Volume 4, Issue 6, June 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Prediction of Blood Donors' Population using Data Mining Classification Technique

Ritika[1],
[1]M.Tech Student
Department of Computer Science
[1]Eternal University, Baru Sahib
Himachal Pradesh-173101, India

Aman Paul[2]
[2]Assistant Professor
Department of Computer Science
[2]Eternal University, Baru Sahib
Himachal Pradesh-173101, India

ABSTRACT- Data mining is a technique that finds relationships and trends in large datasets to promote decision support. Classification is a data mining technique that maps data into predefined classes often referred as supervised learning because classes are determined before examining data. Different classification algorithms are analyzed for the effective classification of data. The aim is to examine different classification algorithms and to find out a classification technique with best accuracy rate & least error for the prediction of blood donors.

KEYWORDS
Data mining, Classification, Classification algorithms, Classification accuracy, Error rate, Weka 3.6.9 tool, CBA (v2.1) tool.

I. INTRODUCTION

Classification is one of the most important data mining techniques. It is a concept or process of finding a model which discovers the class of unknown objects. Actually it maps the data items into one of some predefined classes. Classification model generate a set of rules based on the features of the data in the training dataset. Further these rules can be used for classification of future unknown data items. In the field of machine learning and statistics, classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. The distinct observations are examined into a set of quantifiable properties, known as various explanatory variables, features, etc. These properties may be, ordinal , integer-valued or real-valued. Some algorithms work only in terms of discrete data and need that real-valued or integer-valued data be discretized into groups.

II. CLASSIFICATION ALGORITHMS

1. J48: J48 classifier is a simple C4.5 decision tree for classification and creates a binary tree. The decision tree approach is very helpful in classification problem. According to this technique, a tree is constructed which models the classification process. Once the tree is formed, it is applied to each tuple in the database and results in classification for that tuple [1]. The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a 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 is 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 the target variable, then we terminate that branch and assign to it the target value that we have obtained.

2. Naive Bayes: A naive Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features. In other words, the Naive Bayes algorithm is a simple probabilistic classifier used for calculating a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. The algorithm tends to perform well and learn rapidly in a variety of supervised classification problems [1].

3. BayesNet: A Bayesian network or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases

and symptoms. Graphical models such as Bayesian networks provide a general framework which is used for dealing with uncertainty in a probabilistic setting and thus are well suited to tackle the problem of churn management. Bayesian Networks was coined by Pearl (1985). In Bayesian network, every graph codes a class of probability distributions. The nodes of that graph comply with the variables of the problem domain. Arrows between nodes show relations between the variables. These dependencies are quantified by conditional distributions for every node given its parents.

4. ZeroR: The simplest of the rule based classifiers is the majority class classifier, called 0-R or ZeroR in Weka. The 0-R (zero rule) classifier takes a look at the target attribute and its possible values. It will always output the value that is most commonly found for the target attribute in the given dataset. 0-R as its names suggests; it does not include any rule that works on the non target attributes. So more specifically it predicts the mean (for a numeric type target attribute) or the mode (for a nominal type attribute). Zero-R is a simple and trivial classifier, but it gives a lower bound on the performance of a given dataset which should be significantly improved by more complex classifiers. As such it is a reasonable test on how well the class can be predicted without considering the other attributes. It can be used as a Lower Bound on Performance [2].

5. Ridor: Ridor algorithm generates a default rule first and then the exceptions for the default rule with the least (weighted) error rate. Then it generates the "best" exceptions for each exception and iterates until pure. Thus it performs a tree-like expansion of exceptions. The exceptions are a set of rules that predict classes other than the default. IREP is used to generate the exceptions. It is well known that classification models produced by the Ripple Down Rules (RDR) are easier to update and maintain. They are compact and are capable of providing an explanation of their reasoning which makes them easy to understand for medical practitioners. Ripple Down Rules were initially introduced as an approach that facilitated the maintenance problem in knowledge based systems. Their applications in various domains have been actively investigated. Multiple Classification Ripple Down Rules (MCRDR) are of particular interest for medical applications, since they are capable of producing multiple conclusions for each instance, which may correspond to several diagnoses for one patient [3].

