Specific Usage of Visual Data Analysis Techniques

Snezana Savoska[1], Suzana Loskoska[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. There are many techniques that can be used to support these tasks of data mining processes. The paper contains some specific usage of data visualization technique and tools used for generation of association rules, classification and clustering.

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

1 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 some special algorithms for data analysis and data mining. The data are often automatically recorded via sensors and monitoring systems, via 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. In the data analysis process, the user is directly involved and this is a great advantage of visual data exploration.

Visual data mining integrate the human in the data analysis process. The human perceptual abilities help to the analysis of today's large data sets [1]. Visual data mining process presents the data in some visual form, providing the user to gain insight into the data, draw conclusions, and directly interact with the data. 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 some new hypotheses. The verification of the hypotheses can also be done via data visualization, and may be accomplished by automatic techniques from statistics, pattern recognition, or machine learning. But, 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 it 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 provides more interesting results, especially in cases where automatic algorithms fail. VDE techniques provide a much higher degree of confidence in the findings of the exploration, too. These facts lead to a high demand for visual exploration techniques and make them indispensable in conjunction with automatic exploration techniques.

The visualization techniques are commonly applied in many arreas especialy for supporting data decision proces. There are a number of vizualization techniques that have been developed for some specific usage as data mining tasks. These data mining processes include association rule generation, classification and clastering. There are many techniques that can be used to support these task of data mining processes.

1 Generation of Association Rules

Association rules are statistical relation 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 pattern and trends in databases, but it is important to find out some rule in 70% of cases and define transaction with some probability, called confidence. A second important parameter is the support of this rule, defined as percentage of co-occurence of items in the transactions. An association rule is an implication of the form $X => Y$, 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 overcome this problem and allow an interactive selection of good support and confidence levels. In the Figure 1 is shown SGI Sets Rule Visualizer [6, 2] which maps the left and right hand 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. Using the visualization, the user is able to see groups of related rules and the impact of different confidence and support levels. But we can visualize limited number of rules and the visualization does not support combinations of items on the left or right hand side of the association rules. 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.

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 hand side of the association rules. The bars correspond to the number of items in the considered subset of the database 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 represent 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.

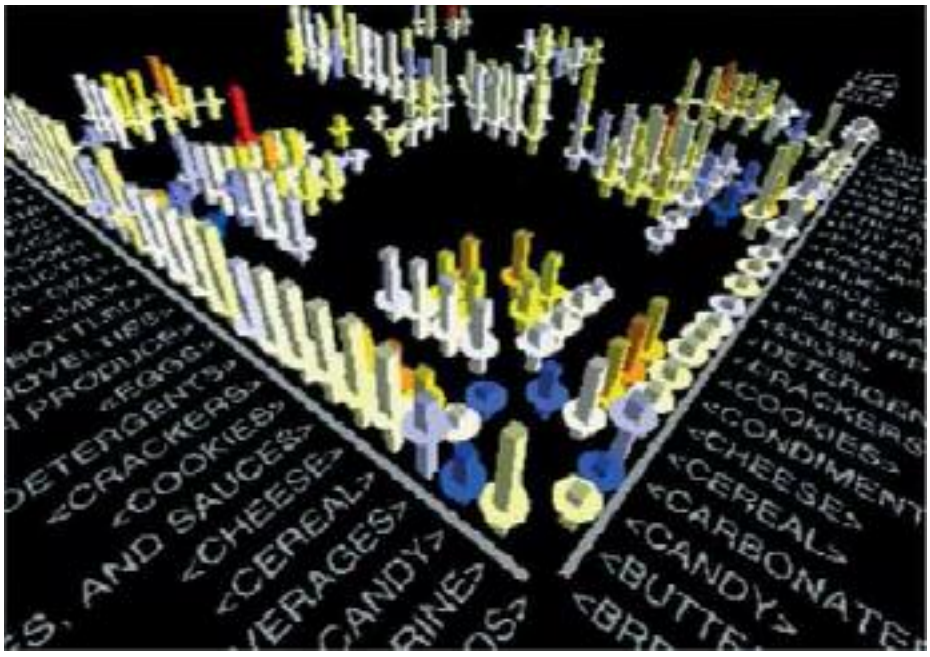
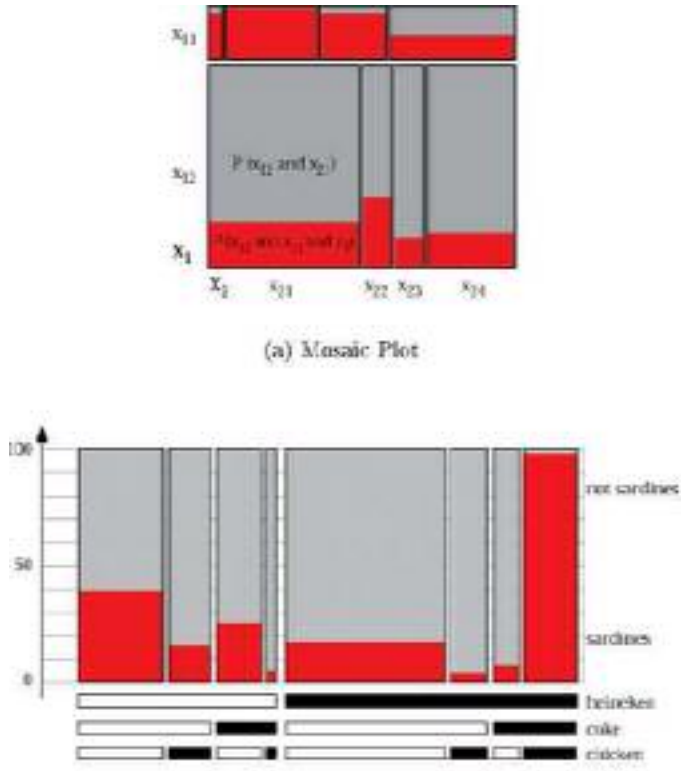


