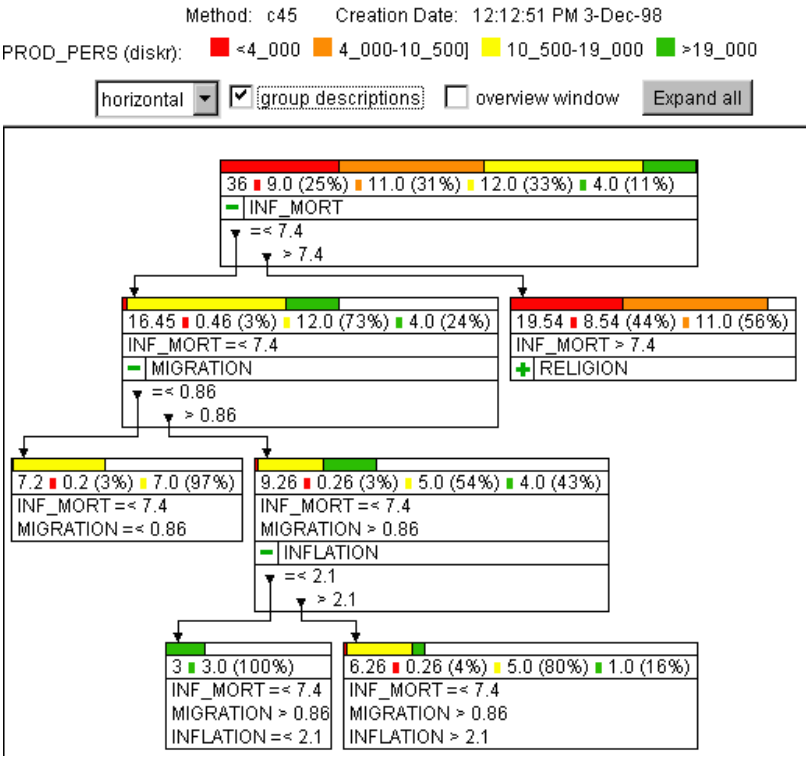

Figure 1. An example classification tree.

Findings of the **C4.5** algorithm strongly depend on (a) the initial partition of the value range of the target attribute, if it is numeric, and (b) the selected set of independent attributes. It is important to note that the knowledge discovery process is inherently iterative. Analysts usually apply some algorithm to data at hand, analyze the results, then change some settings (for example, the partition of the values of the target attribute and/or the set of independent attributes), and repeat. An interactive support of these activities and a convenient user interface are extremely important for the overall success of the exploration.

We are particularly interested in application of KDD techniques to spatially referenced data, like the above-mentioned data about countries. It seems obvious that general KDD algorithms, such as **C4.5**, alone are inadequate for analysis of spatial data:

- they do not take into account relative locations of objects and spatial relationships among the objects and between the objects and their surrounding (unless this information is somehow encoded in the form of 'usual' attributes);
- their results also lack spatial component and therefore are difficult to interpret.

We propose to compensate for these weaknesses through integration of KDD techniques with appropriately designed interactive cartographic software. Presentation of spatially referenced data on a map is an important instrument of analysis of spatial relationships and patterns in data distribution. A map that displays spatial properties of data distribution may help to generate a suitable for KDD attribute that encodes these properties. A map presenting KDD results in the spatial context greatly supports interpretation of these results.

The present paper describes a possible way of collaboration of KDD methods (on an example of **C4.5** algorithm) and interactive cartographic visualization facilities. The next section shows how a map can help in preparation of input and in interpretation of output of a general (not spatial data oriented) KDD algorithm such as **C4.5**. The requirements to the cartographic visualization software to be integrated with KDD techniques are formulated. Then we describe the implementation of the suggested synthesis of KDD and maps on the basis of two existing systems, KEPLER for knowledge discovery in databases and DESCARTES for visual analysis of spatially referenced data by means of cartographic visualization and interactive manipulation of maps. A sample scenario of data analysis using the linked systems is given. At the end the technical side of the link is outlined.