6. PART: PART algorithm combines two general data mining strategies; the divide-and-conquer strategy for decision tree learning and the separate-and-conquer strategy for rule learning. In the divide-and-conquer approach, an attribute is placed at the root node and then the tree is divided by making branches for each possible value of the attribute. For each branch the same process is carried out recursively, using only those instances that reach the branch. In order to build the rules, the separate-and-conquer strategy is employed. A rule is derived from the branch of the decision tree explaining the most cases in the dataset, instances covered by the rule are removed, and the algorithm continues creating rules recursively for the remaining instances until none are left [4].

7. Prism: The Prism algorithm was introduced by Cendrowska .The Prism classification rule induction algorithm promises to induce qualitatively better rules compared with the traditional TDIDT (Top Down Induction of Decision Trees) algorithm. Compared with decision trees Prism is less vulnerable to clashes, it is bias towards leaving a test record unclassified rather than giving it a wrong classification and it often produces many fewer terms than the TDIDT algorithm if there are missing values in the training set. The algorithm generates the rules concluding each of the possible classes in turn. Each rule is generated term by term, with each term of the form 'attribute =value'. The attribute/value pair added at each step is chosen to maximize the probability of the target 'outcome class'.

8. CBA (Classification Based Association): Association rules are used to analyse relationships between data in large databases. Classification involves learning a function which is capable of mapping in stances into distinct classes. Now both the association rule mining and classification rule mining can be integrated to form a framework called as Associative Classification and these rules are referred as Class Association Rules. The use of association rules for classification is restricted to problems where the instances can only belong to a discrete number of classes. The reason is that association rule mining is only possible for categorical attributes. However, association rules in their general form cannot be used directly. We have to restrict their definition. When we want to use rules for classification, we are interested in rules that are capable of assigning a class membership. A class association rule is obviously a predictive task. By using the discriminative power of the Class Association Rules we can also build a classifier [5].

III. RELATED WORK

Some classification techniques are described for Blood Group Donors datasets and data mining classifiers are used to generate decision tree. The ability to identify regular blood donors will enable blood bank and voluntary organizations to plan systematically for organizing blood donation camps in an efficient manner. The primary focus of this research is the development of a system that is essential for the timely analysis of huge Blood Group Donors data sets. The analysis had been carried out using a standard blood group donor's dataset and using the J48 decision tree algorithm implemented in Weka. The research work is used to classify the blood donors based on the sex, blood group, weight and age [6].

A data mining model is build to extract knowledge of blood donor's classification to aid clinical decisions in blood bank centre. J48 algorithm and Weka tool have been used for the complete research work. This study utilized real world data collected from an EDP department of a blood bank centre and used J48 algorithm for the classification of donors, which can help the blood bank owner to make proper decisions faster and more accurately. Through training and

evaluation, the experimental results showed that the generated classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9% [7].

Several characteristics are studied here that determines the presence of cataract and people suffering from cataract are found out from the population of 790 (instances with 11 different attributes) using Weka tool. WEKA tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared. In this study, several algorithms are used such as NaiveBayes, SMO, J48, and REPTree and Random Tree and from the results it is observed that Random Tree algorithm is appropriate for cataract performance. Random Tree gives 84% which is relatively higher than other algorithms [8].

Three different data mining classification methods are used for prediction of breast cancer and the comparison between three algorithms i.e. Decision tree, Bayesian Network and K-Nearest Neighbour algorithms is conducted with help of WEKA (The Waikato Environment for Knowledge Analysis) tool, which is open source software. In order to compare the results, some parameters are used and those are correctly classified instances, incorrectly classified instances, time taken, kappa statistic, relative absolute error, and root relative squared error. This comparison is done for the prediction of cancer. But for superior prediction, the main focus is on accuracy and lowest computing time. This study filtered all algorithms based on lowest computing time and accuracy and this conclusion is drawn that Naïve Bayes is a superior algorithm compared to the two others because it takes lowest time i.e. 0.02 seconds and at the same time is providing highest accuracy [9].

Some discussion and demonstration of different classification and association rule mining algorithms is done here. The main contributions are like theoretical survey on association rule mining algorithms, comparison of traditional classification techniques such as decision trees, Naïve Bayes, RIPPER and others, survey on the use of association rule mining in classification framework, extensive experimental comparison studies using several data sets between five popular traditional classification algorithms, i.e. OneR, decision trees C4.5, PART, Naïve Bayes, and RIPPER, and two well known classifications based on a association rule algorithms, i.e. CBA and MMAC, in term of number of rules produced, runtime and classification rate. Performance evaluation on different data sets shows that there is consistency between C4.5 and PART with respect to rules features, error rate, and size of the resulting classifiers. MMAC algorithm acted constantly well in term of classification accuracy on both artificial data sets, i.e. UCI data, and the real world data sets. Also, OneR algorithm performed really well on real world data sets. The results also pointed out that Naive Bayes and OneR algorithms are the fastest ones to frame the classification system. On the other hand, due to the optimization phase RIPPER is the slowest algorithm in building the classification system [10].

