On Interactive Data Mining

Yan Zhao
University of Regina, Canada

Yiyu Yao
University of Regina, Canada

INTRODUCTION

Exploring and extracting knowledge from data is one of the fundamental problems in science. Data mining consists of important tasks, such as description, prediction and explanation of data, and applies computer technologies to nontrivial calculations. Computer systems can maintain precise operations under a heavy information load, and also can maintain steady performance. Without the aid of computer systems, it is very difficult for people to be aware of, to extract, to search and to retrieve knowledge in large and separate datasets, let alone interpreting and evaluating data and information that are constantly changing, and then making recommendations or predictions based on inconsistent and/or incomplete data.

On the other hand, the implementations and applications of computer systems reflect the requests of human users, and are affected by human judgement, preference and evaluation. Computer systems rely on human users to set goals, to select alternatives if an original approach fails, to participate in unanticipated emergencies and novel situations, and to develop innovations in order to preserve safety, avoid expensive failure, or increase product quality (Elm, *et al*., 2004; Hancock & Scallen, 1996; Shneiderman, 1998).

Users possess varied skills, intelligence, cognitive styles, and levels of tolerance of frustration. They come to a problem with diverse preferences, requirements and background knowledge. Given a set of data, users will see it from different angles, in different aspects, and with different views. Considering these differences, a universally applicable theory or method to serve the needs of all users does not exist. This motivates and justifies the co-existence of numerous theories and methods of data mining systems, as well as the exploration of new theories and methods.

According to the above observations, we believe that interactive systems are required for data mining tasks. Generally, interactive data mining is an integration of human factors and artificial intelligence (Maanen, Lindenberg and Neerincx, 2005); an interactive system is an integration of a human user and a computer machine, communicating and exchanging information and knowledge. Through interaction and communication, computers and users can share the tasks involved in order to achieve a good balance of automation and human control. Computers are used to retrieve and keep track of large volumes of data, and to carry out complex mathematical or logical operations. Users can then avoid routine, tedious and error-prone tasks, concentrate on critical decision making and planning, and cope with unexpected situations (Elm, *et al*., 2004; Shneiderman, 1998). Moreover, interactive data mining can encourage users' learning, improve insight and understanding of the problem to be solved, and stimulate users to explore creative possibilities. Users' feedback can be used to improve the system. The interaction is mutually beneficial, and imposes new coordination demands on both sides.

BACKGROUND

The importance of human-machine interaction has been well recognized and studied in many disciplines. One example of interactive systems is an information retrieval system or a search engine. A search engine connects users to Web resources. It navigates searches, stores and indexes resources and responses to users' particular queries, and ranks and provides the most relevant results to each query. Most of the time, a user initiates the interaction with a query. Frequently, feedback will arouse the user's particular interest, causing the user to refine the query, and then change or adjust further interaction. Without this mutual connection, it would be hard, if not impossible, for the user to access these resources, no matter how important and how relevant they are. The search engine, as an interactive system, uses the combined power of the user and the resources, to ultimately generate a new kind of power.

Though human-machine interaction has been emphasized for a variety of disciplines, until recently it has not received enough attention in the domain of data mining (Ankerst, 2001; Brachmann & Anand, 1996; Zhao & Yao, 2005). In particular, the human role in the data mining processes has not received its due attention. Here, we identify two general problems in many of the existing data mining systems:

1. Overemphasizing the automation and efficiency of the system, while neglecting the adaptiveness and effectiveness of the system. Effectiveness includes human subjective understanding, interpretation and evaluation.
2. A lack of explanations and interpretations of the discovered knowledge. Human-machine interaction is always essential for constructing explanations and interpretations.

To study and implement an interactive data mining system, we need to pay more attention to the connection between human users and computers. For cognitive science, Wang and Liu (2003) suggest a relational metaphor, which assumes that relations and connections of neurons represent information and knowledge in the human brain, rather than the neurons alone. Berners-Lee (1999) explicitly states that "in an extreme view, the world can be seen as only connections, nothing else." Based on this statement, the World Wide Web was designed and implemented. Following the same way of thinking, we believe that interactive data mining is sensitive to the capacities and needs of both humans and machines. A critical issue is not how intelligent a user is, or how efficient an algorithm is, but how well these two parts can be connected and communicated, adapted, stimulated and improved.

MAIN THRUST

The design of interactive data mining systems is highlighted by the process, forms and complexity issues of interaction.

Processes of Interactive Data Mining

The entire knowledge discovery process includes data preparation, data selection and reduction, data pre-processing and transformation, pattern discovery, pattern explanation and evaluation, and pattern presentation (Brachmann & Anand, 1996; Fayyad, *et al*., 1996; Mannila, 1997; Yao, Zhao & Maguire, 2003; Yao, Zhong & Zhao, 2004). In an interactive system, these phases can be carried out as follows:

- Interactive data preparation observes raw data with a specific format. Data distribution and relationships between attributes can be easily observed.
- Interactive data selection and reduction involves the reduction of the number of attributes and/or the number of records. A user can specify the attributes of interest and/or data area, and remove data that is outside of the area of concern.
- Interactive data pre-processing and transformation determines the number of intervals, as well as cut-points for continuous datasets, and transforms the dataset into a workable dataset.
- Interactive pattern discovery interactively discovers patterns under the user's guidance, selection, monitoring and supervision. Interactive controls include decisions made on search strategies, directions, heuristics, and the handling of abnormal situations.
- Interactive pattern explanation and evaluation explains and evaluates the discovered pattern if the user requires it. The effectiveness and usefulness of this are subject to the user's judgement.
- Interactive pattern presentation visualizes the patterns that are perceived during the pattern discovery phase, and/or the pattern explanation and evaluation phase.

Practice has shown that the process is virtually a loop, which is iterated until satisfying results are obtained. Most of the existing interactive data mining systems add visual functionalities into some phases, which enable users to invigilate the mining process at various stages, such as raw data visualization and/or final results visualization (Brachmann & Anand, 1996; Elm, *et al*., 2004). Graphical visualization makes it easy to identify and distinguish the trend and distribution. This is a necessary feature for human-machine interaction,

Related Content

Learning Exceptions to Refine a Domain Expertise
Rallou Thomopoulos (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1129-1136).*
www.irma-international.org/chapter/learning-exceptions-refine-domain-expertise/10963

Temporal Event Sequence Rule Mining
Sherri K. Harms (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1923-1928).*
www.irma-international.org/chapter/temporal-event-sequence-rule-mining/11082

Data Mining in Protein Identification by Tandem Mass Spectrometry
Haipeng Wang (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 472-478).*
www.irma-international.org/chapter/data-mining-protein-identification-tandem/10862

Mining Group Differences
Shane M. Butler (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1282-1286).*
www.irma-international.org/chapter/mining-group-differences/10987

Data Streams
João Gamaand Pedro Pereira Rodrigues (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 561-565).*
www.irma-international.org/chapter/data-streams/10876