Vol 2 Issue 6 March 2013

ISSN No : 2249-894X

Monthly Multidisciplinary Research Journal

Review Of Research Journal

Chief Editors

Welcome to Review Of Research

RNI MAHMUL/2011/38595 **ISSN No.2249-894X**

Review Of Research Journal is a multidisciplinary research journal, published monthly in English, Hindi & Marathi Language. All research papers submitted to the journal will be double - blind peer reviewed referred by members of the editorial Board readers will include investigator in universities, research institutes government and industry with research interest in the general subjects.

Advisory Board

Address:-Ashok Yakkaldevi 258/34, Raviwar Peth, Solapur - 413 005 Maharashtra, India
Cell : 9595 359 435, Ph No: 02172372010 Email: ayisrj@yahoo.in Website: www.isrj.net

Review Of Research
Vol. 2, Issue. 6, March. 2013
ISSN:-2249-894X
Available online at www.ror.isrj.net

ORIGINAL ARTICLE

INTERDISCIPLINARY RESEARCH ON KNOWLEDGE DISCOVERY DATABASES WITH HIGHLY INTERACTIVE HUMAN CANTERED ENVIRONMENTS

LALITA AND VIRENDER SINGH SANGWAN

Associate Professor ,CSE Deptt. Manav Instituite Technology and Manangent , JEVRA(HISAR)
Lecturer in CSE ,Govt. Polytechnic Loharu(Bhiwani)

Abstract:

This paper describe the role of domain expert in the process of Knowledge Discovery in Databases (KDD). In order to get better results, various steps of knowledge discovery process are extended with the domain specific expertise of the users. Intermediate results after each step are provided to the domain expert to check whether the process is proceeding in the right direction or there exists a need to change certain variables [Ankerest M., 2001; Hellerstein J.M. et.al., 1999; Anand S.S. et.al.,1995].Even though the quality of results heavily relies on data preparation. Preparing data includes data selection, data cleaning, data transformation and data validation etc. These initial steps of exploring, visualizing and querying the data, to gain insight into the data are very vital and are better done in an interactive manner.

KEYWORD :
Knowledge Discovery Databases

INTRODUCTION:

Data mining is an interdisciplinary research area which integrates machine learning, artificial intelligence, statistics and database theories. Discovering hidden knowledge from the databases may be time consuming. A possible solution is to allow the domain expert to intervene in the otherwise automatic process and use his domain knowledge to alter parameters on-the-fly so that the final results are near to what the user wants to extract. Credibility and usability of the results of knowledge discovery and data mining process and the degree of user confidence in these results can be enhanced by integrating and utilizing human expertise and his domain knowledge. It is important to note that users possess various skills, intelligence, cognitive styles, frustration tolerance and other mental abilities [Li J. et.al., 2007]. Different users attempt to solve a problem with different preferences, requirements and background knowledge.

Data mining and knowledge discovery system would be better off if 1) the user would be able to pose a query of the form “Give me something interesting that could be useful and 2) the system would be able to discover some knowledge that is useful to the user [Anand S.S., et.al., 1995]. This would require human involvement in the data mining process. Pressing the need for highly interactive human cantered environments would enable both 1) human assisted computer discovery and 2) computer assisted human discovery.

1 HUMAN INVOLVEMENT IN The Process Of KNOWLEDGE DISCOVERY

The basic task of the knowledge discovery and data mining process is to extract knowledge from data such that the resulting knowledge is useful in a given situation. Obviously, only the user can determine whether the resulting knowledge satisfies the requirements or not. Instead of allowing an automated data mining process to iterate in a trial and error manner, a better but largely overlooked way to enhance the

knowledge discovery process is to provide a domain expert support through human involvement [Anand S.S. et.al., 1995].

Human beings possess expertise that is gained by long-time working experience [Li J. et.al. 2007]. But as computers' having human human guidance, expertise and experiences are currently irreplaceable in discovering and understanding knowledge. There are broadly two ways to utilize the human expertise in mining useful patterns: 1) guide the computer how to mine the data i.e. suggestions and 2) tell the computer what is expected as the mining result i.e. demands. Suggestions with regard to how to mine the data better require expertise and experience into data mining process [Ankerest M., 2001]. As we all know, human abilities of abstract thinking, understanding, and intuition outstrip computers. Specifying what to bring out means putting constraints on the system. Quite often users have specific expectations from the mining results. In such cases a user can pass his demands to the system which allows the data mining process to go in the desired direction without exhausting all the possibilities.

