Visual Data Mining from Visualization to Visual Information Mining

Herna L. Viktor
University of Ottawa, Canada

Eric Paquet
National Research Council, Canada

INTRODUCTION

The current explosion of data and information, which are mainly caused by the continuous adoption of data warehouses and 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, addresses 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) (Kou et al., 2007); 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 addressed. 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 as well as how visual data mining can be used to directly mine functionality with specific applications in the emerging field of proteomics.

BACKGROUND

Human beings intuitively search for novel features, patterns, trends, outliers and relationships in data (Han and Kamber, 2006). 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. 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 (Valdes and Barton, 2007).

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 (Blanchard et al., 2007).

Visual data mining is currently an active area of research. Examples of related commercial data mining packages include the *MultiMediaMiner* 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. 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*, *C4.5/See5* and NCBI GEO amongst others (Barret et al., 2007).

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.

DATA AND INFORMATION VISUALIZATION

Data Visualization

Data visualization provides a powerful mechanism to aid the user during both data preprocessing and the actual data mining. 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 (Barret et al., 2007). 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 (Sun et al., 2007). 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. Firstly, their focus, i.e. symbolic versus geometric; secondly their stimulus (2D versus 3D); and lastly, their display (static or dynamic). 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 (Gardia-Osorio and Fyfe, 2008).

For instance, 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 (Guo et al., 2007). They are normally used for 2D data where the X value is not repeated.

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. 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.