Specific Usage of Visual Data Analysis Techniques

Snezana Savoska[1] and Suzana Loskovska[2]

[1] Faculty of Administration and Management of Information systems,
Partizanska bb, 7000, Bitola, Republic of Macedonia
[2] University, Ss. Cyril and Methodius",
Faculty of Electrical Engineering and Information Technologies,
Karpos 2 bb, 1000, Skopje, Republic of Macedonia

Abstract. The visualization techniques are very important tools for data mining processes. They are widely applied in many areas especially in supporting decision making processes. We use visualization tools for rule generation, classification and clustering. The paper presents application of data visualization techniques and tools for generation of association rules, classification and clustering.

Keywords: Visualization, Data mining, Rule generation, Classification, Clastering.

Introduction

Data mining processes are computer intensive and algorithm dependent processes. Visualization tools may be very useful in solving data mining problems. Today's information flow demands the use of special algorithms for data analysis and data mining. The data are often automatically recorded by sensors and monitoring systems, cash and credit card paying machines etc. For all items, many variables are recorded, resulting in data with a high dimensionality. The data are collected because people believe that it is a potential source of valuable information, providing new insights or a competitive advantage [4]. But, finding valuable information hidden in the data, however, is a difficult task. Information visualization tools and visual data analysis can help to deal with the flood of information. A great advantage of visual data exploration is the direct involvement of the user.

Visual data mining integrates the human in the data analysis process. The human perceptual abilities help the analysis of today's large data sets [1]. Visual data mining is especially useful when little is known about the data and when the exploration goals are vague. Visual data exploration can be seen as a hypothesis generation process where the visualizations of the data allow setting new hypotheses. The main advantages of Visual data exploration (VDE) is that it can easily deal with highly non-homogeneous and noisy data, it is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters and can provide a qualitative overview of the data, allowing data phenomena to be isolated for further quantitative analysis.

VDE allows a faster data exploration and a much higher degree of confidence in the exploration findings. These facts lead to a high demand for visual exploration techniques. Visual techniques are especialy applied to support data

decision proces. There are a number of vizualization techniques that have been developed for data mining tasks. These data mining tasks include association rule generation, classification and clastering.

1 Generation of Association Rules

Association rules are statistical relations between two or more items in the data set. The most important usage of this data mining methods is the supermarket basket application. The goal of association rule is to find interesting patterns and trends in databases. It is also important to find out rules in 70% of cases and define transaction with some probability, called confidence. A second important parameter is the rule support, defined as percentage of co-occurence of the transactions items. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subseteq I, Y \in I, X, \neq \$$. The confidence c is defined as the percentage of transactions that contain Y for given X. The support is the percentage of transactions that contain both X and Y.

Visualization techniques are used to provide an interactive selection of rule support and confidence levels. Figure 1 shows SGI Sets Rule Visualizer [6, 2] which maps the left and right sides of the rules to the x- and y-axes of the plot, respectively. It shows the confidence as the height of the bars and the support as the height of the discs. The color of the bars shows the interestingness of the rule. Figure 2 shows two alternative visualizations called mosaic and double Decker plots [2, 3]. The idea is to partition a rectangle on the y-axis according to one attribute and make the size of the regions proportional to the sum of the corresponding data values.

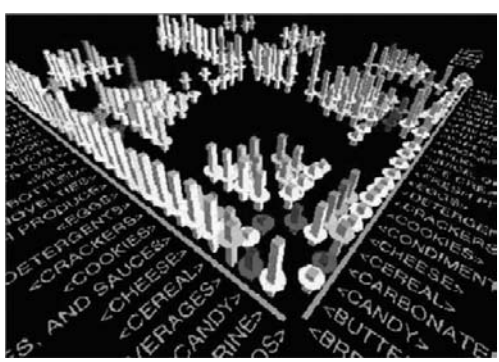

Fig. 1. MineSet's association rule visualizer [2].

