Journal of Computer Science 6 (5): 548-552, 2010
ISSN 1549-3636

Application of CART Algorithm in Blood Donors Classification

T. Santhanam and Shyam Sundaram
PG and Research Department of Computer Science, DG Vaishnav College,
Chennai-600106, Tamil Nadu, India

Abstract: Problem statement: This study used data mining modeling techniques to examine the blood donor classification. The availability of blood in blood banks is a critical and important aspect in a healthcare system. Blood banks (in the developing countries context) are typically based on a healthy person voluntarily donating blood and is used for transfusions or made into medications. The ability to identify regular blood donors will enable blood banks and voluntary organizations to plan systematically for organizing blood donation camps in an effective manner. **Approach:** Identify the blood donation behavior using the classification algorithms of data mining. The analysis had been carried out using a standard blood transfusion dataset and using the CART decision tree algorithm implemented in Weka. **Results:** Numerical experimental results on the UCI ML blood transfusion data with the enhancements helped to identify donor classification. **Conclusion:** The CART derived model along with the extended definition for identifying regular voluntary donors provided a good classification accuracy based model.

Key words: Blood donor, data mining, classification algorithms, decision trees

INTRODUCTION

In the developed world, most blood donors are unpaid volunteers who give blood for a community supply. In developing countries, established supplies are limited and donors usually give blood when family or friends need a transfusion. Many donors donate as an act of charity, but some are paid and in a few cases there are incentives other than money such as paid time off from work. A donor can also have blood drawn for their own future use.

Recruitment of safe donors is a challenging task. It is necessary that people realize that blood donation is their responsibility. No blood bank, hospital or Government can sustain health care without adequate blood from such donors and blood donor organizations play a very crucial role in this endeavor.

Potential donors are evaluated to ensure their blood is safe to use. Donated blood undergoes screening includes testing for diseases that can be transmitted by a blood transfusion, including HIV and viral hepatitis. The donor is also asked about medical history and given a short physical examination to make sure that the donation is not hazardous to his or her health. The frequency of blood donation can vary from days to months and also depend on the laws of the land.

The amount of blood drawn and the methods vary, but a typical donation is 500 c.c of whole blood. The collection is done either manually or with automated equipment.

Blood donor ship: A donation is when a donor gives blood for storage at a blood bank for transfusion to an unknown recipient. A donation is directed when a person, often a family member, donates blood for transfusion to a specific individual. An event where donors come to give blood is sometimes called a blood drive or a blood donor session. These can occur at a blood bank but they are often set up at a location in the community such as a shopping center, workplace, or school. In many developing countries like India there is an increased push towards mobile blood donation camps. Voluntary blood donation programme is the foundation for safe and quality blood transfusion service as the blood collection from voluntary blood donors is considered to be the safest. In order to augment voluntary blood donation in the India specific context, there was a felt need to develop an operational guideline (Government of India, 2007) which can provide all the necessary information on recruitment and retention of voluntary blood donors and guide organizations for this important activity.

Corresponding Author: T. Santhanam, PG and Research Department of Computer Science, DG Vaishnav College, Chennai-600106, Tamil Nadu, India

Categories of blood donors: The following are some of the categories of blood donor (Government of India, 2007). A New Voluntary Donor (NVD) is a voluntary non-remunerated blood donor who has never donated blood before. A Lapsed Voluntary Donor (LVD) is a voluntary non-remunerated blood donor who has given blood in the past but does not fulfill the criteria for a regular donor. A Regular Voluntary Donor (RVD) is a voluntary non-remunerated blood donor who donates blood on a regular basis without any break for a longer duration between two donations.

Review of literature: Masser *et al.* (2009) have developed a framework that helps determining the predictors of the intentions and behavior of established blood donors. Ferguson and Chandler (2005) have used qualitative studies to demonstrate that blood donors describe their behavior using Trans Theoretical Model (TTM). The government of India 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 (Government of India, 2007). Mohamed Mostafa (2009) uses 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. Santhanam and Sundaram (2009) has applied classification algorithms to blood donor ship data and discusses how the classification can be enhanced with a standard dataset.

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. Another study to understand blood donor behavior was undertaken by Schlumpf *et al.* (2007). This study self-administered questionnaire was completed in 2003 by 7905 current donors. With data mining methods, all factors measured by the survey were ranked as possible predictors of actual return within 12 months. Significant factors were analyzed with logistic regression to determine predictors of intention and of actual return.

