Interactive Visual Data Mining

Shouhong Wang
University of Massachusetts Dartmouth, USA

Hai Wang
Saint Mary's University, Canada

INTRODUCTION

In the data mining field, people have no doubt that high level information (or knowledge) can be extracted from the database through the use of algorithms. However, a one-shot knowledge deduction is based on the assumption that the model developer knows the structure of knowledge to be deducted. This assumption may not be invalid in general. Hence, a general proposition for data mining is that, without human-computer interaction, any knowledge discovery algorithm (or program) will fail to meet the needs from a data miner who has a novel goal (Wang, S. & Wang, H., 2002). Recently, interactive visual data mining techniques have opened new avenues in the data mining field (Chen, Zhu, & Chen, 2001; de Oliveira & Levkowitz, 2003; Han, Hu & Cercone, 2003; Shneiderman, 2002; Yang, 2003).

Interactive visual data mining differs from traditional data mining, standalone knowledge deduction algorithms, and one-way data visualization in many ways. Briefly, interactive visual data mining is human centered, and is implemented through knowledge discovery loops coupled with human-computer interaction and visual representations. Interactive visual data mining attempts to extract unsuspected and potentially useful patterns from the data for the data miners with novel goals, rather than to use the data to derive certain information based on *a priori* human knowledge structure.

BACKGROUND

A single generic knowledge deduction algorithm is insufficient to handle a variety of goals of data mining since a goal of data mining is often related to its specific problem domain. In fact, knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns of data mining (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). By this definition, two aspects of knowledge discovery are important to meaningful data mining. First, the criteria of validity, novelty, usefulness of knowledge to be discovered could be subjective. That is, the usefulness of a data pattern depends on the data miner and does not solely depend on the statistical strength of the pattern. Second, heuristic search in combinatorial spaces built on computer and human interaction is useful for effective knowledge discovery. One strategy for effective knowledge discovery is the use of human-computer collaboration.

One technique used for human-computer collaboration in the business information systems field is data visualization (Bell, 1991; Montazami & Wang, 1988) which is particularly relevant to data mining (Keim & Kriegel, 1996; Wang, 2002). From the human side of data visualization, graphics cognition and problem solving are the two major concepts of data visualization. It is a commonly accepted principle that visual perception is compounded out of processes in a way which is adaptive to the visual presentation and the particular problem to be solved (Kosslyn, 1980; Newell & Simon, 1972).

MAIN THRUST

Major components of interactive visual data mining and their functions that make data mining more effective are the current research theme in this field. Wang, S. and Wang, H. (2002) have developed a model of interactive visual data mining for human-computer collaboration knowledge discovery. According to this model, an interactive visual data mining system has three components on the computer side, besides the database: data visualization instrument, data and model assembly, and human-computer interface.

Data Visualization Instrument

Data visualization instruments are tools for presenting data in human understandable graphics, images, or animation. While there have been many techniques for data visualization, such as various statistical charts with colors and animations, the self-organizing maps (SOM) method based on Kohonen neural network (Kohonen, 1989) has become one of the promising techniques of data visualization in data mining. SOM is a dynamic system that can learn the topological relations and abstract struc-

tures in the high-dimensional input vectors using low dimensional space for representation. These low-dimensional presentations can be viewed and interpreted by human in discovering knowledge (Wang, 2000).

Data and Model Assembly

The data and model assembly is a set of query functions that assemble the data and data visualization instruments for data mining. Query tools are characterized by structured query language (SQL), the standard query language for relational database systems. To support human-computer collaboration effectively, query processing is necessary in data mining. As the ultimate objective of data retrieval and presentation is the formulation of knowledge, it is difficult to create a single standard query language for all purposes of data mining. Nevertheless, the following functionalities can be implemented through the design of query that support the examination of the relevancy, usefulness, interestingness, and novelty of extracted knowledge.

1. **Schematics Examination:** Through this query function, the data miner is allowed to set different values for the parameters of the data visualization instrument to perceive various schematic visual presentations.
2. **Consistency Examination:** To cross-check the data mining results, the data miner may choose different sets of data of the database to check if the conclusion from one set of data is consistent with others. This query function allows the data miner to make such consistency examination.
3. **Relevancy Examination:** It is a fundamental law that, to validate a data mining result, one must use external data, which are not used in generating this result but are relevant to the problem being investigated. For instance, the data of customer attributes can be used for clustering to identify significant market segments for the company. However, whether the market segments relevant to a particular product, one must use separate product survey data. This query function allows the data miner to use various external data to examine the data mining results.
4. **Dependability Examination:** The concept of dependability examination in interactive visual data mining is similar to that of factor analysis in traditional statistical analysis, but the dependability examination query function is more comprehensive in determining whether a variable contributes the data mining results in a certain way.

5. **Homogeneousness Examination:** Knowledge formulation often needs to identify the ranges of values of a determinant variable so that observations with values of a certain range in this variable have a homogeneous behavior. This query function provides interactive mechanism for the data miner to decompose variables for homogeneousness examination.

Human-Computer Interface

Human-computer interface allows the data miner to dialog with the computer. It integrates the data base, data visualization instruments, and data and model assembly into a single computing environment. Through the human-computer interface, the data miner is able to access the data visualization instruments, select data sets, invoke the query process, organize the screen, set colors and animation speed, and manage the intermediate data mining results.

FUTURE TRENDS

Interactive visual data mining techniques will become key components of any data mining instruments. More theories and techniques of interactive visual data mining will be developed in the near future, followed by comprehensive comparisons of these theories and techniques. Query systems along with data visualization functions on large-scale database systems for data mining will be available for data mining practitioners.

CONCLUSION

Given the fact that a one-shot knowledge deduction may not provide an alternative result if it fails, we must provide an integrated computing environment for the data miner through interactive visual data mining. An interactive visual data mining system consists of three intertwined components, besides the database: data visualization instrument, data and model assembly instrument, and human-computer interface. In interactive visual data mining, the human-computer interaction and effective visual presentations of multivariate data allow the data miner to interpret the data mining results based on the particular problem domain, his/her perception, specialty, and the creativity. The ultimate objective of interactive visual data mining is to allow the data miner to conduct the experimental process and examination simultaneously through the human-computer collaboration in order to obtain a "satisfactory" result.

Related Content

Deductive Data Warehouses: Analyzing Data Warehouses With Datalog (By Example)
Kornelije Rabuzin (2019). *Emerging Perspectives in Big Data Warehousing (pp. 58-82).*
www.irma-international.org/chapter/deductive-data-warehouses/231008

Employing Neural Networks in Data Mining
Mohamed Salah Hamdi (2005). *Encyclopedia of Data Warehousing and Mining (pp. 433-437).*
www.irma-international.org/chapter/employing-neural-networks-data-mining/10636

Databases Modeling of Engineering Information
Z. M. Ma (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 1182-1204).*
www.irma-international.org/chapter/databases-modeling-engineering-information/7693

Overview of Entity Resolution
 (2014). *Innovative Techniques and Applications of Entity Resolution (pp. 1-14).*
www.irma-international.org/chapter/overview-of-entity-resolution/103240

A Framework for Organizational Data Analysis and Organizational Data Mining
Bernd Knobloch (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 449-468).*
www.irma-international.org/chapter/framework-organizational-data-analysis-organizational/7659