Ankit Bhardwaj, Arvind Sharma, V.K. Shrivastava / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue4, July-August 2012, pp.1303-1309

# Data Mining Techniques and Their Implementation in Blood Bank Sector –A Review

**Ankit Bhardwaj[1], Arvind Sharma[2], V.K. Shrivastava[3]**
[1] M.Tech Scholar, Department of CSE, HCTM, Kaithal, Haryana, India
[2]Head, Department of Computer Science, DAV Kota, Rajasthan, India
[3]Head, Department of I.T, HCTM, Kaithal, Haryana, India

## ABSTRACT

**This paper focuses on the data mining and the current trends associated with it. It presents an overview of data mining system and clarifies how data mining and knowledge discovery in databases are related both to each other and to related fields. Data Mining is a technology used to describe knowledge discovery and to search for significant relationships such as patterns, association and changes among variables in databases. This enables users to search, collect and donate blood to the patients who are waiting for the last drop of the blood and are nearby to death. We have also tried to identify the research area in data mining where further work can be continued.**

**Keywords:** Data mining, KDD, Association Rules, Classification, Clustering, Prediction.

## I. INTRODUCTION

In today's computer age data storage has been growing in size to unthinkable ranges that only computerized methods applied to find information among these large repositories of data available to organizations whether it was online or offline. Data mining was conceptualised in the 1990s as a means of addressing the problem of analyzing the vast repositories of data that are available to mankind, and being added to continuously. Data mining is necessary to extract hidden useful information from the large datasets in a given application. This usefulness relates to the user goal, in other words only the user can determine whether the resulting knowledge answers his goal. The growing quality demand in the blood bank sector makes it necessary to exploit the whole potential of stored data efficiently, not only the clinical data and also to improve the behaviours of the blood donors. Data mining can contribute with important benefits to the blood bank sector, it can be a fundamental tool to analyse the data gathered by blood banks through their information systems. In recent years, along with development of medical informatics and information technology, blood bank information system grows rapidly. With the growth of the blood banks, enormous Blood Banks Information Systems (BBIS) and databases are produced. It creates a need and challenge for data mining. Data mining is a process of the knowledge discovery in databases and the goal is to find out the hidden and interesting information [1]. Various important steps are involved in knowledge discovery in databases (KDD) which helps to convert raw data into knowledge. Data mining is just a step in KDD which is used to extract interesting patterns from data that are easy to perceive, interpret, and manipulate. Several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization, and meta-rule guided mining will be reviewed. The explosive growth of databases makes the scalability of data mining techniques increasingly important. Data mining algorithms have the ability to rapidly mine vast amount of data.

The paper is organised as follows: In section 2 we use the data mining process and several stages of the KDD. Section 3 presents the literature survey. Section 4 discusses about the research methodology to be proposed in our work. Finally, section 5 concludes the paper.

## II. DATA MINING AND KDD PROCESS

Data mining is a detailed process of analyzing large amounts of data and picking out the relevant information. It refers to extracting or mining knowledge from large amounts of data [2]. Data Mining is the fundamental stage inside the process of extraction of useful and comprehensible knowledge, previously unknown, from large quantities of data stored in different formats, with the objective of improving the decision of companies, organizations where the data can be collected. However data mining and overall process known as Knowledge Discovery from Databases (KDD), is usually an expensive process, especially in the stages of business objectives elicitation, data mining objectives elicitation, and data preparation. This is especially the case each time data mining is applied to a blood bank. Data Mining can be defined as the extraction or fetching of the relevant information i.e. Knowledge from the large repositories of data. That's the reason it is also called as Knowledge Mining. However many

synonyms are linked with Data Mining viz. Knowledge Mining from Data, Knowledge extraction, Data/ pattern analysis, Data archaeology and Data dredging. Data Mining is also popularly known as Knowledge Discovery in data bases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in Databases.

Many people treat the data mining as a synonym for another popularly used term, Knowledge Discovery from Data. The following figure (fig.1), shows the data mining as simply an essential step in the process of KDD i.e. Knowledge Discovery from Data[3].