Fig. 1. MineSet's association rule visualizer [2]

2 Classification

Classification is the process of developing a classification model based on a training data set with known class labels. If we want to construct the classification model, the attributes of the training data set must be analyzed. Also an accurate description or model of the classes based on the attributes available in the data set must be developed. The class descriptions are used to classify data for which the class labels are unknown. Classification is sometimes also called *supervised learning* because the training set is used to teach the system how to classify the data. A popular approach is algorithms that inductively construct decision trees. Some approaches use neural networks, genetic algorithms, or Bayesian networks for solving 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).

In this case, the system allows an interactive selection of the attributes and helps the user to understand the decision tree. Also, a more sophisticated approach, which also 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 each attribute bar pixels 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 resulting from this 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 explaining and analyzing the decision tree process. They are the node size (number of training records corresponding to the node), purity of the resulting partitions and class distribution (frequency and location of the training instances of all classes).



(a) Mosaic Plot

(b) Double Decker Plot

Fig. 2. Association rule Visualization [2]

Some standard visualization techniques of a decision tree 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.

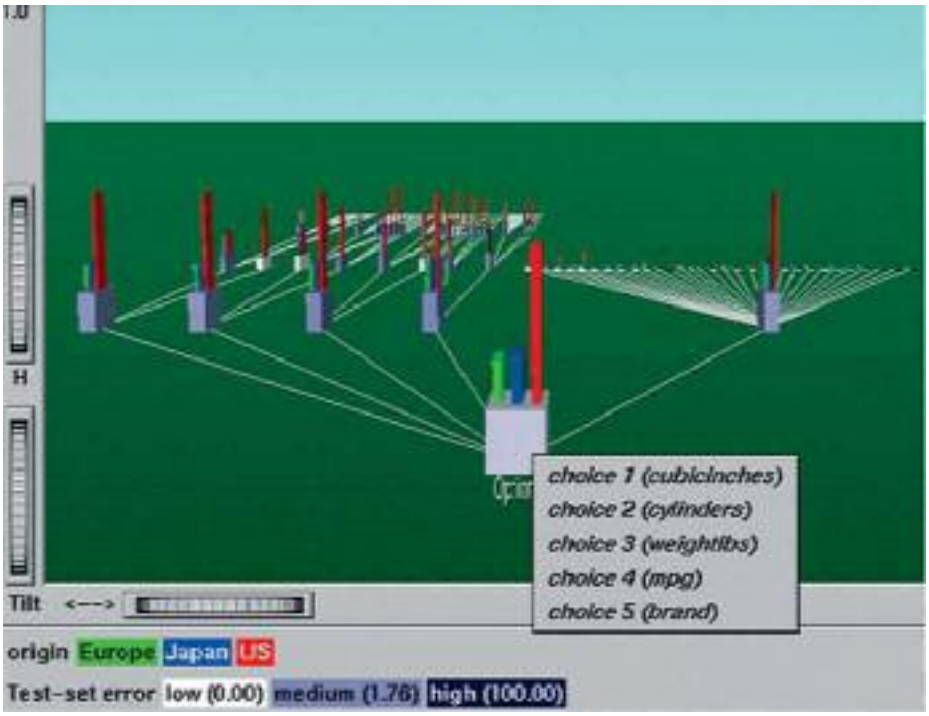


Fig. 3. MineSet's Decision tree Visualizer

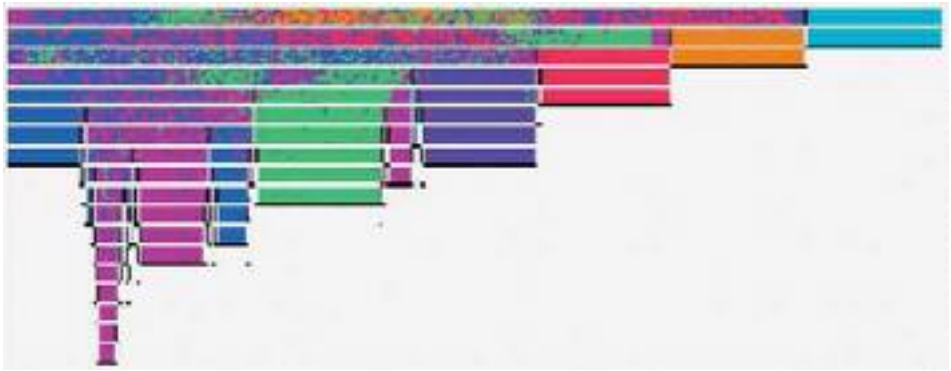


Fig. 4. Visualization of Decision tree for training data set with 19 attributes [2]

3 Clustering

Clustering is the process of partitioning of the data set into homogeneous subsets called clusters [1]. Unlike classification, clustering is implemented as a form of unsupervised learning. The classes are unknown and the training set with class labels is available. 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.

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 the clusters are not clearly separated or data set contains noise.

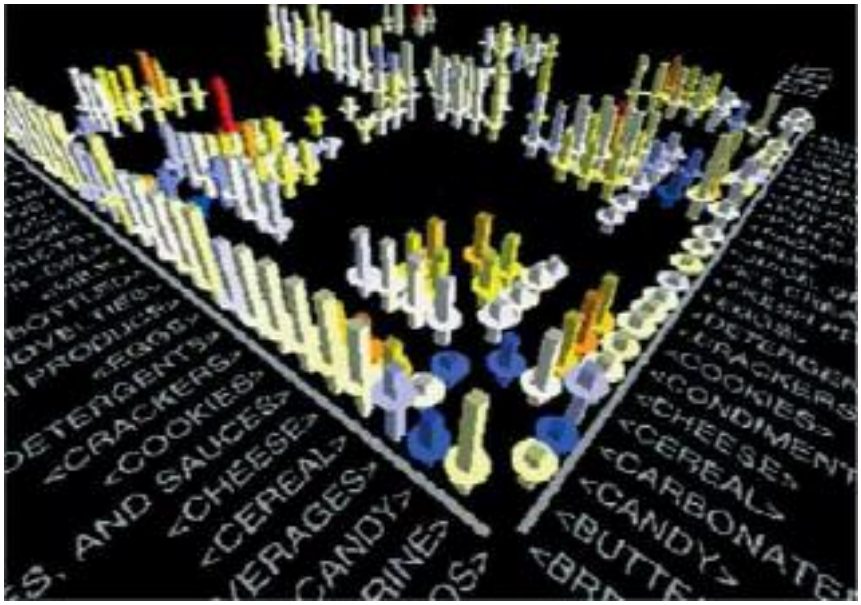


Fig. 5. Visualization based on a projection into 3D space [2]

In this case, more sophisticated visualization techniques are required to guide the clustering process and select the right clustering model and also adjust the parameter values appropriately. We can see an example of a system that uses visualization techniques to help in high-dimensional clustering. This is OPTICS (Ordering Points to Identify the Clustering Structure) [2]. In this technique, the basic idea is to create a one-dimensional (or two-dimensional) ordering of the database representing its

density-based clustering structure (Figure 6). Intuitively, points within a cluster are close in the generated one-dimensional ordering and their reachability distance (Figure 6) is similar. If we jump to another cluster, results in higher reachability distances which provide a visualization of the inherent clustering structure. It is therefore valuable for understanding the clustering process.

