Journal of Theoretical and Applied Information Technology
31st August 2011. Vol. 30 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A COMPARISON OF BLOOD DONOR CLASSIFICATION DATA MINING MODELS

[1]SHYAM SUNDARAM, [2]SANTHANAM T

[1]Research Scholar ,PG & Research, Dept. of Computer Science, DG Vaishnav College,

Chennai 600106, Tamil Nadu,India

[2]Head, PG & Research, Dept. of Computer Science, DG Vaishnav College,

Chennai 600106, Tamil Nadu,India

E-mail: santhanam_dgvc@yahoo.com

ABSTRACT

This study uses decision trees derived from data mining modeling techniques to examine the blood donor classification. The primary goal of this classification model is provide the capability to determine voluntary blood donorship based on blood donation patterns. In specific a comparison is made of two models (one based on a specific donation drive versus the regular voluntary donor patterns) based on a standard data set for blood transfusion. The enhancement of attributes that help enable the determination of voluntary donorship is also a suggested value addition The paper discusses comparison of donorship models using the classification algorithms of data mining which enable representation as decision trees. The analysis provides insight into the development of donor classification which enable blood banks to determine the kinds of donor profiles and manage blood donorship related activities like recruitment and campaigns for blood donations.

Keywords: *Blood Bank, Blood Transfusion, Blood Donor, Data Mining, Classification Algorithms*

1. INTRODUCTION

In the developed world, most blood donors are unpaid volunteers who give blood for a community supply. In poorer countries, established supplies are limited and donors usually give blood when family or friends need a transfusion. Blood donation service is has a number of interdependent operations such as donor registration, donor screening/evaluation, blood collection, blood screening, inventory management and blood dissemination. A donor can also have blood drawn for their own future use. How often a donor can give varies from days to months based on what he or she donates and the laws of the country where the donation takes place. The ability to develop models that enable classification of blood donors will enhance the ability to better manage the demand for blood products and with effective campaigns in the recruitment of voluntary blood donors.

This paper is organized as follows. Section two deals with the introduction to blood donorship and section three explains about the analysis done using classification algorithms and their results and conclusion is given in the final section.

2. BLOOD DONORSHIP

An donation is when a donor gives blood for storage at a blood bank for transfusion to an unknown recipient. 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, school, or house of worship. Voluntary Blood Donation programme is the foundation for safe and quality Blood Transfusion Service as the blood collection from Voluntary non-remunerated blood donors is considered to be the safest. In order to augment Voluntary Blood Donation in developing countries like India is based on a framework and operational guide for organizations for this important activity[2].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.1 RELEVANT PEER RESEARCH

Santhanam et al[1] extended the nominal definition based on a standard dataset to derive a CART based decision tree model based on standard donorship model derived from a standard blood transfusion dataset. This paper also provides a good review of peer research in the domain of data mining relevant to blood donor classification. The original dataset was extended to determine a regular voluntary donor. This extended dataset adopts the framework defined for regular voluntary donorship as defined by a standard reference organization. Using this extended dataset the study dwells into the classification models of a RVD. The findings suggest a mechanism of identifying regular voluntary donors. Masser et al[3] have developed a framework that help determining the predictors of the intentions and behavior of established blood donors. Ferguson et al[4] have used qualitative studies to demonstrate that blood donors describe their behaviour using TTM(Trans Theoretical Model). Mohamedl[5] uses intelligent modeling techniques to examine the effect of various demographic, cognitive and psychographic factors on blood donation in Egypt. This research used a neural network model based on variable sets such are sex, age, educational level, altruistic values, perceived risks of blood donation in the modeling. Another study to understand blood donor behavior was undertaken by Schlumpf et al [6]. 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.

3. BLOOD TRANSFUSION DATASET ANALYSIS

3.1 ABOUT THE DATASET

The blood transfusion dataset (taken from the UCI ML repository)[7] 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 is Prof. I-Cheng Yeh[8].

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). There is an imbalance in that the people who have donated blood in 2007 accounts for only 24% in the dataset. The dataset has been extended to enable classification of RVD.