2 Interactive visualization support to KDD

As we have indicated in the introduction, mapping facilities can support KDD methods in application to spatially referenced data on two stages: (1) preparation of input for KDD techniques with regard to spatial properties of the data and (2) interpretation of KDD results.

2.1 Preparation of input for KDD techniques

All general KDD techniques deal with attributes the values of which are numbers (numeric attributes) or strings (qualitative attributes). Many of the techniques work only with qualitative attributes or, like **C4.5**, require the target attribute to be qualitative. Hence, to make KDD algorithms account for the spatial component of data, one should be able to represent, at least partially, this spatial component by conventional, preferably qualitative, attributes. This may be interactively done with the use of mapping software. Suppose that an analyst is supplied with a map or a collection of maps on a computer screen. The maps present the spatial objects the data under analysis refer to and visualize these data using methods of thematic cartography. There are a number of ways to produce attributes for KDD that reflect various aspects of spatial distribution of objects and data values:

- The analyst assigns marks to spatial objects according to their spatial properties (location, size, etc.) or surrounding. This can be done by manual selection or through the use of spatial queries provided in some GIS, e.g. through building buffer zones. The marks then become the values of the generated attribute. For example, city districts may be divided into 'center' and 'periphery', cities and towns may be marked as located in forest areas, in agricultural regions, or in large city agglomerations, and so on.
- The user divides the territory shown in a map into named regions. The objects are classified according to what region they fit in. From this classification an attribute is derived.
- The user interactively partitions the value range of a numeric attribute into subintervals. The objects are classified according to the subintervals the associated values fit in. The results of the classification are represented on a map, for example, by assigning colors to the intervals and painting objects in these colors. The analyst has tools to change the partition and can immediately observe the results of the changes in the map. S/he manipulates the partition until a spatially coherent pattern is formed in the map. In this way discretization of values of numeric attributes can be done with respect to spatial distribution of the values.

It is necessary to note that the problem of appropriate discretization of numeric attributes is important not only for KDD. It is traditionally addressed in thematic cartography (the procedure of discretization is called in cartography 'classification'). Some cartographic presentation methods, e.g. the choropleth map, are based on previous classification of numeric data. It is well known that selection of different classification schemes results in very diverse views on the same data set [MacEachren 1994]. Once the problem of 'right' classification was decided to be completely solved after Jenks had developed his algorithm of 'optimal classification' [Jenks 1977]. However, statistically optimal solutions are rarely spatially and semantically meaningful. Therefore we find it important to allow the user to change a classification interactively and to observe the results on the map.

Besides the task of generating attributes to reflect spatial features of a data set, preparation of input for KDD techniques includes selection of source and target attributes. Cartographic visualization can provide a basis for the selection. Using data mapping tools, the analyst may obtain presentations of value distributions of various attributes. If distribution of some attribute is found spatially interesting, the analyst can investigate relationships between this attribute and the others using KDD techniques. In particular, this attribute or its discrete derivative can be selected as the target attribute of the **C4.5** method. Cartographic presentation can also help to select the independent attributes for this or other method. These may be the attributes showing some similarities in distribution to that of the target attribute.

2.2 Interpretation of results of KDD

When KDD procedures are applied to spatially referenced data, it is insufficient to view their results in the traditional representation form. For example, a node in a classification tree resulting from the **C4.5** method does not give any idea about locations or spatial proximity of the objects described by this node, whereas such information is very important when one deals with spatially referenced data. To be able to interpret KDD results in the spatial context, it is necessary to have them visually represented on maps.

Results of work of KDD methods most often have one of the following forms:

- classification tree;
- rules;
- groups of objects, e.g. by similarity.

A node of a classification tree or the left part of a rule contains a logical expression concerning values of one or more attributes. In a case of analysis of data associated with spatial objects it may be necessary to see in a map the objects satisfying this logical expression. With a tree node, the map should also visualize the values of the target attribute for these objects. With a rule, the map should show whether the expression in the right part of the rule is true for each of the objects. If the analyst received groups, s/he would be interested to see the spatial distribution of objects belonging to the groups. So, one can see that in each case the map is required to present some group(s) of objects and, possibly, some data associated with the objects of the group(s).

