International Journal of Research in Science and Technology
(IJTBM) 2012, Vol. No. 2, Issue No. IV, Oct-Dec
http://www.ijrst.com/
ISSN: 2249-0604

A GUIDED CLUSTERING TECHNIQUE FOR KNOWLEDGE DISCOVERY

Diksha Sharma[1] Dr. Kalyankar N.V.[2]

[1]Research Scholar, CMJ University, Shillong, Meghalaya
[2]Principal, Science & Engineering, Yashvant College, Nanded, Maharashtra

ABSTRACT

Scalability: *As visualization techniques are used more and more in data mining tools, it becomes important to address scalability issue. Fundamentally, the reason for scalable solutions is to be able to build a good data mining model as quickly as possible. This offers two benefits, a) there is value in being able to deploy and use the model sooner rather than later and b) faster turnaround times yield better models. Time previously spent waiting for results can instead be devoted to finding the model that result in the best and most reliable solution.*

Making data mining tools scalable requires hardware scalability and parallel algorithms. The goal of hardware scalability is to provide high performance by adding modestly priced processor building blocks (or nodes) in such a way that performance scales linearly.

Key words: *- visualization techniques, scalability.*

INTRODUCTION

Extracting hidden patterns in medial domain is becoming increasingly relevant today as records of patients clinical trials and other attributes are available electronically. Traditionally, decision making in health care was based on the ground information, past experiences and funds constraints. But with the use of various data mining and knowledge discovery techniques, a knowledge rich health environment can be created.

Clustering based data mining techniques have been implemented (discussed in Chapter 7) to find clusters in the liver disorder patient's database. Domain users play a useful role in identifying patterns through the use of visualisation tools by using their heuristics. The experiment takes user's feedback torun the clustering algorithm and searches for crisp patient clusters in the liver disorder dataset. Such techniques can go a long way in the analysis of inconsistencies in medical classifications as well as in other domains of health care.

REVIEW OF LITERATURE

There are a number of obstacles to successful operationalisation of health care system guidelines due to incomplete and imprecise information. To evaluate the health care system, particularly in India, such techniques are required which can handle inconsistency and uncertainty in data. A Fuzzy Inference System (FIS) (described in Chapter 8) has been developed to evaluate the performance of immunisation program in different states of India.

The designed FIS classifies the immunisation program of Indian states into 'Well-Organised', 'Satisfactory' and 'Poorly-Organised' categories based on the number of children immunised and infant mortality rate. This type of classification facilitates identification of areas needing attention and more resources. The fuzzy logic rules can also infer critical patterns, results and relationships from the availablehealth care data which will further help in proper monitoring and categorisation of the health programs, supervision and their performances.

Several applications of interactive data mining were explored (and reported in research papers published or accepted for publication in various Journals) as a part of the present research work, in addition to already existing applications of data mining techniques. As far as contemporary work on applications of knowledge and data mining is concerned, one can refer to Chapter 2 for various references. Still a lot needs to be explored as far as human interactivity in the process of knowledge and data mining is concerned. In this section, some areas of interactivity in data mining process are discussed wherein scope of human involvement in the applications of data mining exist for the betterment of the overall knowledge discovery process.

MATERIAL AND METHOD

To involve domain user in the data mining process, an interface must be designed through which user acquires knowledge about the current state of KDD process and is enabled to manipulate a mining algorithm through interaction. Useful data mining should not only be efficient, but also be effective. For data mining algorithms, especially for classification, cluster analysis, decision tree, etc., an interactive processing between machine and user is necessary.

The interaction can have textual or visual interface. In today's visualisation platform, an interactive interface is essential in interactive data mining. A Graphical User Interface (GUI) is a type of user interface item that allows people to interact with programs in more ways than typing such as computers; hand-held devices such as MP3 Players, Portable Media Players or Gaming devices; household appliances and office equipment with images rather than text commands [Window Manager Definition, PC Magazine, 2008]. In this context, interface can be designed for text mining, clustering or association rules. A well designed interface reduces the iterations within the knowledge discovery processloop as well as user trusts more in the results. New modes[2] and modalities[3] in user interface can work to take into account the human psychology and physiology of the users, and make the process of using the system effective, efficient and satisfying.

Through careful observation and analysis, it can be concluded that human beings give their specification in two forms: 1) how the data is to be mined, and 2) what mining results are required. Correspondingly, these two categories are named as suggestions and demands. Demands can be expressedby query predicate and suggestions are expressed by rules. The data mining tools should accommodate more graphic user interfaces, for example, touch user interface[4], object-oriented user interface[5], voice user interfaces[6], natural-language interfaces[7], zooming user interfaces[8], reflexive user interfaces[9] and many more, to help the domain user express their demands more precisely.

