Classifying Blood Donors Using Data Mining Techniques

[1]P.Ramachandran, [2]Dr.N.Girija, [3]Dr.T.Bhuvaneswari,
[1] CSA Dept., SCSVMV University, Enathur, Kanchipuram - 631 561
[2] IT Dept.,Higher College of Technology, Muscat, Ministry of Manpower,
[3] Eng Dept., Dr.M.G.R University, Maduravayal, Chennai-605 013

ABSTRACT
Data mining refers to extracting knowledge from large amount of data. Real life data mining approaches are interesting because they often present a different set of problems for data miners. The process of designing a model helps to identify the different blood groups with available stock in Indian Red Cross Society (IRCS) Blood Bank Hospital Classification techniques for analysis of Blood bank data sets. The availability of blood groups in blood banks is a critical and important aspect in a Blood bank. Blood banks are typically based on a healthy person voluntarily donating blood and used for transfusions or made into medications. The ability to identify regular blood donors will enable blood bank and voluntary organizations to plan systematically for organizing blood donation camps in an efficient manner. The analysis had been carried out using a standard blood group donor's dataset and using the J48 decision tree algorithm implemented in Weka. The research work is used to classify the blood donors based on the sex, blood group, weight and age. This may be achieved through collecting the data utilizing the data mining technique and choosing the most suitable implementation tool for the domain.

Keywords- Data Mining, Blood Donors, Blood groups, Classification, Weka

I. INTRODUCTION

Data mining provides automatic pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods. A Data Warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. A data warehouse is also often viewed as architecture constructed by integrating data from multiple heterogeneous sources to support structured and/or ad-hoc queries, analytical reporting and decision making.

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives, databases and summarizing it into useful form called information - information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational database. Data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.

Data mining can be used to uncover patterns in data but is often carried out only on samples. The mining process will be ineffective if the samples are not a good representation of the larger body of data. Data mining cannot show up patterns that may be presented in the larger body of data if those patterns are not present in the sample being "mined". An important part of the process is the verification and validation of patterns on other samples of data.

Data mining commonly involves four classes of task:

Classification - Arranges the data into predefined groups. For example an email program might attempt to classify an email as legitimate or spam. Common algorithms include Decision Tree Learning, Nearest Neighbor, Bayesian classification and Neural Network.

Clustering - Is like classification but the groups are not predefined, so the algorithm will try to group similar items together.

Regression - Attempts to find a function which models the data with the least error.

Association rule learning - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently brought together and use this information for marketing purposes.

II NEED AND IMPORTANCE OF RESEARCH PROBLEM

Modern world has experienced a dramatic increase in the amount of data stored online. With the widespread use of medical information systems including databases, there is an explosive growth in their sizes; Physicians and Surgeons are faced with a problem of making use of stored data.

Massive healthcare data needs to be converted into information and knowledge, which can help control cost and

maintains high quality of patient care. Without data mining it is difficult to realize the full potential of data collected within healthcare organization as data under analysis is massive, highly dimensional, distributed and uncertain. The traditional manual data analysis has become insufficient and methods for efficient computer assisted analysis indispensable. Application of data mining and knowledge discovery and database techniques are very beneficial but highly challenging in the field of medical and health care.

The data mining algorithm will be tested on blood group donor's data. The blood group donor's database for this research consists of blood groups maintained by *Indian Red Cross Society (IRCS) Blood Bank Hospital*, Chennai.

III. REVIEW OF LITERATURE

T. Santhanam and Shyam Sundaram et al [3] The CART derived model along with the extended definition for identifying regular voluntary donors provided a good classification accuracy based model. Masser *et al* [4] have developed a framework that helps determining the predictors of the intentions and behavior of established blood donors. Anil Rajput, RameshPrasadAharwal, Nidhi Chandel, Devenra Singh Solanki and Ritu Soni et al [5] Application of data mining and knowledge discovery and database techniques are very beneficial but highly challenging in the field of medical and health care.