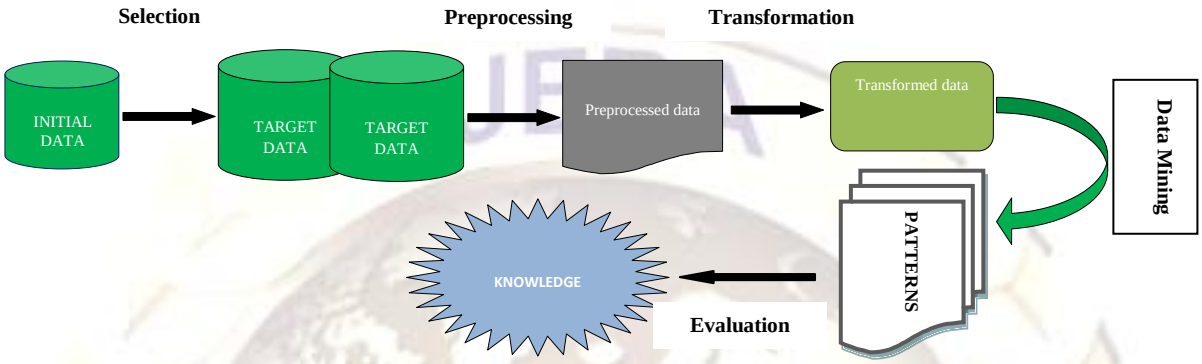


Fig.1: Data mining is the core of KDD Process

The KDD process includes selecting the data needed for data mining process & may be obtained from many different & heterogeneous data sources. Preprocessing includes finding incorrect or missing data. There may be many different activities performed at this time. Erroneous data may be corrected or removed, whereas missing data must be supplied. Preprocessing also include: removal of noise or outliers, collecting necessary information to model or account for noise, accounting for time sequence information and known changes. Transformation is converting the data into a common format for processing. Some data may be encoded or transformed into more usable format. Data reduction, dimensionality reduction (e.g. feature selection i.e. attribute subset selection, heuristic method etc) & data transformation method (e.g. sampling, aggregation, generalization etc) may be used to reduce the number of possible data values being considered. Data Mining is the task being performed, to generate the desired result. Interpretation/Evaluation is how the data mining results are presented to the users which are extremely important because the usefulness of the result is dependent on it. Various visualization & GUI strategies are used at this step. Different kinds of knowledge requires different kinds of representation e.g. classification, clustering, association rule etc.[4]

## 2.1 STEPS OF KDD PROCESS

The knowledge discovery in databases (KDD) process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process contains of the following steps:

**1. Data cleaning**

It also known as data cleansing, it is a fundamental step in which noise data and irrelevant data are removed from the collection.

**2. Data integration**

In this step, multiple data sources, often heterogeneous, may be combined in a common source.

**3. Data selection**

At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

**4. Data transformation**

It is also known as data consolidation, it is a stage in which the selected data is transformed into forms appropriate for the mining procedure.

**5. Data mining**

It is the core step of KDD process in which clever techniques are applied to extract patterns potentially useful.

**6. Pattern evaluation**

In this step, strictly interesting patterns representing knowledge are identified based on given measures.

**7. Knowledge representation**

This is the final step in which the discovered knowledge is visually represented to the user. This is a very essential step that uses visualization techniques to help users understand and interpret the data mining results.

### 2.2 DATA MINING MODELS

The data mining models and tasks are shown in figure 2 given below:

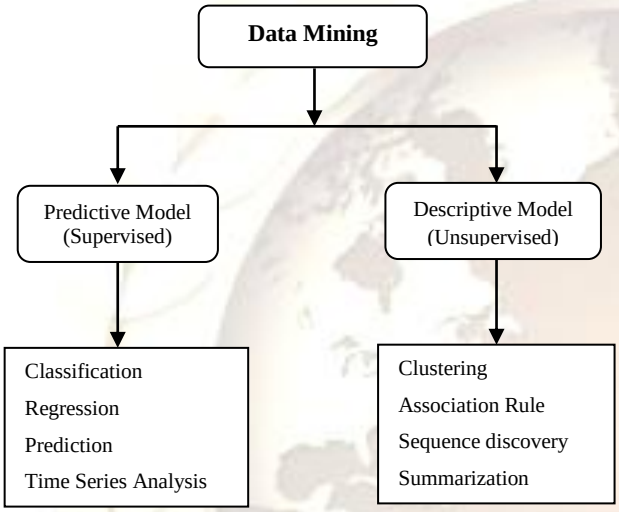


Fig.2: Data Mining Models and Tasks

The predictive model makes prediction about unknown data values by using the known values. The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined.

### 2.3 DATA MINING TASKS

There are different types of data mining tasks depending on the use of data mining results. These data mining tasks are categorised as follows:

**1. Exploratory Data Analysis:** It is simply exploring the data without any clear ideas of what we are looking for. These techniques are interactive and visual.

**2. Descriptive Modeling:** It describes all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

**3. Predictive Modeling:** This model permits the value of one variable to be predicted from the known values of other variables.

