Data Mining in Blood Platelets Transfusion using Classification Rule

Devchand J. Chaudhari
Assistant Professor
Depts. of Computer Science & Engg.
Government College of Engineering,
Chandrapur, Maharashtra, India.

Mamta Ramteke
Visiting Faculty
Depts. of Computer Science & Engg.
Government College of Engineering,
Chandrapur, Maharashtra, India

Manoj G. Lade
Visiting Faculty
Depts. of Computer Science & Engg.
Government College of Engineering,
Chandrapur, Maharashtra, India

ABSTRACT

Data mining provides automatic pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods. 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. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery. The goal of this project will be to develop data mining algorithm for transfusion of blood platelets. This algorithm will be adapted to find conditions under which transfusions were successful and those under which platelet transfusions were unsuccessful. This proposal will be useful for finding cancer patient's dataset.

Keywords
Data Mining, Blood Platelets, Transfusion, Classification.

1. INTRODUCTION

Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is becoming an increasingly important tool to transform these data into information. 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. 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 present 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 namely classification, clustering, regression, and association rule learning.

2. OBJECTIVES

Position The goal of this project will be to develop data mining algorithm for transfusion of blood platelets. In this proposal, our primary goal is to design a model by analyzing patients' medical records, which may include nearly 25 attributes. The attributes include information such as patient's temperature, patient's age, patient's HLA type, donor's HLA type, transfused HLA type, none illness and any other information that may be available. HLA compatibility is the common measure upon which patient's platelets matched with donor platelets. This measure is not sufficient for predicting the success of transfusion. There exist medical conditions that may affect the survivability of transfused platelets. A patient's temperature may influence his/her response to a transfusion. Complications such as septicemia, splenomegaly, and intravascular coagulation may also affect a patient's response to transfusions. The objective of this project will be to mine this platelet transfusion database and generate data that could be used in decision support system. A Classification algorithm and Artificial Neural Network (ANN) will be adopted and applied for mining platelet transfusion database. The classification algorithm will be used to find different conditions under which platelet transfusions were successful and those under which platelet transfusions were unsuccessful. Again the goal of this proposal will be to find those combinations of factors that have resulted in positive and negative responses to platelet transfusions.

3. RESEARCH METHODOLOGY

Classification process consists of training set that are analyzed by a classification algorithms and the classifier or learner model is represented in the form of classification rules. Test data are used in the classification rules to estimate the accuracy. Classification is an important data mining tool that analyses a given training set and develop a model for each class according to the features present in the data. Rule Induction, Decision Tree, and Artificial Neural Network classification techniques are used in this proposal.

3.1 Rule Induction

It is the process of extracting useful 'if-then' rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules.

Knowledge represents has the form:

IF conditions THEN conclusion

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction. In the health care system it can be applied as follows:

(Symptoms)(Previous---history)------> (Cause—of--- disease)

Rule Induction Method has the potential to use retrieved cases for predictions.

Sr. No.	Attribute Name
01	Age(in years)
02	Sex(M/F)
03	Weight(in Kg)
04	Temperature(Low/High/Normal)
05	Complications(Yes/No)
06	Illness(None/Yes)
07	Patients HLA type
08	Donor’s HLA type

Table 1: Shows attributes for blood platelet transfusion

IF_THEN rule induced in the successful or unsuccessful platelet transfusion.

Example1:

IF Temperature=NORMAL **AND** Age<60
THEN
Transfusion=SUCCESFULL

Example2:

IF Complications=NO **AND** Illness=NONE
THEN
Transfusion=SUCCESSFUL

We have applied this technique here because of the ready availability of subjects with some knowledge of the domain that can provide feedback on the explanations as shown in Example1 and 2. This can be used for decision making in healthcare.