Ferguson and Chandler et al [6] have used qualitative studies to demonstrate that blood donors describe their behavior using Trans Theoretical Model (TTM). The government of India et al [7] through the National AIDS Control Organization has developed a detailed Voluntary Blood Donation Programme. This provides an operational guideline which provides some valuable information into the foundation of blood donor ship. Mohamed Mostafa et al [9] use intelligent modeling techniques to examine the effect of various demographic, cognitive and psychographic factors on blood donation in Egypt. This research used variable sets such are sex, age, educational level, altruistic values, perceived risks of blood donation, blood donation knowledge, attitudes toward blood donation and intention to donate blood. Neural network based models were employed in this research. From an India specific context Chaudhary (2009) discusses specific overall governance and controls that have been developed in the area of blood transfusion services.

Bharucha (2005) addresses specific areas of improving blood donor ship/management in the implementation of a quality system. Strategies towards donor recruitment and retention have been presented from a south East Asian perspective. Nevine M. Labib, Michael N. Malek et al [13] Childhood Acute Lymphoblastic Leukemia (also called acute lymphocytic leukemia or ALL) is a cancer of the blood and bone marrow. This type of cancer usually gets worse quickly if it is not treated. It is the most common type of cancer in children. Devchand Chaudhari , Dr. Ravindra , S. Hegadi et al [14] the process of designing a model that can help in blood platelet transfusion database maintained in Maxcare Hospital, which has a great significance in the health care field.

IV. WEKA OPEN SOURCE

The **WEKA** (Waikato Environment for Knowledge Analysis) software was developed in the University of New Zealand. A number of data mining methods are implemented in the WEKA software. Some of them are based on decision trees like the J48 decision tree, some are rule-based like ZeroR and decision tables, and some of them are based on probability and regression, like the Naïve Bye's algorithm.

The data that is used for WEKA should be made into the ARFF (Attribute Relation file format) format and the file should have the extension dot ARFF (.arff). WEKA is a collection of machine learning algorithms for solving real world data mining problems. It is written in Java; WEKA runs on almost any platform and is available on the web at www.cs.waikato.ac.nz/ml/weka.

Fig 1.Weka Software

1 DATA SET

Data is taken from Blood Donors record office of *Indian Red Cross Society (IRCS) Blood Bank Hospital* Chennai. Blood group donors list consist many fields such as Bag number, name, age, sex, blood group, and weight, HIV etc, these data is typed in MS Access. Description Of data is summarized in the table 1.

Attribute	Instance	Numeri	Nominal	Class
5	2387	4	2	6

Tabel.1Data set

2 PREPROCESSING DATA

After data is loaded, it is shown in the 'Preprocess' panel of the Explorer. Summary statistics are available for

every attribute from the dataset. If the attribute is nominal the distribution of the instances according the attribute values is shown. If the attribute is numerical the minimum, maximum, mean and standard deviations are given.

Statistic	Value
Minimum	3055
Maximum	6999
Mean	5678.162
StdDev	938.775

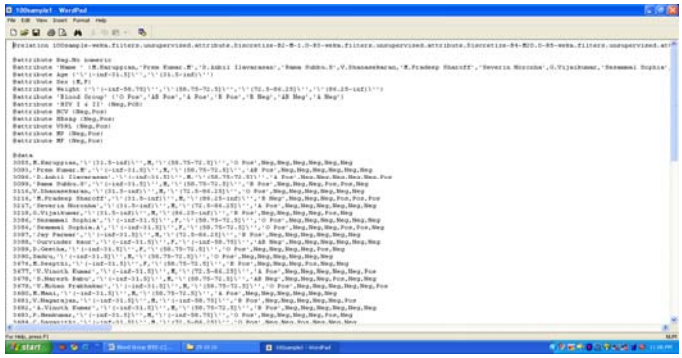

Fig .2 Preprocess Data

The blood groups are available in *Indian Red Cross Society (IRCS) Blood Bank Hospital,* Chennai. The following chart diagram demonstrates the availability of blood groups.