MATERIALS AND METHODS

About the dataset: The blood transfusion dataset (UCI ML repository) (Asuncion and Newman, 2007) is based on donor database of Blood Transfusion Service Center in Hsin-Chu City in Taiwan. The center passes their blood transfusion service bus to one university in Hsin-Chu City to gather blood donated about every three months. This dataset provided by Yeh *et al.* (2008). The data set consists of 748 donors at random from the donor database. These 748 donor data, each one included R (Recency-months since last donation), F (Frequency-total number of donation), M (Monetary - total blood donated in c.c.), T (Time-months since first donation) and a binary variable representing whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood). The people who have donated blood in 2007 accounts for only 24% in the dataset. There is an imbalance in both the blood donated in 2007 and also found in the extended nominal class of regular voluntary donor. It must be noted that the extended nominal class of the regular voluntary donor has been extended based on the current dataset, it factors not just the blood donation in 2007 but also the other attributes.

Data mining tools: There are a number of high quality commercial and open source tools for data mining. In this research Weka (Ian Witten and Eibe Frank, 2005) has been used from the perspectives of direct core usage. This serves as a powerful core tool that allows the ability to load, preprocess and visualize data and also perform standard DM algorithms with sufficient parameterization. These algorithms can either be applied directly to a dataset or called from custom Java code.

Decision trees models: Decision tree learning is a common method used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf (Dunham, 2003). Some of the key advantages of using decision trees are the ease of use and overall efficiency. Rules can be derived that are easy to interpret.

CART algorithm: In this study the use of Classification and Regression Trees (CART) classification algorithm has been attempted (Breiman *et al.*, 1984). Classification tree analysis is

when the predicted outcome is the class to which the data belongs. Regression tree analysis is when the predicted outcome can be considered a real number. CART has been applied in a number of applications in the medical domain (Paul Harper, 2005; Stothers *et al*., 2009; Jin *et al*., 2004).

One of the advantages of using classification trees is their ability to provide easy to understand classification rules. Each node of a classification tree is a rule. The only exception to this would be in cases where the tree is very large and in such cases there may need to be a more specific focus on pruning required to optimize the tree size. Trees are easy off the shelf classifiers that require no variable transformation.

CART builds the tree by recursively splitting the variable space based on the impurity of the variables to determine the split till the termination condition is met. The gini impurity determines how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. Using this algorithm for the classification of a regular voluntary donor, the decision tree in Fig. 1 has been generated. The following is a pseudo procedure (Soman *et al*., 2006):

Step 1: Start with root node (t = 1)
Step 2: Search for a split s* among the set if all possible candidates s that gives the purest decrease in impurity.
Step 3: Split node 1 (t = 1) into two nodes (t = 2, t = 3) using the split s*.
Step 4: Repeat the split search process (t = 2, t = 3) as indicated in steps 1-3 until the tree growing the tree growing rules are met.

RESULTS AND DISCUSSION

The original dataset has been extended with an extended nominal class of Regular Voluntary Donor (RVD) creates based on the classification suggested in Table 1. This allows us to possible create a classification model with a combination of input parameters and most importantly considered in the blood donor classification in India. The creation of creation of this in the dataset allows for models to be built agnostic to pure assumption of donor ship drives attributes. The CART algorithm implemented in Weka uses minimal cost-complexity pruning. The following Fig. 1 shows the classification tree generated.

Table 1: RVD classification

Attribute	Conditional	Value
Recency	<=	6 Months
Frequency	>=	4 Months
Monetary	>=	2000 c.c
Time	>	24 Months

It should be noted that pruning the tree results in making the tree shorter and simpler and at the same time, one has to guard against the possibility of over fitting of data. In this case the RVD based classification tree has reduced the tree complexity of the number of leaf nodes by 7 in comparison to the donated blood attribute based tree.

Figure 2 show the distribution of donated blood in 2007 (X axis) and the classification of the donor based on our extended nominal class of RVD (Y axis).

Table 2 depicts the confusion matrix generated using the RVD extended class.

With the application of this model a dashboard can be provided by blood transfusion management systems that will facilitate a real time status of the types of donors. Figure 3 show a prototype that was created based on the analysis in the study. Similar to RVD we can also adopt other donor ship type for monitoring.

