A comparative study of classification algorithm on blood transfusion

S.Asha Rani¹, Dr. S. Hari Ganesh²

¹Department of Computer science, Bishop Heber College, Trichirapalli, India, ²Department of Computer science, Bishop Heber College, Trichirapalli, India.

Email: bluewhirlpool999@gmail.com¹, hariganesh17@gmail.com²

ABSTRACT

Requirement of blood is increasing gradually due to accidents, surgeries etc. Blood transfusion play an important role in healthcare. Blood donor prediction provide essential details to medical professionals to increase the number of voluntary blood donors in future. Data mining is the process of collecting relevant data from enormous amount of data. This paper focus on analyzing the efficiency of different classification algorithm in data mining using blood transfusion dataset

KEYWORDS: Data mining, Blood Bank, Blood transfusion dataset, Blood donor, classification algorithms

1. INTRODUCTION

Many developed countries, blood donors are mostly regular and volunteer donors and new volunteer donors are also increased gradually day by day. In developing countries, blood donors are paid and not volunteer. They are mostly donate their blood when their relations or friends are in need of blood.

2. RELATED WORK

Arvind Sharma [C], the main purpose of his research work is to predicting the blood donors through their age and blood group. In this paper, measure the accuracy of blood donors prediction using real time data set. WEKA tool used in this research. In conclusion, he said that blood donor’s behaviors and attitude predicted using various data mining techniques.

Vikram Singh [D], in this research work, they explain how they acquire interactive KDD process in less effort with low time consumption by interaction between human and computer system. Blood transfusion dataset from UCI repository is used for this research work.

Shyam Sundaram [E], the main purpose of this research work is to test whether the particular person is RVD (Regular Voluntary Donor) or not.

Alaa Hamouda et al. [F], RBC counted automatically using image processing techniques. In this RBC counting, learning through decision tree techniques shows more accuracy than other techniques.

Shyam Sundaram et al. [G], predict the regular voluntary donors form various locations in India. Compare various locations in India and identify voluntary donors from which location. Blood transfusion dataset from UCI repository used geo location attribute is added randomly with this dataset.

Ivana D. Radojevic et al. [H], water quality analyzed by using coliforms presented in reservoirs. Classification and clustering techniques used for total coliform analysis.

Wen-Chen Lee et al. [I], created the system using classification and clustering techniques for predicting blood donors behaviours. Understanding blood donor’s behaviours increase voluntary blood donor’s rate gradually. Classification techniques such as Naive bayes, NB tree are used for blood donor behavior prediction.

3. DATA SET DESCRIPTION

DB2K7 (Donated Blood in 2007) and RVD (Regular Voluntary Donor) are two models used for testing. RVD model show better accuracy than DB2K7.
Dataset used for testing in this paper is taken from UCI repository. The dataset details collected from 748 donors from donor database. This dataset consist of five attributes.

- Recency (months since last donation)
- Frequency (Total number of donation)
- Monetary (total blood donated in c.c)
- Time (months since first donation)
- Whether the person donated blood in 2007 (Donated blood in 2007: 1 otherwise 0)

**Table 1: Attributes used in dataset**

<table>
<thead>
<tr>
<th>Sl.No</th>
<th>Attributes</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Recency</td>
<td>Numeric</td>
</tr>
<tr>
<td>2</td>
<td>Frequency</td>
<td>Numeric</td>
</tr>
<tr>
<td>3</td>
<td>Monetary</td>
<td>Numeric</td>
</tr>
<tr>
<td>4</td>
<td>Time</td>
<td>Numeric</td>
</tr>
<tr>
<td>5</td>
<td>Whether he/she</td>
<td>Nominal</td>
</tr>
<tr>
<td></td>
<td>donated blood in 2007 (0 or 1)</td>
<td></td>
</tr>
</tbody>
</table>

4. TECHNIQUES

We apply different classification techniques to diabetes dataset and the error results obtained is tabulated in table given below.

**Table 2: Algorithm Comparison based on time taken**

<table>
<thead>
<tr>
<th>Classification Technique</th>
<th>Time taken to Build (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>0.03</td>
</tr>
<tr>
<td>J48</td>
<td>0.01</td>
</tr>
<tr>
<td>Random Tree</td>
<td>0.01</td>
</tr>
</tbody>
</table>

The graphical comparison of various classification algorithms can be given as below.
The Naive Bayesian classifier is based on Bayes’ theorem with independence assumptions between predictors. It is particularly suited for large datasets containing enormous amounts of inputs. Naive Bayes classifiers use all attributes in the dataset and analyze them individually and make it independent of each other. Despite its simplicity, it outperforms many complex algorithms.

Fig 4.3: Naive Bayes

**J48:**

J48 decision tree is the implementation of algorithm ID3 (Iterative Dichotomiser 3). Decision tree is the predictive machine learning model. J48 builds decision trees based on the set of training data.

Decision tree is the most useful and important in classification. Each attribute in the dataset is used. The fundamental idea used in decision tree is to divide the data into smaller subsets for easy classification. Decision trees always provide an easy way of understanding the classification by splitting the data. J48 accepts both continuous and discrete attributes for classification.

Fig 4.4: J48

**RANDOM TREE:**

Random tree that both work with classification and regression problems. It constructs multiple decision trees randomly. While building each tree, this algorithm gets remaining features randomly at each node without any checks such as information gain, gini index, etc. It is a collection of tree predictors called forest. This tree algorithm stops it growth when no more node to split or node is empty. This algorithm does not require any accuracy estimation procedures like Bootstrap.

Fig 4.5: Random tree

5. DISCUSSION
The final attribute used to find whether person donate blood in 2007 or not.

Table 4: confusion matrix

<table>
<thead>
<tr>
<th>TP</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP</td>
<td>FN</td>
</tr>
</tbody>
</table>

TP-True Positive
FP-False Positive
TN-True Negative
FN-False Negative

748 instances used in this dataset

Table 5: Confusion matrix of Random tree:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>Classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>568</td>
<td>2</td>
<td>a=0</td>
</tr>
<tr>
<td>49</td>
<td>129</td>
<td>b=1</td>
</tr>
</tbody>
</table>

a = not donate blood in 2007
b = donate blood in 2007

True positive:
Person not donates blood in 2007 correctly identified as not donated in 2007.
False positive:
Persons whose are not donate blood in 2007 incorrectly identified as donate blood in 2007.
True negative:
Persons donate blood in 2007 correctly identified as donate blood in 2007.
False negative:
Donated persons incorrectly identified as not donate blood in 2007.

Table 6: Confusion matrix of J48:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>Classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>515</td>
<td>55</td>
<td>a=0</td>
</tr>
<tr>
<td>88</td>
<td>90</td>
<td>b=1</td>
</tr>
</tbody>
</table>

Table 7: Confusion matrix of Naive Bayes:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>Classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>529</td>
<td>41</td>
<td>a=0</td>
</tr>
<tr>
<td>146</td>
<td>32</td>
<td>b=1</td>
</tr>
</tbody>
</table>

6. CONCLUSION

This paper is used to comparison of various algorithms in classification and to measure the accuracy with low time consuming for testing.
The algorithm Random tree has shows 93.18% accuracy within short duration when compared with other algorithms in classification. This comparison among algorithms is used to provide proper utilization of best algorithm to give correct solution to problems in short time. This comparison should be used in further proceedings in scientific research and in prediction of blood donors.

7. REFERENCES


[C] Arvind Sharma and P.C.Gupta, 'Predicting the number of blood donors through their age and blood group by using data mining tool', International journal of communication and computer technologies (IJCOTS), vol-1, and issue: 02, September 2012.


[H] 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