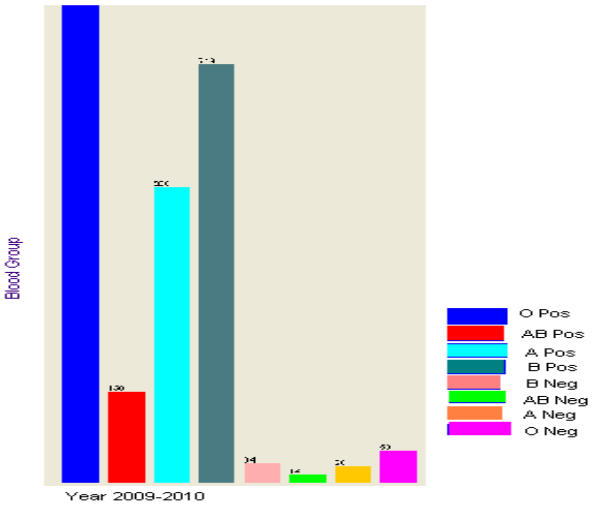

Fig .3 Blood Group Type

3 DECISION TREE GENERATED USING WEKA TOOL

The decision tree algorithm that we used in WEKA, J48, gives us an opportunity to control the confidence factor and training sample size (controlled by the cross-validation option). Our objective is to get a decision tree that minimizes the expected error rate, with the highest amount of correctly classified instances. Give us the highest amount of correct classification; hence the decision tree was generated with 90% confidence and 10-fold cross validation
This decision tree is shown in Figure 4.

In WEKA, the confidence factor is used to address the issue of tree pruning. When a decision tree is being built, many of the branches will reflect anomalies due to noise or outliers in the training data. Tree pruning uses statistical Measures to remove these noise and outlier branches, allowing for the Confidence factor. This means that our dataset did not have much noise or outlier cases, so there was not much to prune faster classification and improvement in the ability of the tree to correctly classify independent test data (Han & Kamber, 2006).

A smaller confidence factor will incur more pruning, so for example if a 98% confidence factor is used, our tree will incur less pruning. We ran WEKA with a very wide range of confidence factors, but the results were not reacting to.

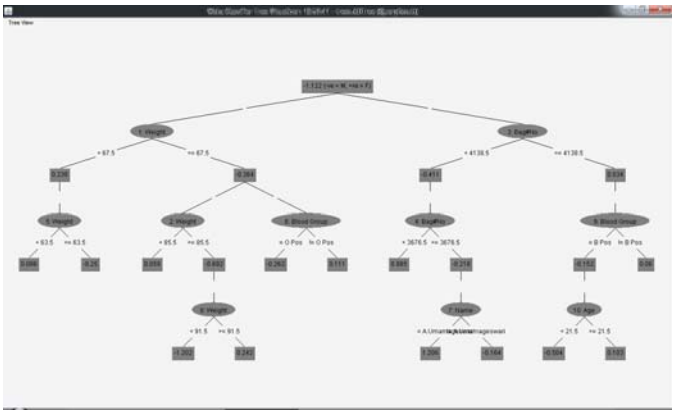

Fig.4 Decision tree in Weka

V. WEKA GENERATED OUTPUT FOR J48 ALGORITHM

```
=== Run information ===
  Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
Sex = M: '(60-80]' (2158.0/876.0)
Sex = F
|   Age = '(-INF-27.75]': '(-INF-60]' (182.0/86.0)
|   Age = '(27.75-38.5]'
|   |   Bag#No <= 6326: '(60-80]' (25.0/4.0)
|   |   Bag#No > 6326: '(-inf-60]' (10.0/4.0)
|   Age = '(38.5-49.25]': '(60-80]' (7.0/3.0)
|   Age = '(49.25-INF)': '(-inf-60]' (4.0/2.0)
Number of Leaves: 6
Size of the tree:       9

==========Confusion Matrix ===========
TP Rate         FP Rate         Precision       Recall
 0.096          0.043           0.464           0.096
 0.952          0.929           0.589           0.952
 ==================================
```

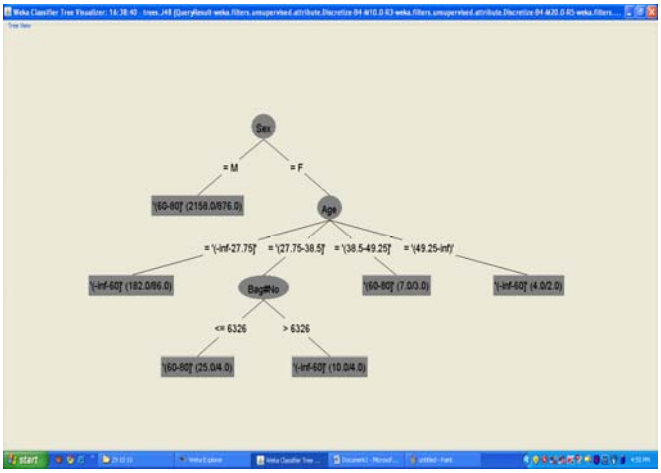

Fig 5. Output for Decision Tree in Weka

VI. CONCLUSION

In this research paper we have described classification techniques for Blood Group Donors datasets. We have used data mining classifiers to generate decision tree. The primary focus of this research is the development of a system that is essential for the timely analysis of huge Blood Group Donors data sets. The traditional manual data analysis has become insufficient and the methods for efficient computer assisted analysis indispensable.