CONCLUSION

It is important to remember that the amount of data to be mined is ever increasing over time. Furthermore, other problems to which you wish to apply data mining may be larger or more complex thanyour current problem. Thus, issues like scalability, security and efficiency should be given more concern in upcoming researches.

For example, by doubling the number of processors, twice as much data in the same time or the same amount of data in half the time can be processed. The hardware vendors have taken two complementary approaches of using multiple CPUs: Symmetric Multi-Processing (SMP) and Massively Parallel Processing (MPP). Data mining algorithms written for a uniprocessor machine won't automatically run faster on a parallel machine; they must be rewritten to take advantage of the parallel processors. In addition to this, the databases can also be made parallel for simultaneous processing. Most data warehouses today make use of parallel database management systems. However, because the data mining process may involve data from multiple databases, or subsets of data from a data warehouse, it is usually a good idea to create a data mart for purposes of mining that data. This avoids the problem of consolidating data every time you want to use it.

Security: Data mining, the discovery of new and interesting patterns in large datasets, has immense potential and applicability, and is generating a lot of interest these days, in public as well as private sectors. Advances in technology and business processes like growth of computer networks used toconnect databases, development of enhanced search related techniques, such as neural networks and advanced algorithms, spread of the client/server computing model allowing users to access centralised data resources from the desktop, and an increased ability to combine data from disparate sources into a single searchable source have focused on the advantages of data mining in the current scenario.

Recently there has been a realisation that data mining has an impact on security, especially the use of data mining to improve security. A significant example is the use of data mining for intrusion detection. A second aspect is the potential security hazards posed when an adversary has data mining capabilities. A lot of work can be done in the area of implementing security in the data mining systems.

Validation: In any knowledge discovery system, it is important to validate and improve the effectiveness of the system through its use in real applications with participation of domain experts. New validation techniques can be developed to evaluate performance by adding meta-knowledge in the data mining algorithm profiles and integrating on-line rules. It is important that you validate your mining models by understanding their quality and characteristics before you deploy them into a production environment.

There are several approaches for assessing the quality and characteristics of a data mining model. The first includes the use of various measures of statistical validity to determine whether there are problems in the data or in the model. Second, data can be separated into training and testing sets to testthe accuracy of predictions. Third, business experts may be asked to review the results

of the data mining model to determine whether the discovered patterns have meaning in the targeted business scenario. New techniques can be worked out to improve the quality and properties of the data mining system.

Artificial Intelligence and Machine Learning

With its roots in statistics, artificial intelligence, machine learning and data-mining have been around since the 1990s. Artificially Intelligent (AI) powered tools are able to deal with uncertain and incomplete data sets. A lot of new applications can be found which can combine all these techniques in many new environments. A few areas are explained here.

Smart Health Care: Technological enhancements have advanced research towards creating smart homes and intelligent environments. Ongoing research focuses on using technology as a means of supporting the independent life of older people and people with disabilities. Numerous intelligent devices are now available, which provide the resident with 24-hour health monitoring. The goal of smart home- care is to develop technologies that allow early detection and remediation of physical and cognitive decline on the sensor-based data in a smart environment. The next big step for smart health-care would be remote diagnosis and administration of medications, based on the data collected from smart homes and environments, avoiding frequent hospital visits. .

Many data mining techniques can prove out to be extremely valuable to the medical industry, particularly in the area of in-office, or online medical diagnosis. Using neural nets as a data mining solution may allow medical facilities of the future to present patients with the option of receiving online medical advice from an artificial intelligence software program. In other words, an artificial intelligence computer program using data mining techniques, diagnoses malignant melanoma just as well as a trained dermatologist.

Auto-expanding Learning Capability: Machine learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. A major focus of machine learning research is to automatically learn to recognise complex patterns and make intelligent decisions based on data. Some machine learning systems attempt to eliminate the need for human intuition in data analysis, while others adopt a collaborative approach between human and machine. Human intuition cannot, however, be entirely eliminated, since the system's designer must specify how the data is to be represented and what mechanisms will be used to search for a characterisation of the data.

The focus of new research can be on 1) *transduction learning* - which tries to predict new outputs based on training inputs, training outputs, and test inputs, and 2) *learning to learn* - which is inductive- bias learning based on previous experience. Furthermore, *reinforcement learning* is another type of learning which explains how to act in a given observation of the world. Every action has some impact on environment, and the environment provides feedback in form of rewards that guides the learning algorithm.