IV. DATASETS AND TOOLS USED

1)Hardware: We conduct our evaluation on Intel Pentium platform which consist of 2 GB RAM and 250 GB hard disk.

2) Software: For this experiment, we used Weka 3.6.9, CBA (v2.1) and window 7 to predict the blood donors' population. Weka is data mining software written in Java Language. WEKA supports many data mining tasks such as data re-processing, classification, clustering, regression and feature selection to name a few.

CBA mines association rules and builds accurate classifiers using a subset of association rules. This system includes CBA for classification based on association and many more features like Table Class Rules, Table Assoc Rules, Training Data File, Data Converter, Discretizer, Feature Selection, Mine: Multi Sup, Mine: Single Sup, Cross Validation, View Rules, View Tree, Prediction.

3) Data Set: The input data set is an integral part of data mining application. The dataset used in my research is the real world data obtained from the IGMC blood bank, Shimla. Blood Donors' dataset consists of 1000 instances with 7 attributes in the area of Health Sciences and none of them contains missing value.

4) Discussion and Experimental results: For classification algorithms, I have explored blood donors' dataset and tried to figure out which classification technique has the best accuracy rate & least error rate for the prediction of blood donors. In Weka, results are computed with two applications i.e. Explorer & Experimenter. In Explorer, accuracy rates & errors rates are computed for different algorithms under three Test Modes namely Training Mode, Cross Validation Mode (10 Folds) and Percentage Split Mode (66%). In Experimenter, different algorithms are compared and their accuracy is obtained at once.

Table 1. Blood Donors dataset used for the experimental results

Data Set	Blood Donors
Instance	1000
Attributes	7
Area	Health Science
Missing values	0

Error Rates are obtained as higher the accuracy rate; least will be the error rate. By comparing the results of Explorer & Experimenter in Cross Validation Mode (10 Folds), we found out that in some cases their results are slightly different. So, we preferred to consider the results of Experimenter in order to make comparison with CBA (Classification Based

Association). A comparison is also done among PART, J48 and CBA algorithm on the basis of varying confidence factors. The results of CBA are computed directly with the help of CBA tool.

In Explorer, by applying Prism, Ridor, J48, ZeroR, PART, NaiveBayes & BayesNet algorithms on blood donors' dataset, the best and worst classification accuracies on the basis of blood group was 38.20% of both PART & BayesNet algorithms and 17.50% of Prism algorithm respectively in Training Mode, the best and worst classification accuracies on the basis of blood group was 37.90% of ZeroR algorithm and 16.10% of Prism algorithm respectively in Cross Validation, the best and worst classification accuracies on the basis of blood group was 34.41% of PART, ZeroR & J48 algorithms and 18.23% of Prism algorithm respectively in Percentage Split. The best and worst classification accuracies on the basis of age was 86.70% of J48 algorithm and 83.60% of ZeroR algorithm respectively in Training Mode, the best and worst classification accuracies on the basis of age was 86.50% of both Ridor & J48 algorithms and 83.60% of ZeroR algorithm respectively in Cross Validation, the best and worst classification accuracies on the basis of age was 83.23% of J48, Ridor & PART algorithms and 79.11% of ZeroR algorithm respectively in Percentage Split. The best and worst classification accuracies on the basis of weight was 97.50% of J48, Ridor & PART algorithms and 95.20% of both ZeroR & Prism algorithms respectively in Training Mode, the best and worst classification accuracies on the basis of weight was 97.50% of PART, Ridor & J48 algorithms and 95.20% of ZeroR algorithm respectively in Cross Validation, the best and worst classification accuracies on the basis of weight was 97.94% of BayesNet, NaiveBayes, PART, Ridor & J48 algorithms and 95.29% of both Prism & ZeroR algorithms respectively in Percentage Split. In Training Mode, the best average Classification Accuracy and minimum Error Rate was 74.10% & 25.90% respectively of PART algorithm. In Cross Validation, the best average Classification Accuracy and minimum Error Rate was 73.86% & 26.14% respectively of J48 algorithm. In Percentage Split, the best average Classification Accuracy and minimum Error Rate was 71.86% & 28.14% respectively of both J48 & PART algorithms.