3.2 Decision Tree

Decision trees are a useful data analysis tool as they are easy to understand and can be easily transformed into rules. The main goal in a decision tree algorithm is to minimize the number of tree levels and tree nodes. It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute). Tree-based models which include classification and regression trees (CART) are the common implementation of induction modeling. Decision tree models are best suited for data mining. They are inexpensive to construct, easy to interpret, easy to integrate with database system and they have comparable or better accuracy in many applications. There are many Decision tree algorithms such as HUNTS algorithm, CART, ID3, C4.5 (a later version ID3 algorithm), SLIQ, and SPRINT. The C4.5 decision tree algorithm uses a measure taken from information theory to help with the attribute selection process.

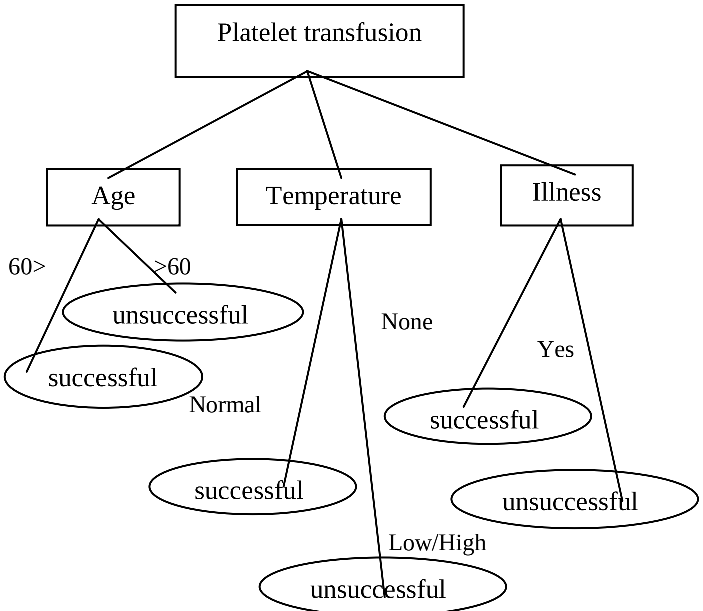

Fig. 1: A decision tree built from the data in Table-2

Sr. No.	Age	Temperature	Illness	Platelet Transfusion
01	50	Normal	None	Successful
02	61	Normal	None	Unsuccessful
03	55	High	None	Unsuccessful
04	56	Low	None	Unsuccessful
05	54	Normal	Yes	Unsuccessful
06	61	Normal	Yes	Unsuccessful
07	62	Normal	None	Unsuccessful
08	55	High	Yes	Unsuccessful
09	35	Low	Yes	Unsuccessful

Table-2: Data set used to build decision tree of Fig. 1

3.3 Artificial Neural Network (ANN):

Artificial Neural Networks are analytical techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations from previous observations after executing a process called learning from existing data. Neural networks are known to produce highly accurate results in medical applications, can lead to appropriate decisions. It has been applied within the medical domain for clinical diagnosis, image analysis and interpretation. In this proposal we realized the classification model with back propagation, which is the most popular neural network learning algorithm.

A change in the neuronal signal function and the layering of fields of neurons requires the derivation of a new learning algorithm. Fortunately, the idea of gradient descent can be put to use again in such a derivation, and the resulting algorithm is the back-propagation algorithm. The back-propagation algorithm is popular for its simplicity of implementation and its ability to quickly generate networks that have the capability to generalize.

3.3.1 Outline of Learning Procedure

The basic procedure of gradient descent based learning algorithm is outlined as follows:

1. Select a pattern X_k from the training set P, and present it to the network.
2. Compute activations and signals of input, hidden and output neurons in that sequence.
3. Compute the error over the output neurons by comparing the generated outputs with the desired outputs.
4. Use the error calculated in Step3 to compute the change in the hidden to output layer weights, and the change in the input to hidden layer weights (including all bias weights), such that a global error measure gets reduced.
5. Update all weights of the network in accordance with the changes computed in Step 4.
6. Repeat Steps 1 through 5 until the global error falls below a predefined threshold.