**4. Discovering Patterns and Rules:** It concern with pattern detection, the aim is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest.

**5. Retrieval by Content:** It is finding pattern similar to the pattern of interest in the data set. This task is most commonly used for text and image data sets.

## III. LITERATURE SURVEY

In this section, lot of research works have been recorded from past few years. They are presented here:

[8] In June 2009, work is carried out on management information system that helps managers in providing decision making in any organization. This MIS is basically all about the process of collecting, processing, storing and transmitting the relevant information that is just like the life blood of the organization that uses a data mining approach.

[7] In June 2009, the research work entitled on web based information system for blood donation, all the information regarding blood donation are available on world wide web i.e. online systems that communicate/interconnect all the blood donor societies in a country using LAN Technology.

[6] In July 2009, the research work entitled on Segmentation, Reconstruction and Analysis of Blood thrombus formation in 3D-2 photons microscopy images. In this work, method and platform are applied to differentiate between the thromby formed in wild type and low FV11 mice. The high resolution quantitative structural analysis using some algorithms that provide a new matrix, that is likely to be critical to categorize and understanding bio-medically relevant characteristics of thromby. Also, in this work, composition of different components on the clot surface and number of voxels in each clot component has been computed.

[5] In December 2009, the work proposed on an application to find spatial distribution of blood donors from the blood bank information system, blood donation and transfusion services are carried out and help the patients to access the availability of blood from anywhere.

[9] In the year 2009, the research was based geographical variation in correlates of blood donor turn out rates-an investigation of Canadian metropolitan areas. In this work, Canadian blood services i.e. an organization that aims the collecting and distributing the blood supply across the country. Accessibility variables are introduced and calculated using a 2-step floating catchment area in order to analyse the accessibility of services. The regression technique is also used.

[10] In the year 2010, the research is accomplished on application of CART algorithm in blood donor's classification. In this work, one of the popular data

mining technique i.e. Classification is used and through the use of CART application, a model is created that determines the donor's behavior.

[11] In the year 2010, the data mining tool is used to extract information from PPI (Protein-Protein Interaction) systems and developed a PPI search system known as PP Look that is an effective tool 4-tier information extraction systems based on a full sentence parsing approach. In this paper, researchers introduced a useful tool that is PP Look which uses an improved keyword, dictionary pattern matching algorithm to extract protein-protein interaction information from biomedical literature. Some visual methods were adopted to conclude PPI in the form of 3D stereoscopic displays.

[12] In December 2010, the work is proposed on binary classifiers for health care databases i.e. a comparative study of data mining classification algorithms in the diagnosis of breast cancer. This work helps in uncovering the valuable knowledge hidden behind them and also helping the decision makers to improve the health care services. In this work, the presented experiment provides medical doctors and health care planers a tool to help them quickly make sense of vast clinical databases.

[15] In February 2011, the research paper presented on classifying blood donors using data mining techniques, the work is performed on blood group donor data sets using classification technique.

[13] In August 2011, the research work carried out on interactive knowledge discovery in blood transfusion data set, the work is done through conducting data mining experiments that help the health professionals in better management of blood bank facility.

[16] In year 2011, the work presented on rule extraction for blood donors with fuzzy sequential pattern mining, fuzzy sequential pattern algorithm is used to extract rules from blood transfusion service center data set that predicts the behaviour of donor in the future.

[17] In August 2011, the research work proposed on a comparison of blood donor classification data mining models, which uses decision tree to examine the blood donor's classification. In this work, comparison between extended RVD based model and DB2K7 procedures are carried out and it was discovered that RVD classification is better than DB2K7 in aspects of recalling and precision capability.

[18] In the September 2011, the research work carried out on real time blood donor management using dash boards based on data mining models, the data mining techniques are used to examine the blood donor classification and promote it to development of real time blood donor management using dash boards with blood profile or RVD profile and geo-location data.

[14] In year 2011, the work was proposed on an intelligent system for improving performance of blood donation. In this work, many important techniques i.e. Clustering, K-means classification is adopted that improves the performance of blood donation services.

[19] In January 2012, the data mining techniques are applied on medical databases and clinical databases which are used to store huge amount of information regarding patient's diagnosis, lab test results, patient's treatments etc. which is a way of mining the medical information for doctors and medical researchers. In all developed countries, it is found that the routine health tests are very common among all adults. It is also determined that precautionary measures are less expensive rather than the treatments and also facilitates a better chance for patient's treatment at earlier stages. In this research work, five medical databases are used and experimental results are computed using data mining software tool.