Using data mining as a strategic tool for organization is not merely a technical challenge, but it also involves matching relevant data and structuring the organization data so that the organization can actually utilize the new knowledge and transform it into an organizational asset.

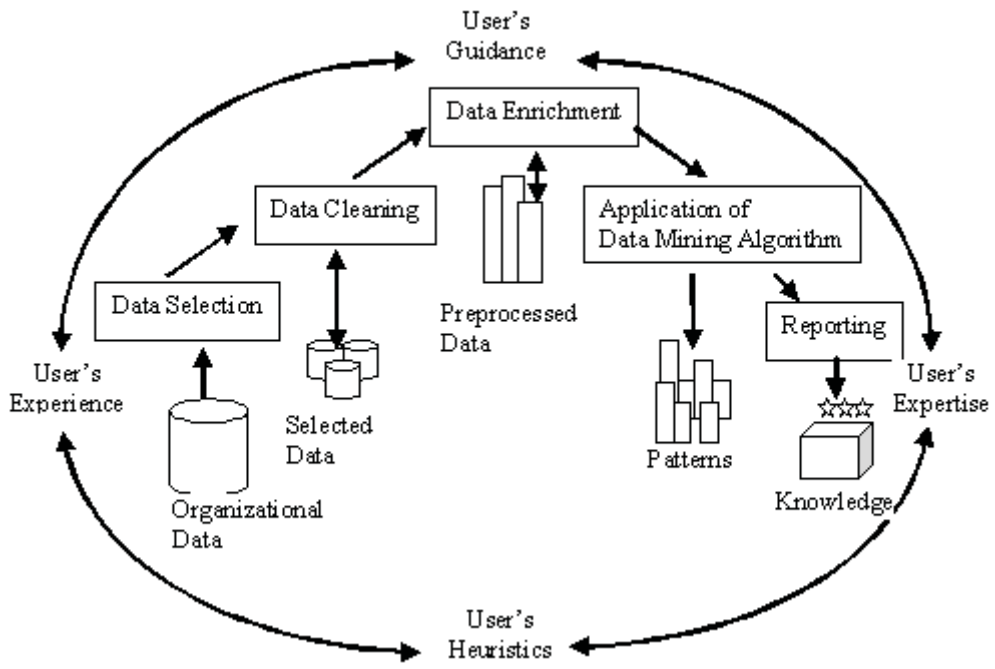


Figure 1.1: Interactive knowledge discovery process.

1.1 Data Selection.

The knowledge discovery process is initiated with information and knowledge requirement. The data mining process is aimed at finding new and interesting patterns and rules in the data, so first of all it is required to ascertain what kind of knowledge is expected out of data. Purpose and nature of data required derives the entire knowledge discovery process. The first phase of knowledge discovery deals with the selection of data for mining purpose. In most cases data are selected from the operational data stored in the information system of the organization [Adriaans P. et.al., 2003]. In order to facilitate the knowledge discovery process, a copy of operational data is drawn and stored in a separate database. Interactive data selection involves selecting the attributes and records in the operational data on which mining will be performed. Domain user plays an important role here as the operational data in different parts of the organization vary in quality. Some departments maintain high quality databases containing information that is vital for their operation, while others may have small data sets built on an adhoc basis but are still valuable from data mining perspective. Some databases are updated on a day-to-day basis; others may contain information that date back to several years.

1.2 Data Cleaning

Real world data tend to be incomplete, noisy and inconsistent. So data must be cleaned before the mining can be performed. Data cleaning routines attempt to fill in missing values, re-duplicate the data and remove inconsistencies in the data.

In the operational data there are many tuples that have no recorded value for several attributes. The mining algorithm might not give accurate results for the data with missing values. There are several ways to deal with missing values 1) ignore the tuple, 2) fill in the missing value manually, 3) use a global constant, 4) compute mean to fill in the missing value or 5) use the most probable value to fill in the missing value [Rohm E. et.al., 2000]. Hence a lot of human interaction is required at this stage. An algorithm based upon pattern analysis techniques could identify the probable duplicate records and present it to the user to make a decision. If duplicate records are not checked out, the discovered patterns might not reflect the actual trend in the data.

Another type of error that frequently occurs is lack of domain consistency. Suppose in the company's operational data if there is a date recorded as 1 January, 1901, although the company probably did not exist at that time. This type of error is particularly damaging because it is hard to trace, but it will greatly influence the type of patterns discovered. Manual reference is required to detect data inconstancies.

