Data Mining with C4.5 and Interactive Cartographic Visualization

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

Abstract

In the paper we consider application of techniques of knowledge discovery in databases (KDD) to spatially referenced data. We propose to combine such techniques with various methods of interactive classification of spatial objects supported by map displays. Maps help both to prepare source data for KDD procedures and to interpret results of their application. The latter function is realized through dynamic linking of spatial and non-spatial displays.

1. Introduction

In this paper we focus on analysis of thematic (attribute) data referring to units of territorial division, such as numbers of occurrences of some disease in municipalities of a country. Very often an analyst needs to detect territorial clusters of "similar" objects and study them in respect to other attributes such as incomes, ages, environmental conditions etc.

Usually spatial objects shown in a map are grouped into classes due to discretization of the value range of a numeric attribute: the whole range is divided into subintervals, which are treated as values. Practically this means that objects with values of the attribute fitting in the same subinterval are represented in the same way in the map, e.g. painted in the same color. It is obvious that even a slight change of subinterval boundaries can result in a very different view on the same data set [1]. Some guidelines for "right" discretization and even an algorithm of "optimal classification" [2] were developed in cartography. Still it is recognized that a statistically optimal classification is rarely spatially and semantically meaningful. Instead of any predefined static classification, it is more promising to apply interactive visual facilities that allow an analyst to manipulate classes and immediately observe the resulting changes in a map in seeking for interesting spatial patterns [3]. We consider several variants of interactive classification in section 2.

Since an interesting classification is found, an analyst usually tries to relate it to other attributes. Purely visual means of analysis are insufficient for this purpose. A promising approach is to employ here the potential of data mining techniques (also referred to as techniques for knowledge discovery in databases, or KDD). We shall demonstrate the possible contribution of general KDD methods (i.e. not specifically designed for spatially referenced data) on the example of the C4.5 classification learning algorithm [4]. The main goal of the algorithm is to discover relationships between a given classification of objects and a set of attributes. The output of the algorithm is a classification tree showing how objects may be assigned to the given classes on the basis of values of these attributes (see an example in Figure 4).

The knowledge discovery process is inherently iterative. An analyst prepares data for application of a data mining procedure (e.g. produces a classification and selects attributes to test for relationships with the classes), then runs the procedure and analyzes the results, then changes some settings (e.g. the classification or/and the set of attributes), and repeats. It is necessary to equip the analyst with convenient facilities to support this iterative process. When data to analyze are spatially referenced, map presentation is required both to prepare them for KDD (e.g. to produce meaningful classes) and to interpret the results of calculations. Usually such results have completely aspatial form. For example, a classification tree gives no information about the spatial aspect of the data set characterized by it. Consideration of the spatial aspect can be enabled through dynamic linking of non-cartographic displays of KDD results with maps. This linking is described in section 3 whereas the next section is devoted to cartographic support for data preparation.

2. Interactive visualization to prepare data for data mining

Seeking to achieve a synergy of two approaches to exploration of spatial data, visual analysis with the use of interactive cartographic displays and KDD methods, we decided to integrate two existing systems. Descartes [5] is a system for cartographic visualization (see also an online version in the WWW [6]). It is able to automatically choose proper visualization techniques for user-selected

data. The user is given interactive tools to manipulate the map displays in order to make them more expressive and thereby reveal interesting features of spatial distribution of data. Kepler [7] is a general KDD tool with plug-in architecture (i.e. it is possible to add new procedures to the system without software reengineering). One of the existing plug-ins is the C4.5 method considered here.

Descartes enables a number of ways to classify spatial objects with a support of interactive map displays:

- interactive discretization of a single numeric attribute (Figure 1);
- interactive cross-classification on the basis of two numeric attributes (Figure 2);
- classification according to a dominant attribute value among several attributes (Figure 3);
- interactive regional aggregation of objects in the map.

Some other methods of map-mediated classification are potentially possible. Suitable for this purpose are spatial queries supported in advanced GIS, for example, "select objects on the coastline" or "select objects lying in 3 miles distance from the road". In Descartes classifications are also automatically done according to values of qualitative attributes or can be produced by the user on the basis of logical queries.