This technique will be applied to the blood group transfusion database maintained in the *Indian Red Cross Society (IRCS) Blood Bank Hospital* Chennai.

This algorithm will be adapted to find conditions under which blood groups are frequently requested during emergency situations.

The future work can be applied to blood type classification, diagnosing diabetic symptoms, classifying blood donor type and diagnosing cancer patients based on blood cells or predicting the cancer types on the blood groups.

ACKNOWLEDGEMENTS

My sincere thanks to Indian Red Cross Society (IRCS) Blood Bank Hospital for providing facility for data collection, Prof.N.R.AnanthaNarayanan, &HOD, Porf.M.Kannan, Prof.V.Ramesh, Prof.S.Babu and other faculty members of Computer Science &Application department for providing valuable suggestions.

REFERENCES

[1] Indian Red Cross Society (IRCS) Blood Bank Hospital, Chennai.

[2] Jiawei Han and Micheline Kamber "Data Mining Concepts and Techniques "2nd Edition.

[3] T. Santhanam and Shyam Sundaram "Application of CART Algorithm in Blood Donors Classification**"** Journal of Computer Science 6 (5): 548-552, 2010 ISSN 1549-3636, pp.01-05 © 2010 Science Publications

[4] Masser, M.B., White, M. Katherine, Hyde and K. Melissa *et al*., 2009. "Predicting blood donation intentions and behavior among Australian blood donors: Testing an extended theory of planned Behavior model" Transfusion, 49: 320-329. DOI: 10.1111/j.1537-2995.2008.01981.x

[5] Anil Rajput, RameshPrasadAharwal, Nidhi Chandel, Devenra Singh Solanki and Ritu Soni" Approaches of Classifications to Policy of Analysis of Medical Data" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.11, November 2009,pp.01-09

[6] Ferguson, E. and S. Chandler, 2005. "A stage model of blood donor behavior: Assessing volunteer behaviors. Health Psychol., 10: 359-372. DOI: 10.1177/1359105305051423

[7] Government of India, 2007. "Voluntary blood donation programme" http://www.nacoonline.org/upload/Final%20Publications/Blood%20Safety/voluntary%20blood%20do nation.pdf

[8] Ian Witten, H. and Eibe Frank, 2005. "Data Mining: Practical Machine Learning Tools and Techniques." 2nd Edn., Morgan Kaufmann, San Francisco. ISBN: 0-12-088407-0, pp: 560.

[9] Mohamed Mostafa, M., 2009." Profiling blood donors in Egypt: A neural network analysis" Expert Syst. Appli., 36: 5031-5038. DOI: 10.1016/j.eswa.2008.06.048

[10] Paul Harper, R., 2005. "A review and comparison of classification algorithms for medical decision making" Health Policy, 71: 315-331. DOI: 10.1016/j.healthpol.2004.05.002

[11] Schlumpf, K.S., S.A. Glynn, G.B. Schreiber, D.J. Wright and W. Randolph Steele *et al*., 2007. "Factors influencing donor return". Transfusion, 48: 264-72. DOI: 10.1111/j.1537-2995.2007.01519.x

[12] Soman, K.P., S. Diwakar and V. Ajay, 2006. "Insight into Data Mining-Theory and Practice" Prentice Hall of India, New Delhi, ISBN: 81-203-2897-3.

[13] Nevine M. Labib, and Michael N. Malek" Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia" World Academy of Science, Engineering and Technology 8 2005, pp: 1-6

[14] Devchand Chaudhari & Dr. Ravindra S. Hegadi" Data Mining in Blood Platelets Transfusion Using Classification Rule:" pp: 1-8

[15] Rossen Dimov et al., Weka: Practical machine Learning Tools and Techniques -April 30, 2007.