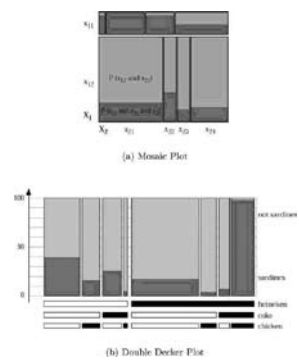

Fig. 2. Association rule visualization [2].

Mosaic plots use the height of the bars instead of their width to show the parameter value. Then each resulting area is split according to a second attribute. The coloring reflects the percentage of data items that fulfill a third attribute. The visualization shows the support and confidence values of all rules of the form X_1, X_2 => Y. Mosaic plots are restricted to two attributes on the left side of the association rule. Double Decker plots can be used to show more than two attributes on the left side. Idea is to display a hierarchy of attributes on the bottom corresponding to the left side of the association rules. The bars correspond to the number of items in the considered database subset

and therefore visualize the support of the rule (the colored areas in the bars correspond to the percentage of data transactions that contain an additional item and therefore represents the support).

Other approaches to association rule visualization include graphs with nodes corresponding to items and arrows corresponding to implications and association matrix visualizations to cluster related rules.

2 Classification

Classification is the process of developing a classification model based on a training data set with known class labels. The class descriptions are used to classify data for which the class labels are unknown. Classification is sometimes called *supervised learning*. A popular approach is the algorithm that inductively constructs decision trees. Some approaches use neural networks, genetic algorithms, or Bayesian networks to solve the classification problem [1]. Usually problem is seen as a black box and for this reason some problems such as over fitting or tree pruning are difficult to tackle. It is why we use visualization techniques to overcome these problems (SGIs MineSet tree visualizer – Figure 3).

The system allows an interactive selection of the attributes and helps the user to understand the decision tree. A more sophisticated approach, which helps in decision tree construction, is visual classification. It shows each attribute value by a colored pixel and arranges them in bars - similar to the Dense Pixel Displays. The attribute bars for each pixel are sorted separately and the attribute with the purest value distribution is selected as the split attribute of the decision tree. Until all leaves correspond to pure classes, the procedure is repeated. The decision tree process is shown in Figure 4. If we compare a standard visualization of a decision tree, we can see that we have additional information that is helpful for explanation and analysis of the decision tree process. They are the node size, purity of the resulting partitions and class distribution.

Some standard visualization techniques of decision trees can provide this information, but this approach clearly fails for more complex information such as the class distribution. In general, visualizations provide a better understanding of the classification models and they can help to interact more easily with the classification algorithms to optimize the model generation and classification process.

3 Clustering

Clustering is the process of partitioning the data set into homogeneous subsets called clusters [1]. Unlike classification, clustering is implemented as a form of unsupervised learning. Many clustering algorithms are density-based and linkage-based methods [3]. Most algorithms use assumptions about the properties of the clusters that are either used as defaults or have to be given as input parameters. But, depending on the parameter values, the user obtains different clustering results.

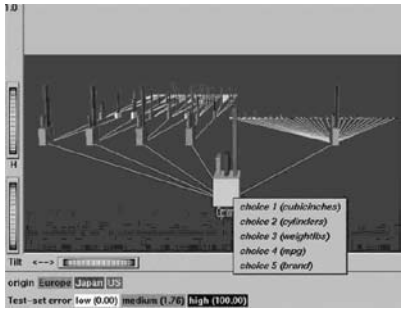

Fig. 3. MineSet's Decision tree Visualizer.

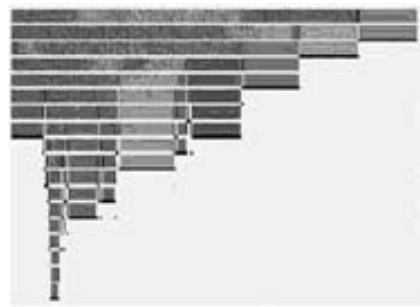

Fig. 4. Visualization of Decision tree for training data set with 19 attributes [2].