3.2 ANALYSIS

The analysis has been done using the WEKA [9] tool with the development of classification models on this dataset. Applying the CART (Classification and Regression Trees) classification algorithm [10][11]. 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 analysis is used to refer to both of the above procedures. The donated blood in 2007 (DB2K7) attribute is converted to nominal values. The resulting decision tree with the application of the CART model is depicted in figure 1. This decision tree is also more complex from the levels of nesting. The analysis compares this model with RVD model[1]. The RVD model has a much simpler decision tree with recency and frequency as the key indicators. The key aspect that differentiates the two models is the notion of using the specific blood donation drive as an indicator or a measure of the RVD.The comparison was carried out with the extended RVD based model[1] with our nominal class(DB2K7) from the dataset.The figure 2 shows the comparison of the models and the improvement using the RVD model.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
Recency < 6.5
| Frequency < 4.5
| Frequency >= 4.5
| | Time < 49.5
| | | Frequency < 14.5
| | | | Time < 18.5
| | | | Time >= 18.5
| | | | | Frequency < 6.5
| | | | | Frequency >= 6.5
| | | Frequency >= 14.5
| | Time >= 49.5
| | | Frequency < 12.5
| | | Frequency >= 12.5
| | | | Frequency < 25.0
| | | | | Time < 57.5
| | | | | Time >= 57.5
| | | | Frequency >= 25.0
Recency >= 6.5
```

Figure 1. CART Classification Tree

The following table 1 contains the confusion matrix based on our model.

TABLE 1: **DB2K7** CONFUSION MATRIX

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Class 0 (not DB2k7)	0.91	0.69	0.81	0.91	0.86	0.69
Class 1 (DB2k7)	0.31	0.09	0.53	0.31	0.39	0.69
Weighted Average	0.77	0.55	0.74	0.77	0.75	0.69

This suggests that the approach taken by the RVD based classification has a better result in the context of developing a better classification profile as seen in the following table 2 of the RVD based model. The RVD classification has a better recall and precision capability over the DB2K7 classification. The RMS error of the RVD is lower than the DB2k7 model.

TABLE 2: **RVD** CONFUSION MATRIX

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Class 0 (not RVD)	1	0.06	1	1	1	0.97
Class 1 (RVD)	0.94	0	1	0.94	0.97	0.97
Weighted Average	1	0.06	1	1	1	0.97

4. CONCLUSION

The model using the DB2K7 model has a 77% accuracy of correct classification in comparison with the RVD based model which has a with a 10 fold classification 99.9%. The analysis show the determination of a RVD has simpler decision tree [1]. Also the better classification accuracy in terms of improved true positive,precision and recall rates of the RVD model over the DB2K7 model suggests it to be a better classification model. Figure 3 shows the comparative comparison of these models in a graphical manner.

Future work will be focused on further refining this model for improved classification of regular blood donors. Also the implementation of this classification model to core blood donor management systems will be looked at. Applications of this model to the management of blood donors and enable the effective campaign planning will be looked at. Further research can also look at fuzzy models in the context of this domain.

REFRENCES:

[1] Shyam Sundaram and T. Santhanam, "Classification of Blood Donors using Data Mining", *Proceedings of the Semantic E-Business and Enterprise Computing (SEEC'09)* , 2009, pp. 145-147.

[2] National AIDS Control Organization Government of India Ministry of Health and Family Welfare, "Government of India Ministry of Health and Family Welfare", *Voluntary Blood Donation Programme* , 2007.

[3] Masser, M. Barbara, White, M. Katherine, Hyde, Melissa K., Terry, Deborah J.,Robinson and G. Natalie, "Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned behavior model", Transfusion, Vol. 49, No. 2, 2009, pp. 320-329.

[4] Eamonn Ferguson and Susie Chandler. "A stage model of blood donor behaviour: Assessing volunteer behaviour", Journal of Health Psychology,, Vol. 10, No. 3, 2005, pp. 359-372.

[5] Mohamed M. Mostafa. "A Profiling blood donors in Egypt: A neural network analysis ", Expert Systems with Applications, Vol. 36, No. 3, 2009, pp. 5031-5038.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[6] Schlumpf, K.S., S.A. Glynn, G.B. Schreiber, D.J. Wright and W. Randolph Steele et al."Factors influencing donor return", Transfusion,Vol. 48 ,2007, pp. 264-72.

[7] A. Asuncion and D.J. Newman. "UCI repository of machine learning databases", www.ics.uci.edu/_mlearn/MLRepository.html, , 2007.

[8] I.C. Yeh , K.J. Yang, Ting and T. Ming."Knowledge discovery on RFM model using bernoulli sequence", Expert Systems with Applications, Vol. 36, No. 5, 2009, pp. 5866-5871.

[9] Ian H. Witten and Eibe Frank."Data Mining: Practical machine learning tools and techniques",Morgan Kaufmann, San Francisco, 2005.

[10] L. Breiman, J. Friedman, R. A. Olshen and C. J. Stone."Classification and regression trees",Wadsworth, 1984.

[11] K.P Soman, Shyam Diwakar, and V. Ajay."Insight into Data Mining – Theory and Practice",Prentice Hall Of India, New Delhi, 2006.

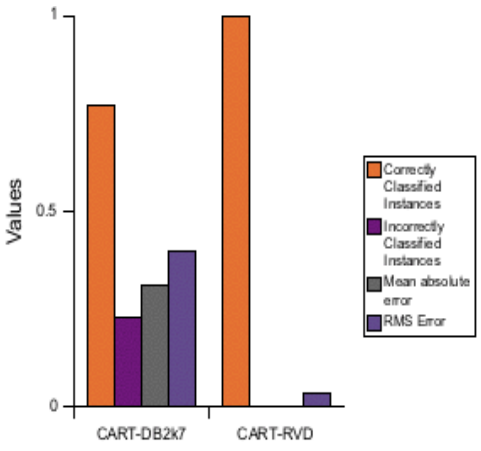

FIGURE 2. COMPARISON OF MODELS

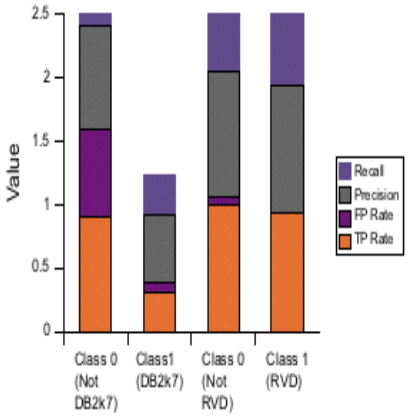

FIGURE 3. MODEL DB2K7 VS. RVD