1.3 Data Enrichment

Data enrichment is performed to optimize the data for mining purpose. At this stage some new attributes may be added in the existing tables to make the data more detailed regarding the facts [Edelstein H.A., 1999]. Presently, independent data marts1 are available that can provide access to stored databases on a variety of subjects. The data can also be arranged from other organizations. Matching the information from third party databases with internal data is cumbersome and does not lend itself to automation. Enriching data by buying additional information is not a simple, straightforward task, and constitutes an ongoing process in the organization. A user may help examine the attributes and tables in the operational and historical data to be mined from the enrichment point of view.

1.4 Interactive Data Mining

At the discovery stage in the KDD process one or more data mining algorithms are executed and it is proposed that the data mining algorithm be run interactively. Machine learning and pattern recognition algorithms are applied to the data that have been pre-processed and transformed into a homogeneous format [Marakas G.M., 2003]. Data mining algorithm may do well on one part of the data set while others may not yield that much useful result. Depending on the nature of the data, suitable data mining techniques need to be selected on-the-fly, therefore presenting ample avenues for user interaction during the mining phase of knowledge discovery. Data mining involves fitting the data to a model that describes its behaviour. The choice of optimal parameters for modelling process is an iterative process and is better done with user involvement. The results of data mining process are subjected to interpretation by a domain expert and data miner. Domain experts judge the results within domain context, while data miners use data mining criteria to evaluate the results. Before deployment, the modellers evaluate the models for accuracy and generality.

The goal of interactive system design is to integrate user's experience and domain knowledge into the entire data mining process.

1.5 Process Reporting

The reporting stage combines two distinct functions 1) analysis of the generated data mining model and 2) application of the results of the data mining model to new data. During the last leg of knowledge discovery process three activities are carried out namely: 1) reporting, 2) analysis and 3) visualizations. The objective of this phase is to evaluate the model with respect to problem solving perspective. It is advisable to test and evaluate the models by applying it on real-life problems. If the results fulfil the problem objectives, data mining application can be deployed, otherwise, all the phases of knowledge discovery such as data preparation, modelling and evaluation have to be repeated with changed parameters.

2 KNOWLEDGE DISCOVERY IN BLOOD TRANSFUSION DATA SET

Data mining, in contrast to traditional data analysis, is discovery driven. Most of the recent data mining tools provide automatic pattern recognition and attempt to uncover patterns in data that are difficult to detect with traditional statistical methods. However, the increasing size of result sets does not mean increasing discoveries of knowledge embedded in the data. Most of the result patterns are not meaningful to the domain users at all. The users have to take another painful process to select really interesting patterns in

the result set.

2 . Pre-processing the Blood Transfusion Dataset

A case study of real-world database has been considered to implement a human-interactive knowledge discovery project. Since the process of knowledge discovery begins with data selection, the researcher chose blood transfusion dataset for the data mining experiment. The target dataset has been donated by Prof. I-Cheng Yeh, Department of Information Management, Chung-Hua University, Hsin Chu, Taiwan and contains the data of the Blood Transfusion Service Centre. The dataset was accessed from UCI Machine Learning Repository website http://archive.ics.uci.edu:80/ml/datasets.html. There are 5 attributes and 748 instances in the dataset. An RFMTC2 model was built from the dataset of 748 donors [Yeh C. et.al., 2009].

The description of the dataset is given in Table

Attribute Name	Description
Regency	months since last blood donation.
Frequency	total number of donation.
Monetary	total blood donated in c.c..
Time	months since first donation/
Donated	a binary variable representing whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood).

Table 3.1: Description of Blood Transfusion Dataset.
(Source: Blood Transfusion Dataset - http://archive.ics.uci.edu:80/ml/datasets.html)

The Blood Transfusion Service Centre faces a classification problem with regard to blood donated by the donors. The Centre collects different parameters (Table 3.1) for donors and predicts whether the donor will donate blood in March 2007. The blood transfusion dataset was pre-processed before performing experiments in consultation with a domain expert. The data was first converted into desired MS-EXCEL format and cleaned for missing values while maintaining the integrity and intactness of the dataset. Furthermore the 'monetary' attribute was normalized to make it suitable for mining.

2.2 Data Mining in Blood Transfusion Dataset

After pre-processing the dataset, the next step was to apply data mining algorithm(s). There are various data mining techniques available with their suitability dependent on the domain application. The data mining may help in answering several important and critical questions related to present application domain such as: 'What factors are more crucial to predict the blood donating frequency of a person?' or 'What is the average time gap between two blood donations of a person?' However there is a concern of patient privacy. The role of data mining is not to practice the outcomes but to fetch useful information and knowledge, so that better understanding and health care can be provided.