[20] In February 2012, the work is carried out through the Health Care Applications to diagnose different diseases using Red Blood Cells counting. In this work, the researchers presented the automatic process of RBC count from an image and some of the data mining techniques like Segmentation, Equalization and K-means clustering are used for preprocessing of the images. Some diseases which are related to sickness of RBC count also predicted.

[21] In this work, the main purpose of the system is to guide diabetic patients during the disease. Diabetic patients could benefit from the diabetes expert system by entering their daily glucoses rate and insulin dosages; producing a graph from insulin history; consulting their insulin dosage for next day. It's also tried to determine an estimation method to predict glucose rate in blood which indicates diabetes risk. In this work, the WEKA data mining tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared.

## IV. RESEARCH METHODOLOGY

### 4.1 Data Mining Techniques in Blood Bank Sector

In our proposed methodology, we suppose a blood bank system has the following objectives:

- Increase blood donor's rate and their attitude.
- Attract more blood donors to donate blood.
- Assist blood bank professionals in making policy on the acquisition of blood donors and new blood banks.

There are several major data mining techniques had been developed and used in data mining research projects recently including classification, clustering, association, prediction and sequential patterns. These techniques and methods in data mining need brief mention to have better understanding.

#### 4.1.1 Classification

The estimation and prediction may be viewed as types of classification. The problem usually is evaluating the training data set and then we apply to the future model.

The following table1 shows different classification algorithms:

| Type | Name of algorithm |
|---|---|
| Statistical | Regression<br>Bayesian |
| Distance | Simple distance<br>K-nearest neighbors |
| Decision Tree | ID3<br>C4.5<br>CART<br>SPRINT |
| Neural Network | Propagation<br>NN Supervised learning<br>Radial base function network |
| Rule based | Genetic rules from DT<br>Genetic rules from NN<br>Genetic rules without DT and NN |

Table 1: Classification Algorithms

#### 4.1.2 Clustering

Clustering technique is grouping of data, which is not predefined. By using clustering technique we can identify dense and sparse regions in object space. The following table2 shows different clustering algorithms:

| Type | Name of algorithm |
|---|---|
| Similarity and distance measure | Similarity & distance measure |
| Outlier | Outlier 3333 |
| Hierarchical | Agglomerative, Divisive |
| Partitional | Minimum spanning tree<br>Squared matrix<br>K-means<br>Nearest Neighbor<br>PAM<br>Bond Energy<br>Clustering with Neural Network |
| Clustering large database | BIRCH<br>DB Scan<br>CURE |
| Categorical | ROCK |

Table 2: Clustering Algorithms

#### 4.1.3 Association Rule

The central task of association rule mining is to find sets of binary variables that co-occur together frequently in a transaction database, while the goal of feature selection problem is to identify groups of that are strongly correlated with each other with a specific target variable. Association rule has the several algorithms like: Apriori, CDA, DDA, interestingness measure etc. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule can be 'if a customer buys a dozen eggs, he is 80% likely to also purchase milk.' An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

The association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. The *support* is an indication of how frequently the data items appear in the database. The *confidence* indicates the number of times the if/then statements have been found to be true.

Types of Association Rule Techniques are as follows:

- Multilevel Association Rule
- Multidimensional Association Rule
- Quantitative Association Rule

#### 4.1.4 Prediction

The prediction as it name implied is one of the data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, prediction analysis technique can be used in blood donors to predict the behaviour for the future if we consider donor is an independent variable, blood could be a dependent variable. Then based on the historical data, we can draw a fitted regression curve that is used for donor's behaviour prediction. Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction.

Types of Regression Techniques are as follows:

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

#### 4.1.5 Sequential Patterns

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

## V. CONCLUSION

This paper presents an overview of data mining and its techniques which have been used to extract interesting patterns and to develop significant relationships among variables stored in a huge dataset. Data mining is needed in many fields to extract the useful information from the large amount of data. Large amount of data is maintained in every field to keep different records such as medical data, scientific data, educational data, demographic data, financial data, marketing data etc. Therefore, different ways have been found to automatically analyze the data, to summarize it, to discover and characterize trends in it and to automatically flag anomalies. The several data mining techniques are introduced by the different researchers. These techniques are used to do classification, to do clustering, to find interesting patterns. In our future work, the data mining techniques will be implemented on blood donor's data set for predicting the blood donor's behaviour and attitude, which have been collected from the blood bank center.

