Comparative study of Data Mining Approaches for blood platelet transfusion

Dr.Hari Ganesh S[1], Vanitha. K[2]
[1]*Asst. Professor,* [2]*M.Phil. Scholar,*
Department of Computer Applications,
Bishop Heber College (Autonomous),
Trichirappalli-620 017

ABSTRACT

Data mining is the extracting knowledge from the large amount of data. The four types of data mining techniques are classification, clustering, regression, and association rule. Data mining are used to many applications. There are medical, survillence, fraud detection and marketing. Platelet is the circulates of blood in our body when they blood vassals are damaged and recognized. Platelets are called thrombocyte and it is count low and normal. Platelets are reduced of our body the cancer, heavy bleeding, and leukaemia diseases are affected. The goal of the project using breast cancer dataset used in classification algorithm. Which is the best accuracy is given the algorithm is analyzed.

KEYWORDS: Data mining, Blood Platelet Transfusion, Naive Bayesian,
Decision Table, J48.

1.1 INTRODUCTION

Data mining is the extracting knowledge from the large amount of data. Data mining is defined the hidden data in the database. Data mining provides automatic pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods. Data mining in two important process, validation and verification .Data mining is the four important techniques classification, clustering, regression, and association rule. The classification algorithms are mainly used to the data set. Classifications are predefined groups or classes. For example, the person is guessing other person age.

1.2 CLASSIFICATION

Classification is the predefined groups or classes. Classification is the guessing the data and it require the class based on the attribute value. One of the classifications of

ISSN: 2278 – 1323

pattern recognition where an input pattern is classified in several classes based to similar to the predefined classes. Classification used for loan approval, detecting faults in industry application, classifies the financial market trends. There are different types of classification algorithm are used in the dataset. Naive Bayesian, Decision Tree, J48 algorithm are using the data set.

1.3 BLOOD PLATELET TRANSFUSION

Platelet is the very small cells that are found the blood, when blood vessels are damaged. The platelet is identified and the new blood vessels are created. Platelet is circulates of blood in our body. Platelet is known as thrombocyte. The illnesses are affected of platelets because the number of platelet in the blood is lower than normal. The illness such as cancer, leukaemia, or certain blood disorder or because of a side effect of chemotherapy treatment.

1.3.1 DATASET

The platelet is lower than our blood the illness is affected. The breast cancer dataset. The breast cancer dataset are using ten attributes and the attributes are Age, Menopause, Tumor size, Inv node, Node caps, Deg malig, Breast, Breast quad, Irradiat, Class.

1.4 NAIVE BAYESIAN

Naive Bayesian classifier is a simple probabilistic classifier**.** It is predicted class membership probabilities such as the probability such as given a tuple in particular class. Naive Bayesian classifier comparable to decision tree and selected neural network classifiers. When the large database can be applied for Bayesian classifier was high speed and accuracy. Bayesian classifier assume the effect of the attribute value and the given the independent values of the other attribute value. That the assumption is called the class conditional independence.

1.5 DECISION TABLE

The Decision table most rudimentary way of representing the output of machine learning is to make it just the same as the input .it is the decision table. For example of the decision table is using the algorithm is very accuracy of the table.

1.6 J48

The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a

ISSN: 2278 – 1323

decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for that target variable, then we terminate that branch and assign to it the target value that we have obtained.

2. DATA DESCRIPTION

The data set are taken from the breast cancer dataset from UCI repository .because the platelet count is lower than normal the illness are affected such as cancer , leukaemia and blood disorder. The dataset are consisting of ten attributes in breast cancer. The detail and description are given below,

NO	Name of the Attribute	Description
1	Age	Age of the patients
2	Menospuse	Natural change in women
3	Tumor -size	Size of the tumor of patients.
4	Inv -nodes	Inv node of patients
5	Node -caps	Lumph node capsule
6	Deg -malig	Degree of malignancy
7	Breast	Checkup of breast
8	Breast -quad	Quadrants of breast
9	Irradiat	Patient irradiat
10	Class	Class of the patient

Table 1: Attribute Table

The attribute are given the breast cancer data types .The data type are used only nominal. Nominal data type is list of predefined values.

