Visualization techniques for data mining

Viktor, Herna L.; Paquet, Eric

Encyclopedia of Data Warehousing and Mining, pp. 1190-1195, 2005

Visualization Techniques for Data Mining

Herna L Viktor[1] and Eric Paquet[2]

[1]School of Information Technology and Engineering, University of Ottawa, Ontario, Canada

E-mail: hlviktor@site.uottawa.ca

[2] Visual Information Technology, National Research Council, Ottawa, Ontario, Canada

E-mail: Eric.Paquet@nrc-cnrc.gc.ca

INTRODUCTION

The current explosion of data and information, mainly caused by data warehousing technologies as well as the extensive use of the Internet and its related technologies, has increased the urgent need for the development of techniques for intelligent data analysis. Data mining, which concerns the discovery and extraction of knowledge chunks from large data repositories, is aimed at addressing this need. Data mining automates the discovery of hidden patterns and relationships that may not always be obvious. Data mining tools include classification techniques (such as decision trees, rule induction programs and neural networks) (Han and Kamber, 2001); clustering algorithms and association rule approaches, amongst others.

Data mining has been fruitfully used in many of domains, including marketing, medicine, finance, engineering and bioinformatics. There still are, however, a number of factors that militate against the widespread adoption and use of this new technology. This is mainly due to the fact that the results of many data mining techniques are often difficult to understand. For example, the results of

a data mining effort producing 300 pages of rules will be difficult to analyze. The visual representation of the knowledge embedded in such rules will help to heighten the comprehensibility of the results. The visualization of the data itself, as well as the data mining process should go a long way towards increasing the user's understanding of and faith in the data mining process. That is, data and information visualization provide users with the ability to obtain new insights into the knowledge, as discovered from large repositories.

This paper describes a number of important visual data mining issues and introduces techniques employed to improve the understandability of the results of data mining. Firstly, the visualization of data prior to and during data mining is addresses. Through *data* visualization, the quality of the data can be assessed throughout the knowledge discovery process, which includes data preprocessing, data mining and reporting. We also discuss *information* visualization, i.e. how the knowledge, as discovered by a data mining tool, may be visualized throughout the data mining process. This aspect includes visualization of the results of data mining as well as the learning process. In addition, the paper shows how virtual reality and collaborative virtual environments may be used to obtain an immersive perspective of the data and the data mining process.

BACKGROUND

Human beings intuitively search for novel features, patterns, trends, outliers and relationships in data (Han and Kamber, 2001). Through visualizing the data and the concept descriptions obtained (e.g. in the form of rules), a qualitative overview of large and complex data sets can be obtained. In addition, data and rule visualization can assist in identifying regions of interest and appropriate parameters for more focused quantitative analysis (Grinstein and Ward, 2001). The user can thus

get a "rough feeling" of the quality of the data, in terms of its correctness, adequacy, completeness, relevance, etc. The use of data and rule visualization thus greatly expands the range of models that can be understood by the user, thereby easing the so-called "accuracy versus understandability" tradeoff (Thearling et al, 2001).

Data mining techniques construct a model of the data through repetitive calculation to find statistically significant relationships within the data. However, the human visual perception system can detect patterns within the data that are unknown to a data mining tool. This combination of the various strengths of the human visual system and data mining tools may subsequently lead to the discovery of novel insights and the improvement of the human's perspective of the problem at hand. Visual data mining harnesses the power of the human vision system, making it an effective tool to comprehend data distribution, patterns, clusters and outliers in data (Han and Kamber, 2001).

Visual data mining is currently an active area of research. Examples of related commercial data mining packages include the *DBMiner* data mining system, *See5* which forms part of the RuleQuest suite of data mining tools, *Clementine* developed by Integral Solutions Ltd (ISL), *Enterprise Miner* developed by SAS Institute, *Intelligent Miner* produced by IBM, and various other tools (Han and Kamber, 2001). Neural network tools such as *NeuroSolutions* and *SNNS* and Bayesian network tools including *Hugin*, *TETRAD*, and *Bayesware Discoverer*, also incorporates extensive visualization facilities. Examples of related research projects and visualization approaches include *MLC++*, *WEKA*, *AlgorithmMatrix* and *C4.5/See5*, amongst others (Han and Kamber, 2001, Fayyad et al, 2001).

Visual data mining integrates data visualization and data mining and is closely related to computer graphics, multimedia systems, human computer interfaces, pattern recognition and high performance computing.