New Techniques in Data Mining

Since data mining is a growing area, the techniques are constantly changing, as new improved methods are discovered. There are many techniques within data mining that aim to accomplish different tasks. This subsection throws light on the new techniques of data mining which are adaptive and imitate the working of human brain.

REFERENCES

Adriaans P. and Zantinge D. (1999). Introduction to Data Mining and Knowledge Discovery. 3rd Edition Potomac, MD: Two Crows Corporation.

Adriaans P. and Zantinge D. (2003). Data Mining. Pearson Education, Seventh Indian Reprint, 2003.

Agrawal R. and Srikant R. (1994). Fast Algorithms for Mining Association Rule. ***Proceedings of the 20th*** *International Conference on Very Large Databases (VLDB), 487 – 499.*

Agrawal R., Faloutsos C. and Swami A. (1993). Efficient similarity search in sequence databases. Proceedings of the Fourth International Conference on Foundations of Data Organisation and Algorithms, Chicago, Vol. 730, 69-84.

Agrawal R., Imielinski T. and Swami A. (1993). Mining association rules between sets of items in large databases. ***Proceedings of the 1993 ACM SIGMOD International Conference on Management of*** *Data, Washington DC, 207-216.*

Ahmed S.R. (2007). Applications of Data Mining in Retail Business. International Conference on Information Technology: Coding and Computing, Las Vegas, Nevada, Vol. 2.

Anand S.S., Bell D.A. and Hughes J.G. (1995). The Role of Domain Knowledge in Data Mining. ***Proceedings of the Fourth International Conference on Information and knowledge management, 37-43.***

Ankerest M. (2001). Human Involvement and Interactivity of the Next generation's Data Mining Tools. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. Santa Barbara, CA.

Ankerest M., Ester M. and Kriegel H.P. (2000). Towards an Effective Cooperation of the User and the Computer for Classification. Proceedings of 6th International conference on Knowledge Discovery and Data Mining, Boston, MA.

Aruna P., Puviarasan N. and Palaniappan B. (2005). An Investigation of Neuro-Fuzzy Systems in Psychosomatic Disorders. Expert Systems with Applications. Vol. 28, 673-679.

Bates J.H.T. and Young M.P. (2003). Applying Fuzzy Logic to Medical Decision Making in the Intensive Care Unit. American Journal of Respiratory and Critical Care Medicine, Vol. 167, 948-952.

Bayrak C., Kolukisaoglu H. and Chia-Chu Chiang. (2006). Di-Learn: Distributed Knowledge Discovery with Human Interaction. IEEE International conference on Systems, Man and Cybernetics, Taipei, Taiwan, Vol. 4, 3354 – 3359.

Ding C. and He X. (2004). k-Means Clustering via Principal Components Analysis. ACM Proceedings of the 21st International Conference on Machine Learning, Vol. 69, page 29.

Edelstein H.A. (1999). Introduction to Data Mining and Knowledge Discovery (3rd Edition) Potomac, MD: Two Crows Corp.

Ester M., Kriegel H.P. and Sander J. (2001). Algorithms and Applications for Spatial Data Mining. Published in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis.

Fawcett T. and Provost F. (1997). Adaptive Fraud Detection. Data Mining and Knowledge Discovery, Vol. 1(3):291-316.

Fayyad U. M., Piatetsky-Shapiro G. and Smyth P. (1996). From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data mining, AAAI Press, 1-34.

Fayyad U., Piatetsky-Shapiro G. and Smyth P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of ACM, Vol. 39, 27-34.

Forgionne G.A., Gagopadhyay A. and Adya M. (2000). Cancer Surveillance Using Data Warehousing, Data Mining, and Decision Support Systems. Topics in Health Information Management, Proquest Medical Library, Vol. 21(1), 21-34.

Frank H., Klawonn F., Kruse R. and Runkler T. (1999). Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. New York: John Wiley.

Frawley W.J., Piatetsky-Shapiro G. and Matheus C.(1996). Knowledge Discovery in Databases: An Overview. Knowledge Discovery in Databases, AAAI Press/MIT Press, Cambridge, MA., Menlo Park, C.A, 1-30.

G. Fort and Lambert-Lacroix S. (2005). Classification using Partial Least Squares with Penalised Logistic Regression. England: Bioinformatics-Oxford, 21(7), 1104-1111.