The CART decision tree in Fig. 1 shows a simple structure with the root (frequency) and leaf node recency. The other key observations being that the monetary attribute does not figure in the decision tree.

```
Frequency<18.5: FALSE (727.0/0.0)
Frequency>= 18.5
    |Recency<8.5: TRUE (17.0/0.0)
    |Recency>= 8.5: FALSE (4.0/0.0)
```

Fig. 1: CART classification tree

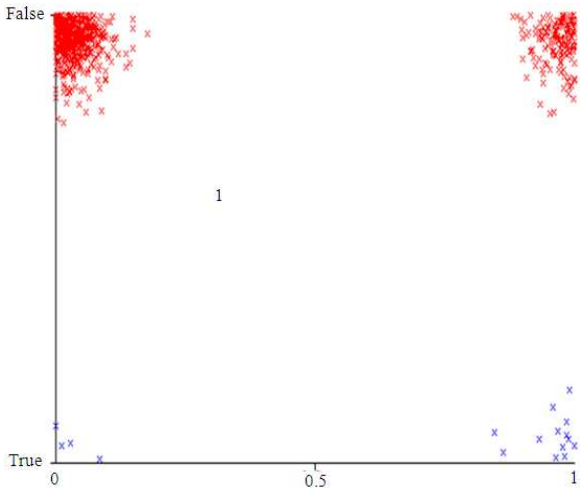

Fig. 2: Donated blood in 2007(X) Vs RVD (Y)

Table 2: RVD confusion matrix

	TP	FP	Precision	Recall
Class not RVD	1.00	0.059	0.99	1.00
Class RVD	0.94	0.000	1.00	0.94

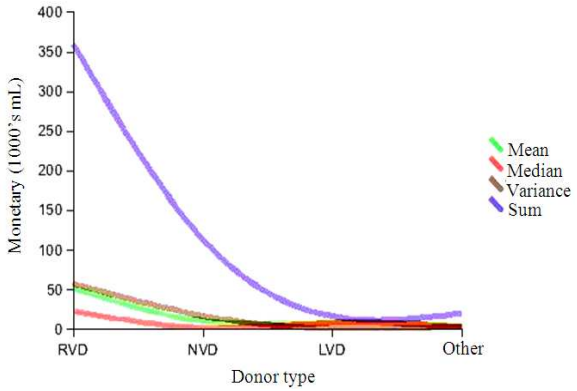

Fig. 3: Donor pool status dashboard

This is in conjunction with the study (Schlumpf *et al*., 2007) which identified higher prior donation frequency as predictor.

The CART derived model has a good classification accuracy detailed in Table 2. This gives the details in terms of a good recall and precision rates.

CONCLUSION

In addition the model needs to be customized to geographic/demographic specific attributes. The dashboard is an ample evidence of the capability of the concepts discussed and can be extended to real-time systems. This can be extended as well to the existing blood banks systems.

The key benefit of the creating an extended RVD definition based on the donor definition (along with the application of CART) provides a standard model to determine the donor behavior and provides the capability to build a classification model. This additional nominal class can be easily computed based on the definition.

This demonstrates the ability to create a dashboard to manage the status of blood donors that will help blood transfusion centers manage the blood bank.

Future work will be focused on using the other classification algorithms of data mining. It is a known fact that the performance of an algorithm is dependent on the domain and the type of the data set. Hence, the usage of other classification algorithms like machine learning will be explored in future.

REFERENCES

Asuncion, A. and D.J. Newman, 2007. UCI repository of machine learning databases. http://www.ics.uci.edu/_mlearn/MLRepository.html

Bharucha, Z.S., 2005. Donor management in South-East Asia region (SEAR). Dev. Biologic. Standard., 120: 145-53. http://www.biomedsearch.com/nih/Donormanagement-in-South-East/16050168.html

Breiman, L., J. Friedman, R.A. Olshen and C.J. Stone, 1984. Classification and Regression Trees. Wadsworth, ISBN: 0534980546, pp: 358.

Chaudhary, R., 2009. Blood transfusion services in India. Transfusion Apheresis Sci., 40: 1-72. DOI: 1016/j.transci.2008.11.012

Dunham, M.H., 2003. Data Mining Introductory and Advanced Topics. Prentice Hall of India, New Delhi, ISBN: 81-7758-880-X.

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

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

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.

Jin, H., Y. Lu, S.T. Harris, M. Dennis Black and K. Stone *et al*., 2004.Classification algorithms for hip fracture prediction based on recursive partitioning methods. Med. Dec. Mak., 24: 386-398. DOI: 10.1177/0272989X04267009

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

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

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

Santhanam, T. and S. Sundaram, 2009. Classification of blood donors using data mining. Proceedings of the Semantic E-Business and Enterprise Computing (SEEC'09), Kingston University, London, pp: 145-147.

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

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.

Stothers, L., R. Guevaraa and A. Macna, 2009. Classification of male lower urinary tract symptoms using mathematical modeling and a regression tree algorithm of noninvasive near-infrared spectroscopy parameters. Eur. Urol., 57: 179-362. DOI: 10.1016/j.eururo.2009.05.004

Yeh, I.C., K.J. Yang, Ting and T. Ming, 2008. Knowledge discovery on RFM model using Bernoulli sequence. Expert Syst. Appli., 36: 5866-587. DOI: 10.1016/j.eswa.2008.07.018