3.3.2 Backpropagation Learning Algorithm

Given are P training pairs

$$\{z_1, d_1, z_2, d_2, ..., z_p, d_p\},$$

Where z_i is $(I \times 1)$, d_i is $(K \times 1)$, and $i = 1,2,...,P$. Note that the $I'th$ component of each z_i is of value -1, since input vectors have been augmented. Size J-1 of the hidden layer having output y is selected. Note that the $J'th$ component of each y is of value -1, since hidden layer outputs have also been augmented; y is $(J \times 1)$ and o is $(K \times 1)$.

Algorithm:

Step1: $\eta > 0$, $E_{\max}$ chosen.

Weights W and V are initialized at small random values; W is $(K \times J)$, V is $(J \times I)$.

$$q \leftarrow 1, p \leftarrow 1, E \leftarrow 0$$

Step2: Training step starts here. Input is presented and the layers' output is computed.

$$z \leftarrow z_p, d \leftarrow d_p$$

$$y_j \leftarrow -f(v_j^t z), \quad \text{for } j = 1,2,..., J$$

where v_j is the $j'th$ row of V, and

$$o_k \leftarrow f(w_k^t y), \quad \text{for } k = 1,2,..., K$$

where w_k is the $k'th$ row of W.

Step3: Error value is computed as:

$$E \leftarrow \frac{1}{2}(d_k - o_k)^2 + E, \quad \text{for } k = 1,2,..., K$$

Step4: Error signal vectors δ_o and δ_y of both layers are computed.

where δ_o is $(K \times 1)$ and δ_y is $(J \times 1)$.

The error signal terms of the output layer in this step are

$$\delta_{ok} = \frac{1}{2}(d_k - o_k)(1 - o_k^2), \quad \text{for } k = 1,2,..., K$$

The error signal terms of the hidden layer in this step are

$$\delta_{yj} = \frac{1}{2}\left(1 - y_j^2\right)\sum_{k=1}^{K}\delta_{ok} w_{kj}, \quad \text{for } j = 1,2,..., J$$

Step5: Output layer weights are adjusted:

$$w_{kj} \leftarrow w_{kj} + \eta\,\delta_{ok} y_j, \quad \text{for } k = 1,2,..., K \text{ and } j = 1,2,..., J$$

Step6: Hidden layer weights are adjusted:

$$v_{ji} \leftarrow v_{ji} + \eta\,\delta_{yj} z_i, \quad \text{for } j = 1,2,..., J \text{ and } i = 1,2,..., I$$

Step7: If $p < P$ then $p \leftarrow p + 1, q \leftarrow q + 1$, and go to Step 2; otherwise go to Step 8.

Step8: The training cycle is completed.

If $E < E_{\max}$ terminate the training session. Output weights are W, V, q, and E.

If $E > E_{\max}$, then $E \leftarrow 0, p \leftarrow 1$, and initiate the new training cycle by going to Step2.

For best results, patterns should be chosen at random from the training set.

3.3.3 Bayes' Theorem

As we have discussed in Back-Propagation Learning Algorithm, feedforward neural network outputs can be shown to find proper interpretation of conventional statistical pattern recognition concepts. In this section, we explore important statistical pattern recognition concepts and their connection to standard feedforward neural networks. The primary task of any pattern recognition machine is to appropriately classify patterns into one of various classes which may or may not be known in advance. More often than not, the classes in consideration will overlap with one another. When classes overlap, the all important issue is to find an optimal placement of the discriminant function so as to minimize the number of misclassifications on the given data set, and simultaneously minimize the probability of misclassification on unseen patterns. The problem at hand is to design a rule for pattern classification that minimizes the probability of misclassification on unseen patterns. This problem is solved in classical statistical pattern recognition using Bayes' theorem.