The blood transfusion dataset was mined by two different techniques. Firstly, outlier mining3 was applied to the dataset to find those records which are considerably dissimilar, exceptional or inconsistent with respect to the remaining data. The second data mining technique that has been used was decision tree4. The following subsections explain these two experiments.

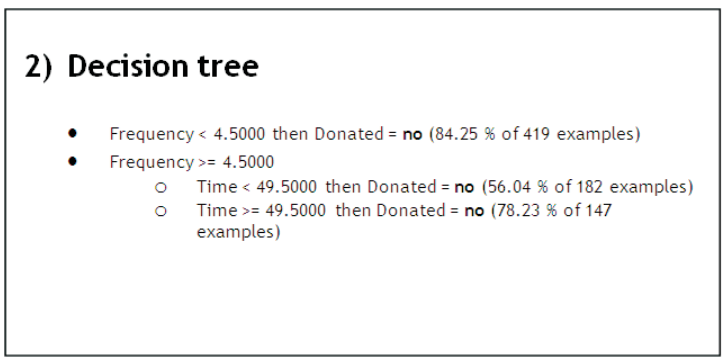
2.2.1 Outlier Mining in Blood Transfusion Dataset

Outliers in medical databases can be caused by measurement errors or may be the result of inherent data variability. For example, the display of a person's age as '-1' could be the result of a typographical error. Alternatively, outliers may be the results of inherent data variability such as abnormal value of blood pressure could be a major indication of a patient's critical situation. TANAGRA - a free, open-source, user-friendly software product developed by Ricco Rakotomalala, has been used to conduct the experimentation. Tanagra supports a host of analytical functions such as binary logistic regression, k-nearest neighbor, neural network trained with back-propagation, Quinlan's ID3 (Iterative Dichotomiser 3), linear discriminate analysis, and naive Bayesian classifier [Rakotomalala R., 2003].

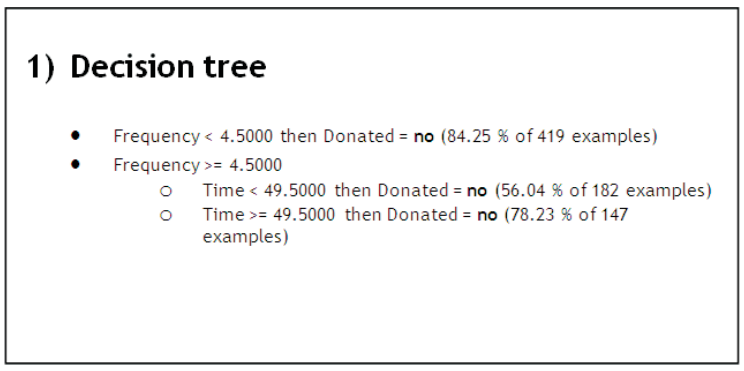
Outliers represent abnormal value which may be an indication of hidden information in the database. Studying the extraordinary behaviour of outliers helps uncovering the valuable knowledge hidden behind them and aiding the decision makers to improve the health care services.

2.2.2 Decision Tree Technique in Blood Transfusion Dataset

ID3 decision tree algorithm has been run using TANAGRA data mining software. In this application domain, the purpose was to find the input factors which determine the blood donation of a person. Therefore, 'Donated' attribute was set as the target attribute, and 'Frequency' and 'Time' were set as input to the algorithm. The Figure 2.1 shows the results of the ID3 decision tree algorithm on the training dataset (randomly selected sample from the database). By observing the results, one can see that if 'Frequency' is less than 4.5000, then, 84.25% chances are that the person will not donate the blood in the current month. Figure 1.2: Decision tree for the user-specified attributes.



The knowledge derived this way can be used to form rules in the whole database. Such rules can help the domain users to predict the future cases in the same domain. The following rule i.e. Rule2 can be formed



from the decision tree constructed (Figure 2.2).
If Frequency < 2. 2THEN Donated IS 'NO' Rule2