Two things complicate presentation of KDD results on a map. First, the groups to be shown are typically rather numerous. Second, they usually greatly overlap. Therefore in most cases it is impossible to represent the whole output in a single map so that the map is legible and productive for analysis. A feasible solution is generation of maps presenting parts of the output. However, it is unlikely that the user would be happy to receive at once as many maps as there are tree nodes or rules, and it is even more unlikely that s/he will be able to manage these maps and make any use from them. Hence, a map visualizing a part of the output should appear only on demand, when the analyst focuses on this part in her/his investigation of the KDD results. It would be very convenient for the user if the map is automatically generated when s/he points at a tree node, a rule, or a group s/he would like to analyze in the spatial context. Another possible solution is a dynamic link between the presentation of KDD results with a map showing all the objects. When the user points at a part of the KDD output, the objects belonging to the corresponding groups are highlighted in the map. The link can be bi-directional: when the user clicks on some object in the map, the nodes/rules/groups it fits in are highlighted in the presentation of the KDD results.

2.3 Requirements to visualization software to support KDD application to spatial data

The requirements to the visualization software can be in an evident way derived from the above-described role it is expected to play when integrated with KDD procedures:

- A map displayed on the screen by this software should allow direct user's manipulation to mark objects and divide a territory into regions.
- The software should include facilities for interactive discretization of numeric attributes with immediate reflection of any changes in a map.
- It should be possible to view two or more maps simultaneously to compare spatial distributions of values of several attributes.

- The visualization module should be able to generate maps automatically on demand. This feature allows the user to concentrate on data analysis instead of being concerned how to present the data or results of KDD procedures.
- The visualization system should intelligently apply the established principles of graphic and cartographic presentation to ensure that the maps appropriately and effectively encode the information they represent.

A suitable candidate for a visualization component in the KDD-cartography complex is our system DESCARTES [Andrienko and Andrienko 1998a and 1999, see also a demo version of the system running in WWW at the URL http://allanon.gmd.de/and/java/iris/]. On the basis of its knowledge base on cartographic presentation, the system automatically builds maps applying proper visualization techniques to user-selected data depending on their characteristics. DESCARTES includes easy to use interactive tools that allow the user to change visual appearance of maps and supplementary graphics. In particular, there are map-mediated tools for interactive classification (discretization) of a numeric attribute (see Figure 2) and for cross-classification of two numeric attributes. In the system it is possible to link several presentations together by simultaneous highlighting of the same objects. These features satisfy our requirements to the cartographic visualization system to support the knowledge discovery process.

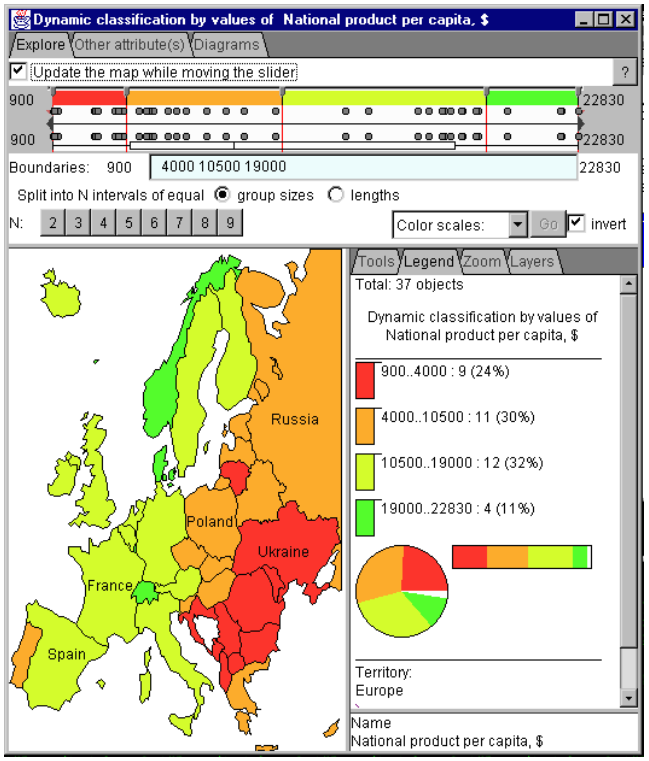

Figure 2. The interface of the tool for interactive classification in DESCARTES.