MAIN TRUST OF THE PAPER

Data and information visualization will be further explored in terms of their benefits for data mining.

Data Visualization

Data visualization provides a powerful mechanism to aid the user during both data preprocessing and the actual data mining (Foong, 2001). Through the visualization of the original data, the user can browse to get a "feel" for the properties of that data. For example, large samples can be visualized and analyzed (Grinstein and Ward, 2001). In particular, visualization may be used for outlier detection, which highlights surprises in the data, i.e. data instances that do not comply with the general behavior or model of the data (Han and Kamber, 2001, Pyle, 1999). In addition, the user is aided in selecting the appropriate data through a visual interface. Data transformation is an important data preprocessing step. During data transformation, visualizing the data can help the user to ensure the correctness of the transformation. That is, the user may determine whether the two views (original versus transformed) of the data are equivalent. Visualization may also be used to assist users when integrating data sources, assisting them to see relationships within the different formats.

Data visualization techniques are classified in respect of three aspects (Grinstein and Ward, 2001). Firstly, their focus, i.e. symbolic versus geometric; secondly their stimulus (2D versus 3D); and lastly, their display (static or dynamic) (Fayyad et al, 2001). In addition, data in a data repository can be viewed as different levels of granularity or abstraction, or as different combinations of attributes or dimensions. The data can be presented in various visual formats, including box plots, scatter plots, 3D-cubes, data distribution charts, curves, volume visualization, surfaces or link graphs, amongst others (Grinstein and Ward, 2001).

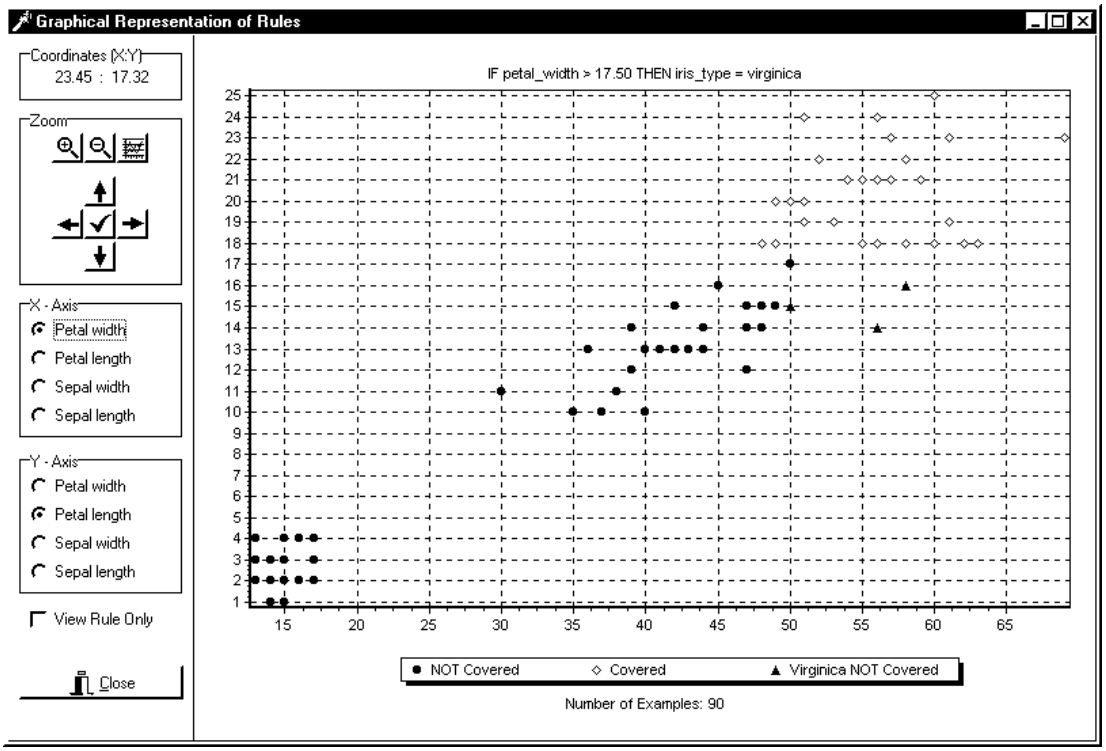

Figure 1: The Vizimine data and information visualization tool (Viktor et al, 2003)

Scatter plots refer to the visualization of data items according to two axes, namely X and Y values. According to Hoffman and Grinstein (2001), the scatter plot is the most popular visualization tool,

since it can help find clusters, outliers, trends and correlations. Figure 1 shows an example of a scatter plot used in the ViziMine system, which visualizes both the data set and the subset of the data covered by a particular rule as discovered by a data mining tool (Viktor et al, 2003). The figure shows the scatter plot for the rule *IF (petal-width > 17.50) then Iris-type = Virginica*. This data set involves the classification of Irises into one of three types. The figure shows the Virginica data instances covered by the rule, the data instances from the other two types of Irises not covered by the rule. In addition, it also importantly indicates those Virginicas, which were not covered by the rule.