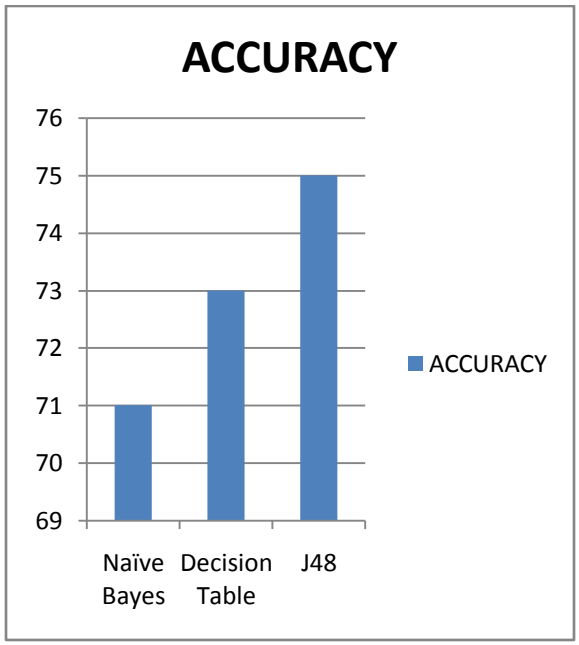
3. EXPERIMENTAL RESULTS

The given three types of algorithms like Naïve Bayesian, Decision Table, J48 and BF tree are applied on the breast cancer data set in WEKA and the performance of the algorithm are given based various factors. The performance can be obtained based on

ISSN: 2278 – 1323

the time taken to build the tree and correctly classified instances.

Fig 3.1 Performance of the Algorithms based on the accuracy level



X-Axis: Classification of Algorithms

Y-Axis: Percentage level

The dataset consists of 286 instances and they are applied as a test case in the classification algorithms. The performance of the algorithms can be known from the instances that are correctly classified. The instances which are correctly classified using the WEKA tool can be given as below,

Table 3.2 Number of instances correctly classified

Name of the Algorithm	Number of correct instance	Accuracy
Naïve Bayesian	205	71.6%
Decision Table	210	73.4%
J48	215	75.5%

4. DISCUSSION

The above three algorithms predicts the class label. The final output will be patterns which are used to find out whether the person is affected by the breast cancer or not. A Confusion Matrix is a useful visualization tool for analyzing the classifier accuracy. Structure of the confusion matrix can be given as below

Table 4.1 Structure of the Confusion Matrix

TP	**TN**
FP	**FN**

Where

- **TP** is True Positive: breast cancer patients correctly identified as the breast cancer Disease.
- **FP** is False Positive: Healthy people incorrectly identified as breast cancer Disease.

ISSN: 2278 – 1323

- **TN** is True Negative: Healthy people correctly identified as healthy.
- **FN** is False Negative: breast cancer Disease patients incorrectly identified as healthy.

The Confusion Matrix for the classification algorithms such as Naive Bayes, Decision Table and J48 can be given as follows based on the execution of the algorithm using WEKA tool.

Table 4.2 Confusion Matrix for Naïve Bayesian

168	**33**
48	**37**

Table 4.3 Confusion Matrix for Decision Table

186	**15**
61	**24**

Table 4.4 Confusion Matrix for J48

193	**8**
62	**23**

5. CONCLUSION

Data mining is the extracting knowledge from large amount of data. The classification algorithms are only used for the data set. Three type of classification algorithms are using the techniques, there are Naive Bayesian, Decision Table, J48 algorithm. The algorithm are applied for the weka tool and given the good performance and good accuracy. The J48 algorithm is given the best accuracy in 75.5% .Three classification techniques are only taken from weka tool. If you are implemented in many classification techniques are used.

6. REFERENCE

[1] P.Ramachandran, Dr.N.Girija, Dr.T.Bhuvaneswari, "Classifying Blood Donors Using Data Mining Techniques" published in IJCSET|Feb 2011|Vol 1, Issue 1, 10-13.

[2] Devchand J.Chaudhari, Mamta Ramteke Manoj G.Lade, "Data Mining in Blood Platelets Transfusion using Classification Rule" Published in IJCA.

[3] Harleen Karur and Siri Krishan Wasan "Empirical Study on Applications of Data Mining Techniques in Healthcare" published in Journal of Computer Science 2(2):194-200, 2006. ISSN 1549-3636

[4] John N.Weinstein, Kurt W.Kohn, "Neural Computing in Cancer Drug

ISSN: 2278 – 1323

Development: Predicting Mechanism of Action" Vol .258

[5] Ankit Bhardwaj, Arvind Sharma V.K.Shrivastava, "Data Mining Techniques and Their Implementation in Blood Bank Sector-A Review published in IJERA, ISSN: 2248-9622, Vol.2 Issued4, July –August 2012, pp.1303-1309.

[6] Jiawei Han, Micheline Kamber, Jian pei.," Data mining concepts and techniques", 3rd edition, page no:350-354

[7] Margaret H.Dunham.,"Data Mining Introductory and Advanced Topics".

ISSN: 2278 – 1323