A pilot implementation of the KDD-cartography link is done by integrating DESCARTES with the system KEPLER [Wrobel et al., 1996] which is a general KDD tool having a plug-in architecture (this means that it is possible to add new methods to the system without software reengineering).

3 An example analysis scenario

Suppose that a hypothetical user wishes to analyze economic and demographic data about the countries of Europe. She lets DESCARTES visualize the distribution of values of the attribute 'gross national product per capita'. On a map built by the system the analyst observes a clear-cut division into east and west. She decides to investigate which of the other attributes are related to this attribute and show similar properties in spatial distribution of their values. For this purpose the analyst chooses to use the **C4.5** algorithm with 'gross national product per capita' as the target attribute.

The selected attribute is numeric, and it is therefore necessary to discretize it. This is done with the tool for interactive classification, as shown in Figure 2. The user changes the number and boundaries of the intervals until the resulting classes of objects form prominent regions seen in the map. Then she commands the system to generate an attribute with values corresponding to the classes and to start the **C4.5** method with this attribute as the target and all the other attributes as the independent ones. On this request DESCARTES generates a temporary table with a column containing values of the new attribute and passes this table to KEPLER along with an instruction to start the **C4.5** method. KEPLER starts the corresponding procedure and passes it the data received from DESCARTES. The result of the work is the classification tree shown in Figure 1. Activation of KEPLER, data transfer, and the work of the **C4.5** method take about 15 seconds (measured on a PC Pentium II 266MHz with tables containing from 30 to 300 rows and from 10 to 40 columns).

After the classification tree is displayed on the screen, the user can click on its nodes and receive map presentation of objects satisfying the conditions shown inside the nodes. Thus, the map in Figure 3 presents geographic distribution of the countries corresponding to the left second-level node of the tree shown in Figure 1. This node contains the condition 'INF_MORT=<7.4' (infant mortality rate is less than or equal to 7.4 deaths per 1000 live births). It is well visible that the countries satisfying this condition are all in the Western Europe and have very high values of gross national product per capita (the minimum is 13120 $). Colors in which the countries are painted are those assigned to the classes resulting from the discretization of the target attribute.

When the user clicks on another node, a new map representing the node selected last replaces the previous map. The user may also preserve the previous map from being removed from the screen (using a special checkbox in the 'Tools' tab of the map window). This makes it possible to compare maps corresponding to two or more nodes.

It is also possible to see how the **C4.5** algorithm further subdivides the countries belonging to the selected node. The considered node (the left second-level node in Figure 1) shows, besides the condition 'INF_MORT=<7.4', that the next attribute taken for the classification is 'MIGRATION' (net migration rate). The group of countries represented by the node is subdivided into countries with the migration rates below or equal to 0.86 and those having the migration rates higher than 0.86. When DESCARTES is requested to represent this division, it builds the map shown in Figure 4. In this map the countries satisfying the condition of the node are, like in the map in Figure 3, painted in the colors of the corresponding intervals of the target attribute. Dark superimposed circles mark the countries with net migration rates below or equal to 0.86. It is well seen that all the marked countries belong to the third class, with respect to the target attribute. These 7 countries become separated from the rest of the group on the next tree level.

The attributes selected by the **C4.5** algorithm are the ones enabling effective division of the countries into the classes derived from the discretization of the target attribute. Hence, there should be some relationships between these attributes and the target attribute. In order to study these relationships, the analyst may be interested to see how values of the attributes selected by **C4.5** are distributed across the classes. This is possible within DESCARTES. The tool for interactive classification (Figure 2) offers the user an opportunity to investigate relationships of the base attribute of the classification (in this case, the one used as the target for **C4.5**) with other attributes. The user can select one or more attributes, and the system will then calculate the statistics of value distribution of these attributes for each class. The statistical results are presented to the user in the form of averaged 'portraits' of the classes and by 'box and whiskers' plots (see Figure 5). All the statistics are immediately recalculated when the classes change.