Other data visualization techniques include 3D-cubes are used in relationship diagrams, where the data are compared as totals of different categories. In surface charts, the data points are visualized by drawing a line between them. The area defined by the line, together with the lower portion of the chart, is subsequently filled. Link or line graphs display the relationships between data points through fitting a connecting line (Paquet et al, 2000). They are normally used for 2D data where the X value is not repeated (Hoffman and Grinstein, 2001).

Advanced visualization techniques may greatly expand the range of models that can be understood by domain experts, thereby easing the so-called accuracy-versus-understandability trade-off (Singhal et al, 1999). However, due to the so-called "curse of dimensionality", which refers to the problems associated with working with numerous dimensions, highly accurate models are usually less understandable, and vice versa. In a data mining system, the aim of data visualization is to obtain an initial understanding of the data and the quality thereof. The actual accurate assessment of the data and the discovery of new knowledge are the tasks of the data mining tools. Therefore, the visual display should preferably be highly understandable, possibly at the cost of accuracy.

The use of one or more of the above-mentioned data visualization techniques thus helps the user to obtain an initial model of the data, in order to detect possible outliers and to obtain an intuitive assessment of the quality of the data used for data mining. The visualization of the data mining process and results is discussed next.

Information Visualization

According to Foster and Gee (2001), it is crucial to be aware of what users require for exploring data sets, small and large. The driving force behind visualizing data mining models can be broken down into two key areas, namely understanding and trust (Singhal et al, 1999, Thearling et al, 2001). Understanding means more than just comprehension; it also involves context. If the user can understand what has been discovered in the context of the business issue, he will trust the data and the underlying model and thus put it to use. Visualizing a model also allows a user to discuss and explain the logic behind the model to others. In this way, the overall trust in the model increases and subsequent actions taken as a result are justifiable (Thearling et al, 2001).

The art of information visualization can be seen as the combination of three well defined and understood disciplines, namely cognitive science, graphics art and information graphics. A number of important factors have to be kept in mind when visualizing both the execution of the data mining algorithm (process visualization), e.g. the construction of a decision tree, and displaying the results thereof (result visualization). The visualization approach should provide an easy understanding of the domain knowledge, explore visual parameters and produce useful outputs. Salient features should be encoded graphically and the interactive process should prove useful to the user.

The format of knowledge extracted during the mining process depends on the type of data mining task and its complexity. Examples include classification rules, association rules, temporal sequences and casual graphs (Singhal, 1999). Visualization of these data mining results involves the presentation of the results or knowledge obtained from data mining in visual forms, such as decision trees, association rules, clusters, outliers and generalized rules. For example, the Silicon Graphics (SGI) MineSet 3.0 toolset uses connectivity diagrams to visualize decision trees, and simple Bayesian and decision table classifiers (Han and Kamber, 2001, Thearling et al, 2001). Other examples include the Evidence Visualizer, which is used to visualize Bayesian classifiers (Becker et al, 2001); the DB-Discover system that uses multi-attribute generalization to summarize data (Hilderman, 2001); and the NASD Regulation Advanced Detection System, which employs decision trees and association rule visualization for surveillance of the NASDAQ stock market (Senator et al, 2001).

Alternatively, visualization of the constructs created by a data mining tool (e.g. rules, decision tree branches, etc.) and the data covered by them may be accomplished through the use of scatter plots and box plots. For example, scatter plots may be used to indicate the points of data covered by a rule in one color and the points *not* covered by another color. The ViziMine tool uses this method, as depicted in Figure 1 (Viktor et al, 2003). This visualization method allows users to ask simple, intuitive questions interactively (Thearling et al, 2001). That is, the user is able to complete some form of "what if" analysis. For example, consider a rule *IF (petal-width > 17.50) then Iris-type = Virginica* from the Iris data repository. The user is subsequently able to see the effect on the data point covered when the rule's conditions are changed slightly, e.g. to *IF (petal-width > 16.50) then Iris-type = Virginica.*

Future trends

Three-dimensional visualization has the potential to show far more information than two-dimensional visualization, while retaining its simplicity. This visualization technique quickly reveals the quantity and relative strength of relationships between elements, helping to focus attention on important data entities and rules. It therefore aids both the data preprocessing and data mining processes.