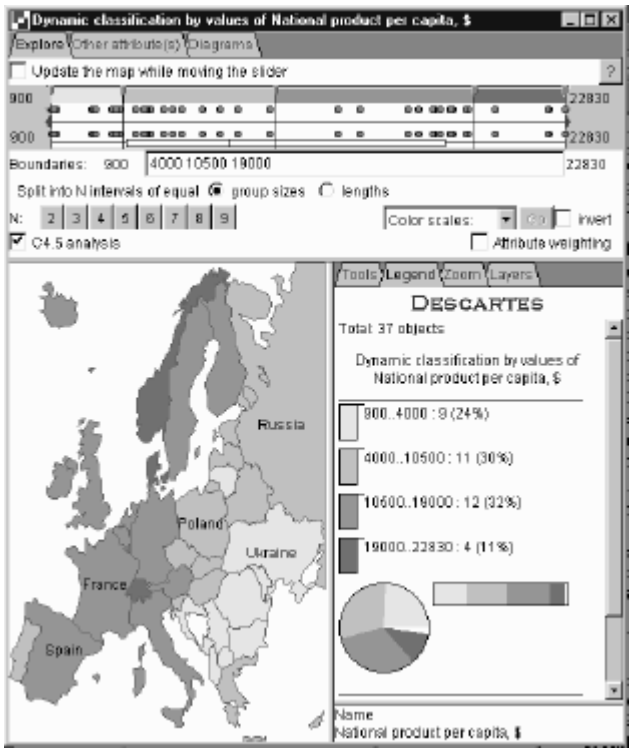

Figure 1. Interactive classification of European countries according to values of the numeric attribute "National product per capita"

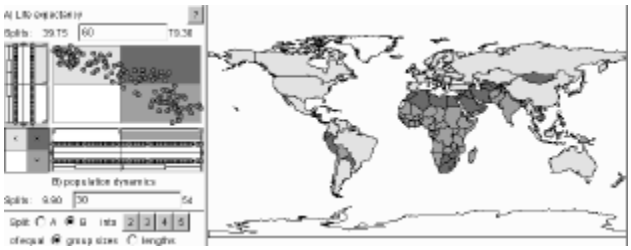

Figure 2. Cross-classification of countries of the world on the basis of two attributes, "Birth rate" and "Life expectancy". The light shade corresponds to countries with high life expectancy and small birth rate, middle dark to countries with small life expectancy and high birth rate, and the darkest to high values of both attributes. The scatter-plot on the left is dynamically linked with the map through simultaneous highlighting of objects pointed at by the cursor either in the map or in the plot.

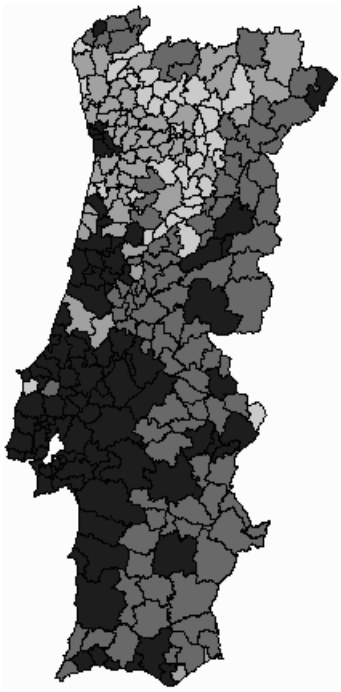

Figure 3. Classification of the municipalities of Portugal according to domination of different age groups. In the map one can observe prominent regions with domination of middle-age population (the darkest shade, mostly in the southwest), of old population (a bit lighter shade, in the east), and of children and young population (the light shades, in the northwest).

In any case of classification the classes together with other data about the classified objects can be transmitted to Kepler for application of KDD methods. Technically, Descartes builds a table to be imported by Kepler. One column of the table contains the identifiers of the spatial objects. The classes they belong to are indicated in

another column. The user selects the attributes from the source data set that should participate in data mining, and the values of these attributes are included in the table.

Results of C4.5 method have the form of classification tree that is displayed in Kepler as shown in Figure 4.