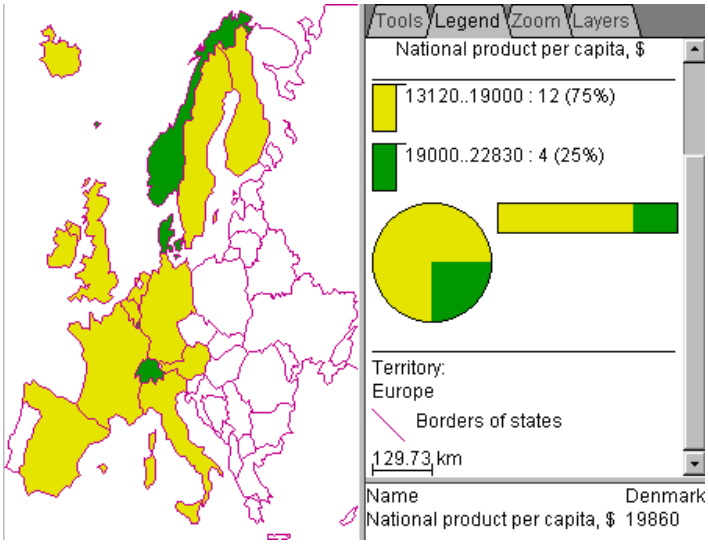

Figure 3. Map presentation of a tree node.

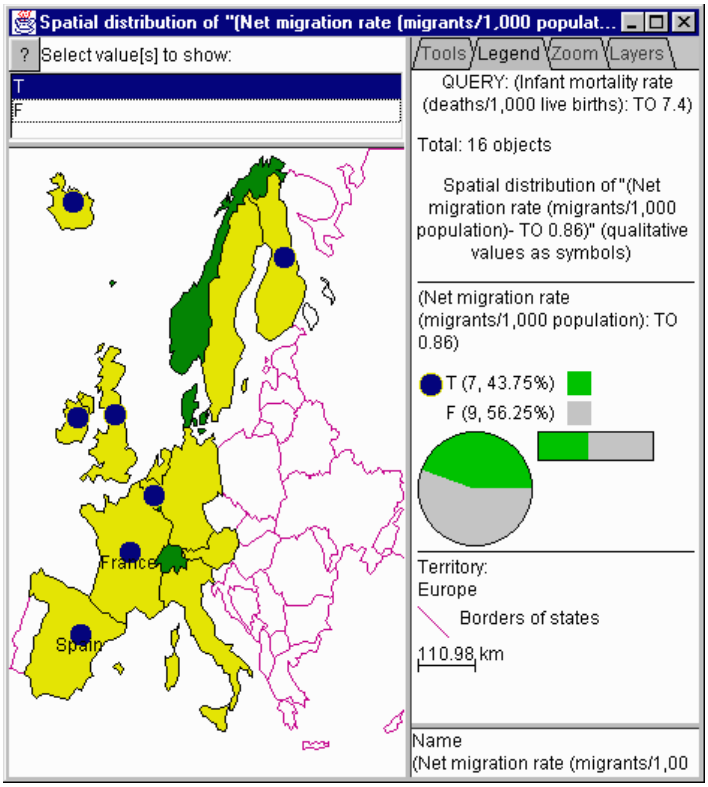

Figure 4. Further division of the tree node presented in Figure 3.