In two dimensions, data representation is limited to bidimensional graphical elements. In three dimensions both two and three-dimensional graphical elements can be utilized. These elements are much more numerous and diversified in three dimensions than in two. Furthermore, three-dimensional representations (or descriptors) can be either volumetric or surface-based depending on whether the internal structure is of interest or not. A surface-based representation only takes into account the outer appearance or the shell of the object while a volumetric approach assigns a value to each volume element. The latter approach is quite common in biomedical imagery such as CAT scanning.

Many techniques are available to visualize data in three dimensions (Harris, 2000). For example, it is very common to represent data by glyphs (Hoffman and Grinstein, 2001, Fayyad et al, 2001). A glyph can be defined as a three-dimensional object suitable for representing data or subsets of data. The object is chosen in order to facilitate both the visualization and the data mining process. The glyph must be self-explanatory and unambiguous. Glyphs can have various attributes such as their color and scale. When using these attributes to describe a glyph, a so-called content-based

descriptor is constructed. Even if most glyphs are rigid objects, non-rigid and articulated objects can be used as well. It is then possible to use the deformation and the pose of the glyph in order to represent some specific behavior of the data set. Furthermore, glyphs can be animated in order to model some dynamic process.

Three-dimensional visualization can be made more efficient by the use of virtual reality (VR). A virtual environment (VE) is a three-dimensional environment characterized by the fact that it is immersive, interactive, illustrative and intuitive. The fact that the environment is immersive is of great importance in data mining. In traditional visualization, the human subject looks at the data from outside, while in a VR environment the user is part of the data world. This means that the user can utilize all his senses in order to navigate and understand the data. This also implies that the representation is more intuitive. VR is particularly well adapted to representing the scale and the topology of various sets of data. That becomes even more evident when stereo visualization is utilized, since stereo vision allows the analyst to have a real depth perception. This depth perception is important in order to estimate the relative distances and scales between the glyphs. Such estimation can be difficult without stereo vision if the scene does not correspond to the paradigms our brain is used to processing. In certain cases, the depth perception can be enhanced by the use of metaphors.

Collaborative Virtual Environments (CVEs) can be considered as a major breakthrough in data mining (Singhal et al, 1999). By analogy, they can be considered as the equivalent of collaborative agents in visualization. Traditionally, one or more analysts perform visualization at a unique site. This operational model does not reflect the fact that many enterprises are distributed worldwide and so are their operations, data and specialists. It is consequently impossible for those enterprises to

centralize all their data mining operations in a single center. Not only must they collaborate on the data mining process, which can be carried out automatically to a certain extent by distributed and collaborative agents, but they must also collaborate on the visualization and the visual data mining aspects.

CONCLUSION

The ability to visualize the results of a data mining effort aids the user to understand and trust the knowledge embedded in it. Data and information visualization provide the user with the ability to get an intuitive "feel" for the data and the results, e.g. in the form of rules, that is being created. This ability can be fruitfully used in many business areas, for example for fraud detection, diagnosis in medical domains and credit screening, amongst others.

Virtual reality and collaborative virtual environments are opening up challenging new avenues for data mining. VR is perfectly adapted to analyze alphanumerical data and to map them to a virtually infinite number of representations. Collaborative virtual environments provide a framework for collaborative and distributed data mining by making an immersive and synergic analysis of data and related patterns possible. In addition, there is a wealth of multimedia information waiting to be data mined. With the recent advent of a wide variety of content-based descriptors and the MPEG-7 standard to handle them, the fundamental framework is now in place to undertake this task (MPEG-7, 2004). The use of virtual reality to effectively manipulate and visualize both the multimedia data and descriptors opens up exciting new research avenues.

REFERENCES

Becker, B., Kohavi, R., & Sommerfield, D. (2001). Visualizing the Simple Bayesian Classifier, In Fayyad, U., Grinstein, G.G., & Wiese, A. (eds.), *Information Visualization in Data Mining and Knowledge Discovery*, pp.237-250, San Francisco: Morgan Kaufmann.

Fayyad, U., Grinstein, G.G. & Wierse, A. (2001). *Information Visualization in Data Mining and Knowledge Discovery*, San Francisco: Morgan Kaufmann.