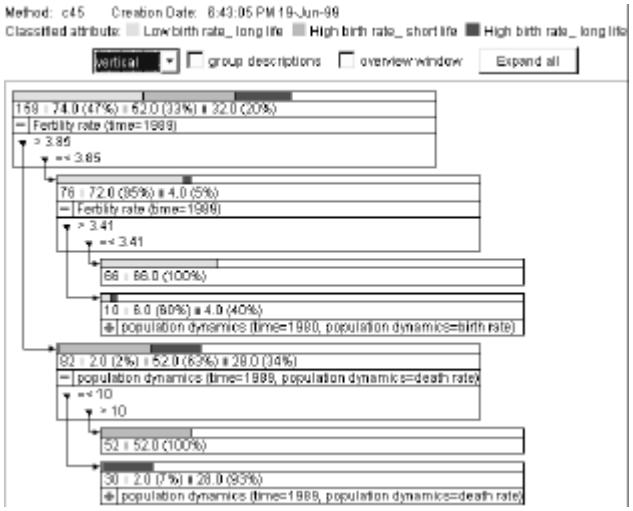

Figure 4. Classification tree calculated for the map in Figure 2. It shows that all 52 countries with high birth rate and small life expectancy can be characterized by fertility rate higher than 3.85 and death rate more than 10.

3. Map-supported analysis of KDD results

A tree resulting from C4.5 contains no information about spatial features of the classified objects. To properly analyze the results got, one should combine consideration of the tree with map visualization. The most natural and straightforward thing that is needed for analysis is the possibility to highlight in the map the objects fitting in any selected node of the tree.

Visualization capabilities of Descartes allow a number of other interesting analyses inspired by KDD results received. Thus, the attributes mentioned in the nodes of a tree are the ones that allow the most consistent with the specified classes partition of the objects. So, they are, probably, somehow related to the classification. The analyst can examine these attributes in Descartes using their map presentation or system-provided summary statistics of distribution of values across the classes (Figure 5).

It is also possible to see how members of a class are distributed among tree nodes. Thus, the map in Figure 6 shows the countries belonging to the class "low birth rate and high life expectancy". Painted in the lighter shade are the countries with fertility rates below 3.41, i.e. those fitting in the 3rd tree node in Figure 4 (counting from the top). Darker painting marks the members of the class that do not fit in this node.

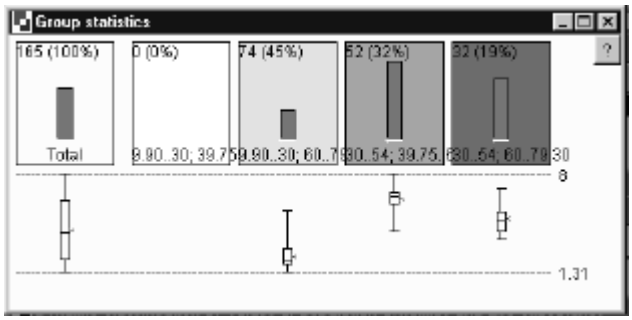

Figure 5. Distribution of values of "Fertility rate" (the attribute from the root node of the tree in Figure 4) across the classes shown in Figure 2.

Figure 6. The countries with low birth rates and high life expectancies in relation to fertility rates.

Results of C4.5 computations may be also represented as a set of classification rules. For example, Figure 7 shows two selected rules from the rule set produced by C4.5 for the classification of countries according to trade balance. The general form of the rule is "<class> if <condition>", where <condition> consists of logical statements about values of attributes.

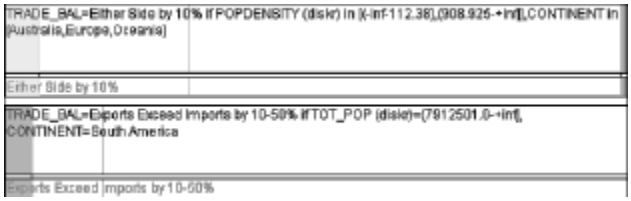

Figure 7. Classification rules.