Other interesting approach is 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 the user direct interaction in the crucial steps of clustering process. They are: 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 (Figure 7).

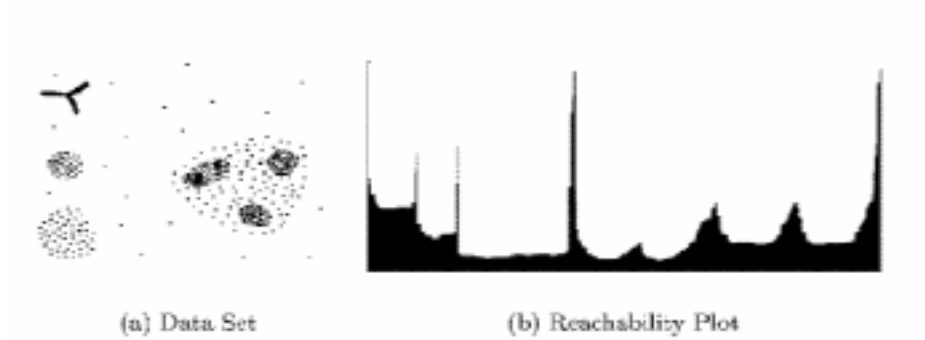


Fig. 6. OPTIC Visual clustering [2]

The separator tree represents the clustering model produced in the clustering process. The abstract iconic displays in Figure 8 visualize the partitioning potential of a large number of projections. The properties are based on histogram information of the point density in the projected space and the number of icons corresponds to the number of peaks in the projection. Their colors correspond to the number of data points belonging to the maximum from dark colors for large maxima to bright colors for small maxima. The measure of how well a maximum is separated from the others is reflected by the shape of the icon [5, 2]. The degree of separation varies from sharp spikes for well-separated maxima to blunt spikes for weak-separated maxima.

In this case, the visualizations are used to decide which dimensions are taken for the partitioning. The partitioning can be specified interactively directly within the visualizations, allowing the user to define non-linear partitioning.

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. These techniques have many applications such as fraud detection. We have tight integration of visualization techniques with traditional techniques from such disciplines as statistics, machine learning, operations research, and simulation. This

integration of visualization techniques and these more established methods 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 tightly integration of the data managing systems that manage the vast amount of relational and semi structured information, including database management and data warehouse systems. This time, 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. It is valuable in an economic sense but also stimulates and delights the user. 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.

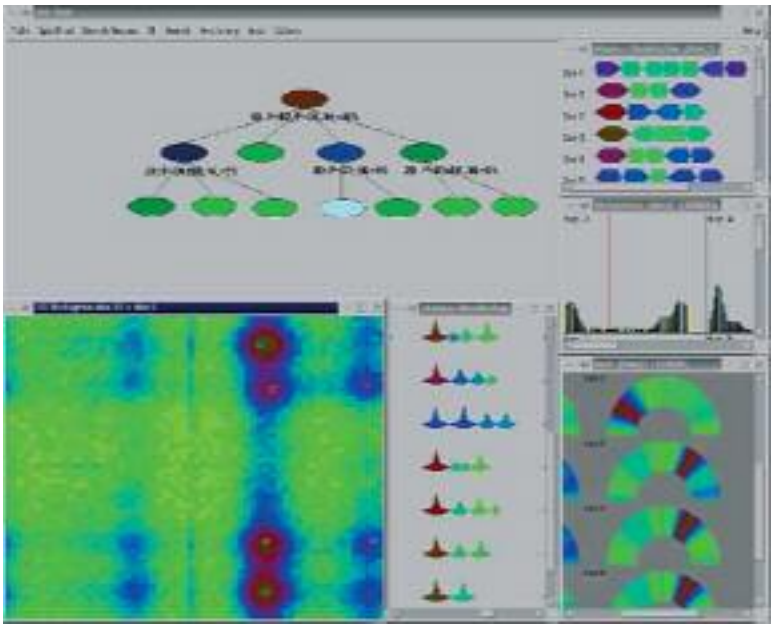


Fig. 7. HD-Eye a screen-shot [2] shows different visualizations of projections and the separator tree with different representation of multi-dimensional projections and color-based 2D density plot

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): 1.2, H.3, G.3,1.5.1,1.4, J.2, J.l, J.3, F.4.1, F.l, 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
6. http://www.the-data-mine.com/bin/view/Software/MineSethttp