3.3.3.1 Notion of Prior

The first piece of information we need to know is the prior probability of any pattern belonging to a class. The prior probability $P(\ell_k)$ of a pattern belonging to a class ℓ_k is measured by the fraction of patterns in that class assuming an infinite number of patterns in the training set. To be accurate, we need to talk about the infinite limit, where we assume that this fraction is the same as what we would observe had there been an infinite number of samples of data. Prior probabilities influence our decision to assign an unseen pattern to a class. In the extreme case when no other information is available-for example when we are not even allowed to see the pattern, we can at best assign a pattern to a class with the highest prior probability. Priors are like priming factors that influence subsequent decisions.

3.3.3.2 Bayes' Theorem for Continuous Variables

Often, instead of working with discrete variables, we find it expedient to work in the continuous domain. Probabilities for discrete intervals of a feature measurement are then replaced by probability density function $p(x)$ such a that the probability of obtaining a feature between two limits χ_l and χ_u is given by the area under the density function $p(x)$ between these two limits is given by

$$P(\chi_l \leq x \leq \chi_u) = \int_{\chi_l}^{\chi_u} p(x)dx$$

Then we can re-express Bayes' Theorem in terms of a continuous variable as

$$P(\ell_k \mid x) = \frac{p(x \mid \ell_k)P(\ell_k)}{p(x)}$$

where $p(x \mid \ell_k)$ is the class conditional density function and $p(x)$ is the

unconditional density function given by

$$p(x) = \sum_{k=1}^{C} p(x \mid \ell_k)P(\ell_k)$$

where the summation is over all C classes.

4. Conclusion and Future Work

The primary focus of this research is the development of a system that is essential for the timely analysis of huge medical data sets. The traditional manual data analysis has become insufficient and methods for efficient computer assisted analysis indispensable. This technique will be applied to the blood platelet transfusion database maintained in the Maxcare Hospital.

Classification method can also be applied on Digital mammography images (for tumor detection in breast cancer**)** to predict a class of categories (normal, benign or malign). The concept of Classification method can also applied in the study of Diabetes. Healthcare administers would like to know how to improve outcomes as much as possible. Although, present proposal will be develop using Classification rule such as Rule induction, Decision tree, Back-propagation algorithm, and Bayes' theorem, we will extend this project by using other algorithm like nearest neighbor, Bayesian classification.

The research makes use of a data mining tool, Clementine, so as to apply Decision Trees technique. We feed it with data extracted from real-life cases taken from specialized Cancer Institutes. Relevant medical cases details such as patient medical history and diagnosis are analyzed, classified, and clustered in order to improve the disease management. The research makes use of a data mining tool, Clementine, so as to apply Decision Trees technique. We feed it with data extracted from real-life cases taken from specialized Cancer Institutes. Relevant medical cases details such as patient medical history and diagnosis are analyzed, classified, and clustered in order to improve the disease management.

3. REFERENCES

[1] Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.

[2] Miller, A., B. Blott and T. Hames, 1992. Review of neural network applications in medical imaging and signal processing. Med. Biol. Engg. Comp., 30: 449-464.

[3] Miller, A., 1993. The application of neural networks to imaging and signal processing in astronomy and medicine. Ph.D. Thesis, Faculty of Science, Department of Physics.

[4] Weinstein, J., K. Kohn and M. Grever *et al*., 1992. Neural computing in cancer drug development: Predicting mechanism of action. Science, 258: 447-451.

[5] Romeo, M., F. Burden, M. Quinn, B. Wood and D. McNaughton, 1998. Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer.

[6] Zaiane, Osmar R, A. Maria-luiza and A. Coman, 2001. Application of data mining techniques for medical image classification.

[7] Sultan Ahmed, A. Hegam, "Classical and incremental classification in data mining process " International Journal of Comp. Sci. and Network security, Vol. 7, 2007.

[8] J.Han and M .Kamber, "Data Mining: concept and techniques "first edition, Harcoart India private Limited. 2001.

[9] M. Kantaradzic, Data Mining: Concepts, Models, Methods, and Algorithms, IEEE Press and John Wiley, New York, NY. 2003.

[10] Harleen and S. K. Wasan, "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science: 194-200, ISSN 1549-2006.