Rules can also be analyzed with support of mapping facilities of Descartes. For any selected rule the system can visualize the satisfaction of its left and right parts by objects. In the map it is possible to see which objects are correctly classified by the rule (both parts are true), which are misclassified (the premise, i.e. condition, is true while the consequence, i.e. assignment to the class, is false), and which class members remain uncovered (the consequence is true while the premise is false). For example, Figure 8 shows a fragment of the visualization of the second rule from Figure 7. The signs mark countries for that either the premise or the consequence of the rule is true. Presence of

a lighter sector in a sign indicates truth of the premise. A darker sector shows that the consequence is true. There are seven cases of correct classification marked by signs having both sectors and two cases of non-coverage where the signs have only dark sectors.

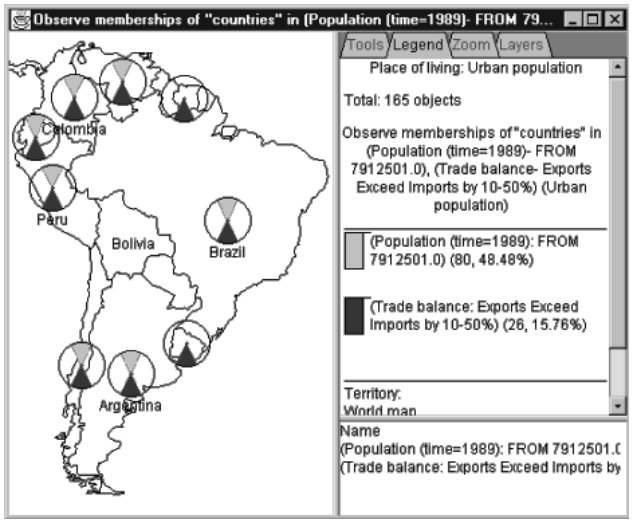

Figure 8. Satisfaction of the premise and the consequence of a classification rule by countries of South America.

Map-supported analysis of KDD results is facilitated by means of dynamic linking between non-cartographic displays in Kepler and maps in Descartes:

- pointing at a geographical object in a map results in tree nodes including the object being highlighted;
- pointing at a tree node highlights in the map the objects fitting in this node.

4. Conclusions

The examples considered support the idea that cartographic visualization and KDD are complementary instruments that can be used for analysis of spatial data. Interactive maps allow:

- data preview;
- formation of initial hypotheses;
- map-mediated discretization of attributes that produces interesting spatial patterns;
- a number of other ways to classify spatial objects.

Through these operations an analyst can select and prepare data for application of KDD methods. These methods help to reveal relationships among attributes, the task that can hardly be done using purely visual means.

Since results of general KDD methods are usually aspatial, they need to be related to spatial context. This is done through map presentation of findings obtained (related attributes, subgroups of objects, conditions in tree nodes or rules etc.) and through dynamic linking between graphical displays of KDD results and maps.

5. Implementation notes

Both Descartes and Kepler have client-server architectures. The server of Descartes is a C++ program, that of Kepler is written in Prolog and C. Both clients are implemented in Java. Descartes server sends commands to Kepler server via a socket connection. Data to be processed by Kepler are stored in files. To achieve the dynamic link between graphical displays, the clients communicate through another socket connection.

The integrated system is available for Windows and Unix platforms.

6. Acknowledgments

We are grateful to all members of the AiS.KD research team for numerous discussions concerning the content of the work. Our special thanks to Dr. D. Wettschereck (Dialogis GmbH) and Dr. A. Savinov (GMD) for the help in the implementation from Kepler side.

7. References

[1] MacEachren, A.M., *Some Truth with Maps: a Primer on Symbolization & Design*. Association of American Cartographers, Washington, 1994.

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

[3] Egbert, S.L. and Slocum, T.A. "ExploreMap: an exploration system for choropleth maps", *Annals of the Association of American Geographers*, **82,** pp.275-288, 1992

[4] Quinlan, Q.R., *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1993.

[5] G. Andrienko and N. Andrienko, "Interactive Maps for Visual Data Exploration", *International Journal of Geographical Information Science*, 13 (4), 1999, pp.355-374.

[6] G. Andrienko and N. Andrienko, "Descartes: knowledge-based system for visual data exploration", URLs http://allanon.gmd.de/and/java/iris/ and http://ais.gmd.de/descartes/IcaVisApplet/, 1999.

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