Our hypothetical expert selects the attributes from the lowest-level nodes of the left branch of the classification tree: infant mortality rate, net migration rate, and inflation rate. DESCARTES immediately creates a new window titled 'Group statistics' (see Figure 5). In the rectangles painted in the colors corresponding to the classes one can see diagrams representing average values of the selected attributes for the classes. The system has used the 'radial bars' presentation technique because the attributes are incomparable (the principles of selection of presentation methods in DESCARTES are explained in [Andrienko and Andrienko 1998a and 1999]). Below the averaged 'portraits' of the classes there are 'box and whiskers' plots showing variations of values within classes. The exact minimum, maximum, mean, median, and quartile values for each class are shown to the user after she clicks on the corresponding rectangle in the 'Group statistics' window (see the two windows at the bottom of the Figure 5).

Viewing the group statistics convinces the analyst that the attributes selected by the **C4.5** method are indeed related to the target attribute.

Upon a request, the system can also represent individual values of these attributes for each country by diagrams superimposed on the classification map. This will, in particular, allow the analyst to find which country in the third class has the value of net migration rate unusual for the class and causing a deviation from the general trend (demonstrated by the corresponding 'box and whiskers' plot). This country is Ireland with net migration equal to –2.22.

For comprehensive data analysis one usually needs to apply a KDD method iteratively. In our case, the user can modify the partition of the value range of the target attribute and to study the impact of this change on the results of classification.

In the described hypothetical scenario the interactive maps and facilities provided by DESCARTES have supported the following activities of the analyst:

1. Preparation to analysis by means of KDD techniques.
   - 1.1. Data preview.
   - 1.2. Orientation of the investigation, i.e., in our case, selection of the target attribute.
   - 1.3. Discretization of values of the target attribute with respect to their spatial distribution.
2. Study and interpretation of results of KDD procedures.
   - 2.1. Viewing the spatial distribution of the groups formed by the applied KDD algorithm.
   - 2.2. Passing from general group descriptions to individual instances, i.e. seeing what spatial objects form this or that group.
   - 2.3. Detecting the instances that contradict the discovered regularities.
   - 2.4. Probing the relationships between attributes found by means of KDD.

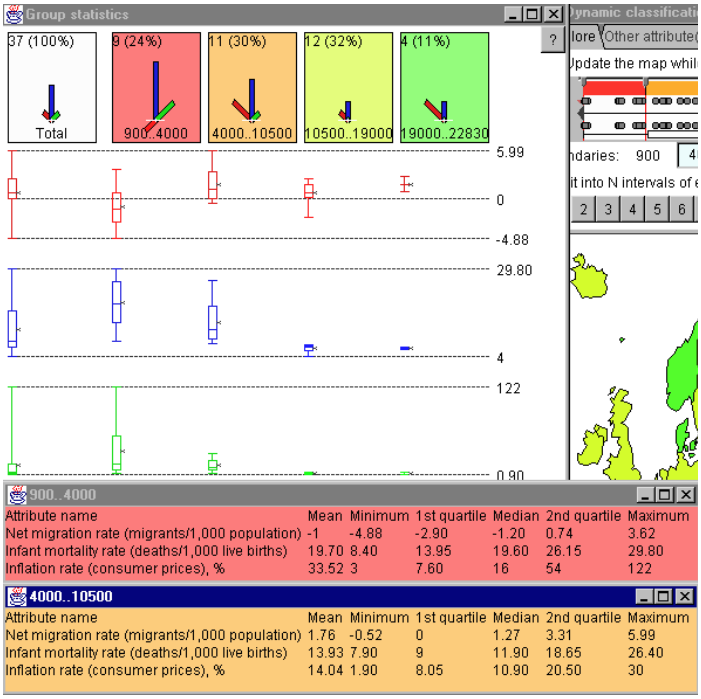

Figure 5. The summary statistics characterizing distribution of values of the attributes from the lowest-level nodes of the classification tree (Figure 1) across the classes of objects formed according to the values of the target attribute.

We continue the work on integration of interactive maps with KDD methods in our project of an integrated environment for knowledge extraction from spatially referenced datasets [Andrienko and Andrienko 1998b].