Foong, D.L.W. (2001). A Visualization-Driven Approach to Strategic Knowledge Discovery, In Fayyad, U., Grinstein, G.G., & Wiese, A. (eds.), *Information Visualization in Data Mining and Knowledge Discovery*, pp.181-190, San Francisco: Morgan Kaufmann.

Foster, M., & Gee, A.G. (2001). The Data Visualization Environment, In Fayyad, U., Grinstein, G.G., & Wiese, A. (eds.), *Information Visualization in Data Mining and Knowledge Discovery*, pp.83-94, San Francisco: Morgan Kaufmann.

Grinstein, G.G., & Ward, M.O. (2001). Introduction to data visualization. In Fayyad, U., Grinstein, G.G., & Wiese, A. (eds.), *Information Visualization in Data Mining and Knowledge Discovery*, pp.21-26, San Francisco: Morgan Kaufmann.

Han, J., & Kamber, M. (2001). *Data Mining Concepts and Techniques*, San Francisco: Morgan Kaufmann.

Harris, R.L. (2000). *Information Graphics: A Comprehensive Illustrated Reference*, Oxford: Oxford University Press.

Hilderman, R.J., Li, L., & Hamilton, H.J. (2001). Visualizing Data Mining Results with Domain Generalization Graphs, In Fayyad, U., Grinstein, G.G., & Wiese, A. (eds.), *Information Visualization in Data Mining and Knowledge Discovery*, pp.251-269, San Francisco: Morgan Kaufmann.

Hoffman, P.E., & Grinstein, G.G. (2001). A Survey of Visualization for High-Dimensional Data Mining, In Fayyad, U., Grinstein, G.G., & Wiese, A. (eds.), *Information Visualization in Data Mining and Knowledge Discovery*, pp.47-82, San Francisco: Morgan Kaufmann.

MPEG-4 and MPEG-7 <http://mpeg.telecomitalialab.com>

Paquet, E., Robinette, K.M., & Rioux, M. (2000). Management of Three-dimensional and Anthropometric Databases: Alexandria and Cleopatra, *Journal of Electronic Imaging,* 9, 421-431.

Pyle, D. (1999). *Data Preparation for Data Mining,* San Francisco: Morgan Kaufman.

Senator, T.E., Goldberg, H.G., & Shyr, P. (2001). The NASD Regulation Advanced Detection System, In Fayyad, U., Grinstein, G.G., & Wiese, A. (eds.), *Information Visualization in Data Mining and Knowledge Discovery*, pp.363-371, San Francisco: Morgan Kaufmann.

Singhal, S., et al. (1999). *Networked Virtual Environments: Design and Implementation*, Reading, MA: Addison-Wesley.

Thearling, K., et al (2001). Visualizing Data Mining Models, In Fayyad, U., Grinstein, G.G., & Wiese, A. (eds.), *Information Visualization in Data Mining and Knowledge Discovery*, pp.205-222, San Francisco: Morgan Kaufmann.

Viktor, H.L., Paquet, E., & Le Roux, J.G., (2003), Cooperative Learning and Virtual Reality-based Visualization for Data Mining", In Wang, J., (ed.), *Data mining: Opportunities and Challenges*, Chapter 3, pp.55-79, Hershey, PA: IRM Publishers.

TERMS AND DEFINITIONS

Collaborative virtual environment. An environment that actively supports human-human communication in addition to human-machine communication and which uses a virtual environment as the user interface.

Curse of dimensionality. The problems associated with information overload, when the number of dimensions is too high to visualize.

Data visualization. The visualization of the data set through the use of a techniques such as scatter plots, 3D cubes, link graphs and surface charts.

Dimensionality reduction. The removal of irrelevant, weakly relevant, or redundant attributes or dimensions through the use of techniques such as principle component analysis or sensitivity analysis.

Information visualization. The visualization of data mining models, focusing on the results of data mining and the data mining process itself. Techniques include rule-based scatter plots, connectivity diagrams, multi-attribute generalization and decision tree and association rule visualization.

Multimedia data mining. The application of data mining to data sets consisting of multimedia data, such as 2D images, 3D objects, video and audio. Multimedia data can be viewed as integral data records, which consist of relational data together with diverse multimedia content.

Virtual reality. Immersive, interactive, illustrative and intuitive representation of the real world based on visualization and computer graphic.

Visualization. The graphical expression of data or information.

Visual data mining. The integration of data visualization and data mining. Visual data mining is closely related to computer graphics, multimedia systems, human computer interfaces, pattern recognition and high performance computing.