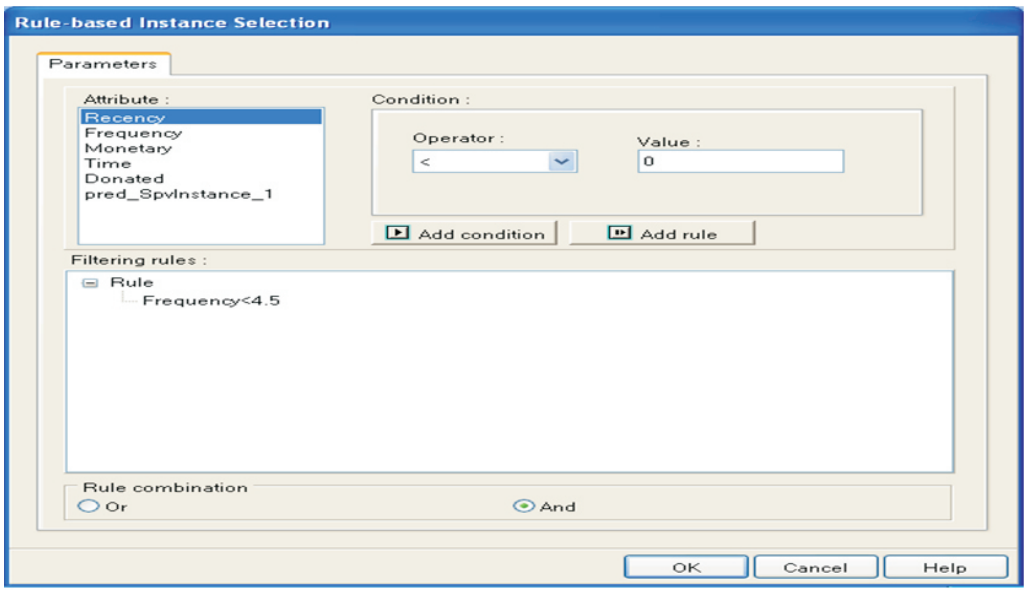


Figure Design of for instance selection.

Now this rule can be used to fetch the records from the database and check generality of the rule with respect to the whole database. Figure shows the records from the blood transfusion database. Such data mining techniques can discover valuable information from the massive health related datasets and can support the health care planners in decision making.

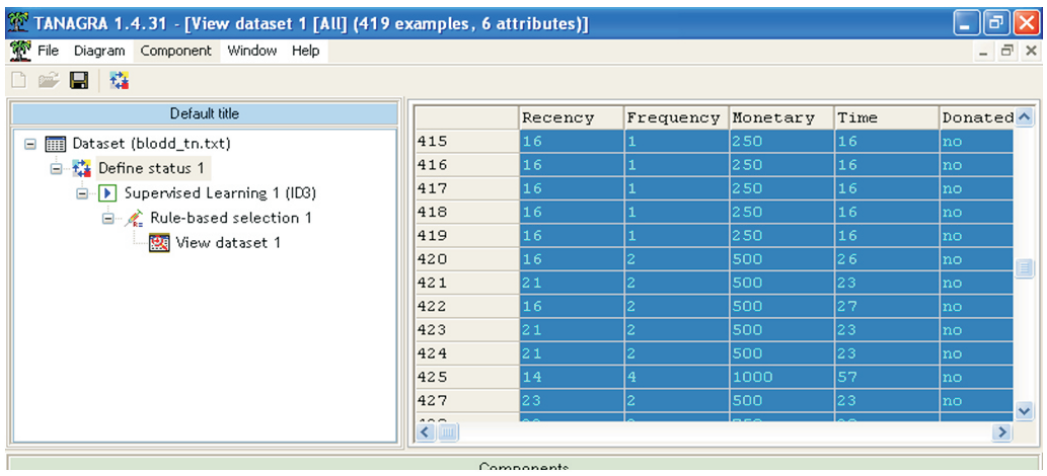


Figure1.2: Instance selection

4 CONCLUSION

The usefulness of the final results can be enhanced by roping in the Domain Expert and hence, there is a need to articulate the role of Domain Expert in the rule discovery process. The interactive data mining experiments carried out in this chapter could help the health professionals in better management of blood bank facility. Management in blood transfusion service is concerned with identification and selection of prospective blood donors, adequate collection of blood, and ensuring the safest and most appropriate use of blood/blood components.

REFERENCES:

1) Adriaans P. and Zantinge D. (2003). Data Mining. Pearson Education, Seventh Indian Reprint, 2003.
2) 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.
3) Agrawal R., Faloutsos C. and Swami A. (1993). Efficient similarity search in sequence databases. Proceedings of the Fourth International Conference on Foundations of Data Organization and Algorithms, Chicago, Vol. 730, 69-84.
4) 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.
5) Ahmed S.R. (2007). Applications of Data Mining in Retail Business. International Conference on Information Technology: Coding and Computing, Las Vegas, Nevada, Vol. 2.
6) 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.
6) 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.
7) 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.

•