4 Implementation

Both DESCARTES and KEPLER have client-server architectures. The server parts perform data management and processing, computations, and (in DESCARTES) map design. The client parts provide the user interface. The server of DESCARTES is implemented in C++, and that of KEPLER – in Prolog and C. The client parts are both implemented in Java. In each system the client and the server communicate through a TCP/IP socket connection. To make the two systems work together, we implemented additional links between the servers and between the clients. The architecture and the functioning of the resulting complex are schematically shown in Figure 6.

The systems DESCARTES and KEPLER are commercially available from Dialogis GmbH (http://www.dialogis.com/).

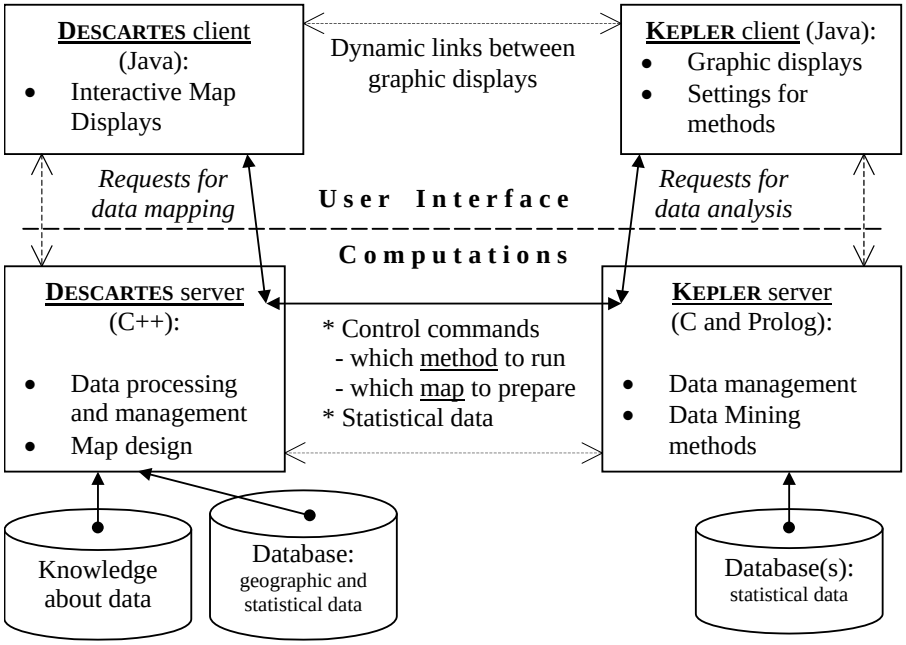

Figure 6. The architecture of the integrated system

Acknowledgments

We are grateful to all members of the Knowledge Discovery research team (GMD, Institute for Autonomous Intelligent Systems) for numerous discussions concerning our work. We express our special thanks to Dr. D.Wettschereck (Dialogis GmbH) for the help in the implementation of the link between the systems from the KEPLER's side.

References

Andrienko, G.L., and Andrienko, N.V. (1998a). 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, New York, pp.66-75

Andrienko, G.L., and Andrienko, N.V. (1998b). Knowledge Extraction from Spatially Referenced Databases: a Project of an Integrated Environment. Unpublished position paper presented on Varenius Workshop on Status and Trends in Spatial Analysis (Santa-Barbara, December 1998), Available at the URL http://allanon.gmd.de/and/sb98/gis-kdd.html

Andrienko, G.L., and Andrienko, N.V. (1999). Interactive Maps for Visual Data Exploration. International Journal of Geographical Information Science, 13(4).

Jenks, G.F. (1977). Optimal Data Classification for Choropleth Maps. Occasional Paper No.2, Department of Geography, University of Kansas.

Quinlan, J.R. (1993). C4.5 Programs for Machine Learning. Morgan Kaufmann, 1993

MacEachren, A.M. (1994). Some Truth with Maps: a Primer on Symbolization & Design. Washington: Association of American Cartographers.

Wrobel, S., Wettschereck, D., Sommer, E., and Emde, W. (1996). Extensibility in Data Mining Systems. In Proceedings of KDD'96 2nd International Conference on Knowledge Discovery and Data Mining. AAAI Press, pp.214-219