The impact of different algorithms and parameters settings in two or three dimensional space can be explored easily using simple visualizations of the resulting clusters (for example, x-y plots). In higher dimensional space, the impact is much more difficult to understand and for these reasons, some higher-dimensional techniques try to determine two or three dimensional projections of the data that retain the properties of the high-dimensional clusters as much as possible. Figure 5 shows a three-dimensional projection of a data set consisting of five clusters. This approach works well with small to medium dimensional data sets. But, it is difficult to apply this approach to large high-dimensional data sets, especially if clusters are not clearly separated or data sets contain noise.

In this case, more sophisticated visualization techniques are required to guide the clustering process and select the right clustering model and adjust the parameter values. An example for a system that uses visualization techniques to help in high-dimensional clustering is OPTICS (Ordering Points to Identify the Clustering Structure) [2]. It creates an one-dimensional (or two-dimensional) ordering of the database representing its density-based clustering structure (Figure 6). Intuitively, points within a cluster are closer in the generated one-dimensional ordering and their reachability distance (Figure 6) is similar. Points from another cluster have higher reachability distances. This technique is valuable for understanding the clustering process.

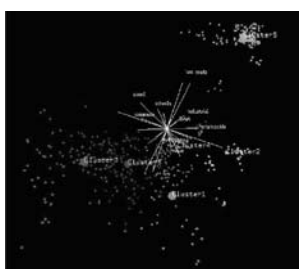

Fig. 5. Visualization based on a projection into 3D space [2].

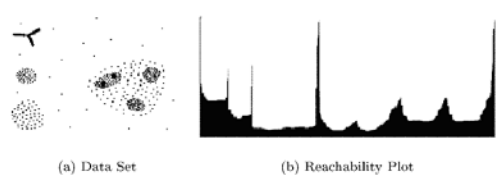

Fig. 6. OPTIC Visual clustering [2].

Other interesting approach is developed in the HD-Eye system [2, 3]. The HD-Eye system considers the clustering problem as a partitioning problem and supports a tight integration of advanced clustering algorithms and state-of-the-art visualization techniques. These techniques allow a user direct interaction in

the crucial steps of clustering process. Steps include: selection of dimensions to be considered, the selection of the clustering paradigm, and the partitioning of the data set. They provide the best separators for partitioning the data [2].

Conclusions

The exploration of large data sets is an important but difficult problem. Information visualization techniques and visual data exploration has a high potential to solve these problems. There is a tight integration of visualization techniques with traditional techniques from disciplines such as statistics, machine learning, operations research, and simulation. This integration would combine fast automatic data analysis algorithms with the intuitive power of the human mind, improving the quality and speed of the data analysis process. Also, there is a tight integration with the data managing systems, including database management and data warehouse systems. The goal is to bring the power of visualization technology to all of us for better, faster, and more intuitive exploration of very large data resources. The visualizations are also used to decide which dimensions are taken for the partitioning and users can create partitions interactively, directly within the visualization. Data mining processes where we use visualization tools are rule generation, classification and clustering.

References

1. Cao J., Yu P.S., Zhang C.,Zhang H., Data Mining for Business Application, Springer, 2007
2. Berthold M., Hand D.J., editors, Intelligent Data Analysis, Second edition, Springer, 2007, ACM Computing Classification (1998), Pages 423-428;
3. Krzanowski W.J.. Principles of Multivariate Analysis: A User's Perspective. Number 3 in Oxford Statistical Science Series. Oxford University Press, Oxford, 1988.
4. Brunk C., Kelly J., Kohavi R., MineSet: An Integrated System for Data Mining, KDD-97 Proceedings. Copyright © 1997, AAAI (www.aaai.org)
5. Ao.S.I., Reiger B., Chen S.S., Advances in Computational Algorithms and Data Analysis, Springer 2009