## REFERENCES

*[1] Badgett RG: How to search for and evaluate medical evidence. Seminars in Medical Practice 1999, 2:8-14, 28.*

*[2] Jaiwei Han and Micheline Kamber, 'Data Mining Concepts and Techniques', Second Edition, Morgan Kaufmann Publishers.*

*[3] 'Data Mining Introductory and Advanced Topics' by Margaret H. Dunham.*

*[4] Sunita B Aher, LOBO L.M.R.J., International Conference on Emerging Technology Trends (ICETT)-2011 Proceedings published by International Journal of Computer Applications (IJCA).*

*[5] B.G. Premasudha et al., 'An Application to find spatial distribution of Blood Donors from Blood Bank Information',Vol. II, Issue No.2, July-December 2009.*

*[6] Jian Mu et al., 'Segmentation, Reconstruction, and analysis of blood thrombus formation in 3 D-2photon microscopy images', EURASIP Journal on Advances in Signal Processing, 10 July 2009.*

*[7] Abdur Rashid Khan et and al. 'Web based Information System for Blood Donation', International Jounal of digital content Technology and its applications,Vol. 3, Issue No.2, July 2009.*

*[8] G. Satyanaryana Reddy et al; 'Management Information System to help Managers for providing decision making in any organization'.*

*[9] T. Santhanm and Shyam Sunderam: 'Application of Cart Algorithm in Blood donor's classification', Journal of computer Science Vol. 6, Issue 5.*

*[10] Zhangetal. 'PPLook: an automated data mining tool for protein–protein interaction', BMC Bio-informatics- 2010.*

*[11] Dr. Varun Kumar, Luxmiverma; 'Binary classifiers for Health Care databases - A comparative study of data mining classification algorithms in the diagnosis of Breast cancer', IJCST, Vol. I, Issue 2, December 2010.*

*[12] Vikram Singh and Sapna Nagpal; 'Interactive Knowledge discovery in Blood Transfusion Data Set'; VSRD International Journal of Computer Science and Information Technology; Vol. I, Issue 8, 2011.*

*[13] Wen-ChanLee and Bor-Wen Cheng; 'An Intelligent system for improving performance of blood donation', Journal of Quality, Vol. 18, Issue No. II, 2011.*

*[14] P.Ramachandran et al. 'Classifying blood donors using data mining techniques'; IJCST, Vol. I, Issue1, February 2011.*

*[15] F. Zabihi et al. 'Rule Extraction for Blood donators with fuzzy sequential pattern mining'; The Journal of mathematics and Computer Science; Vol. II, Issue No. I, 2011.*

*[16] Shyam Sundaram and Santhanam T: 'A comparison of Blood donor classification data mining models', Journal of Theoretical and Applied Information Technology, Vol.30, No.2, 31 August 2011.*

*[17] Shyam Sundaram and Santhanam T; 'Real Time Blood donor management using Dash boards based on Data mining models', International Journal of Computing issues, Vol.8, Issue5, No.2, September 2011.*

*[18] Prof. Dr. P.K. Srimani et al. 'Outlier data mining in medical databases by using statistical methods', Vol.4, No.1, January 2012.*

*[19] Alaa hamouda et al; 'Automated Red Blood Cells counting', International Journal of Computing Science, Vol.1, No.2, February 2012.*

*[20] Ivana D. Radojevic et al. 'Total coliforms and data mining as a tool in water quality monitoring', African Journal of Microbiology Research, Vol. 6(10), 16 March 2012.*

*[21] P.Yasodha, M. Kannan; Analysis of a Population of Diabetic Patients Databases in Weka Tool; International Journal of Scientific & Engineering Research Volume 2, Issue 5, May 2011.*

.

## Author's Profile

Ankit Bhardwaj received the B.Tech. Degree in Information Technology from Maharishi Dayanand University, Rohtak, Haryana, India in 2009 and pursuing M.Tech (Comp.Sc. & Engg) from Haryana College of

Technology and Management, Kaithal, Kurukshetra University, Kurukshetra. His research interest lies in the area of Data Mining.

Arvind Sharma received the M.Sc. Degree in Computer Science from Maharishi Dayanand University, Rohtak, Haryana, India in 2003, M.C.A. Degree from JRNRV University Udaipur, Rajasthan, India, M.Phil. Degree in Computer Science with specialization in Web Data Mining Applications from the Alagappa University, Karaikudi, Tamil Nadu, India and pursuing Ph.D. Computer Science from Jaipur National University, Jaipur, Rajasthan, India. He has published many technical papers in International and National reputed journals and conferences. His research interest lies in the area of Data Mining and Web Applications.

V.K. Shrivastava is working as an Associate Professor and HOD, Department of Information Technology, Haryana College of Technology and Management, Kaithal, Haryana, India. He has published many research papers in International and National reputed journals and conferences. His research interest lies in the area of Digital Signal Processing.