In Experimenter, by applying Prism, Rider, J48, ZeroR, PART, NaiveBayes & BayesNet algorithms on blood donors' dataset, the best Classification accuracy & least Error Rate on the basis of blood group were found to be 37.90% & 62.10% respectively of ZeroR algorithm. The best Classification accuracy & least Error Rate on the basis of age were found to be 86.41% & 13.59% respectively of J48 algorithm. The best Classification accuracy & least Error Rate on the basis of weight were found to be 97.50% & 2.50% respectively of both PART & J48 algorithm.

PART, J48 & CBA were compared by varying the confidence factors and it was found that by varying the confidence factor, Classification Accuracy does vary. Every time CBA obtained maximum accuracy. But with 40% confidence, it obtained maximum Classification Accuracy that was of 74.66%.

Finally, the Accuracy & Error Rate of all the algorithms were compared with the Accuracy & Error Rate of CBA & it was found that CBA obtained maximum Accuracy i.e. 74.66% and minimum Error Rate i.e. 25.34%.

So, we considered CBA for our prediction of blood donors because the Classification Accuracy obtained through CBA model was 74.66%. It means that the model is able to predict the values 75% (approx.) correctly which is quite good. So, if this model is used to find out the donating blood decisions of new donors, the probability will be 0.75.

V. CONCLUSION

Data mining is "the science of extracting useful information from large databases". It is used to determine knowledge out of data and presenting it in a form that is easily understood to humans. Data mining techniques have been used for industrial, commercial and scientific purposes. Classification is one of the main tasks of data mining with broad applications to classify the various kinds of data used in nearly every field of our life. Basically, it is used to classify the item according to the features of the item with respect to the predefined set of classes. In this work, different classification algorithms like Naïve Bayes, Bayes net, J48, Prism, PART, Ridor, ZeroR and CBA are discussed and compared. These algorithms are applied on blood donors' dataset to find out their accuracy and error rate. The accuracy is computed as the sum of true positive and true negative over the sum of true positive, true negative, false positive and false negative OR total number of correctly classified instances over total number of instances. The error rate is computed as the sum of false positive and false negative over the sum of true positive, true negative, false positive and false negative OR total number of incorrectly classified instances over total number of instances. With the help of achieved classification accuracy and error rates, different classification algorithms are compared and the one with best accuracy and worst error rate i.e. CBA is used for the prediction of blood donors'.

REFERENCES

[1] Tina R. Patil and S. S. Sherekar (2013) "*Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification",* International Journal Of Computer Science And Applications, vol. 2, 6, 2013, pp. 256-261.

[2] Inamdar S. A., Narangale S.M. and Shinde G. N. (2011) "*Preprocessor Agent Approach to Knowledge Discovery Using Zero-R Algorithm",* International Journal of Advanced Computer Science and Applications, vol. 2, 12, 2011, pp. 82-84.

[3] A.V. Kelarev, R. Dazeley, A. Stranieri, J.L Yearwood and H.F. Jelinek, "*Detection of CAN by Ensemble Classifiers based on Ripple Down Rules*" Centre for Informatics and Applied Optimization, School of SITE, University of Ballarat, Australia.

[4] Adem Karahoca, Dilek Karahoca and Nizamettin Aydın, "*Benchmarking the Data Mining Algorithms with Adaptive Neuro-Fuzzy Inference System in GSM Churn Management*" Software Engineering Department, Bahcesehir University, Turkey.

[5] Shekhawat P. B. and Dhande S. S. (2011) "*A Classification Technique using Associative Classification",* International Journal of Computer Applications, vol. 20, 5, 2011, pp. 20-28.

[6] Ramachandran P, Girija N, Bhuvaneswari T (2011) "*Classifying Blood Donors Using Data Mining Techniques",* International Journal of Computer Science & Information Technologies, vol. 1, 1, 2011, pp. 10-13.

[7] Sharma A and Gupta P. C. (2012) "*Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool",* International Journal of Communication and Computer Technologies, vol. 1, 6, 2012, pp. 6-10.

[8] Sugandhi C, Yasodha P and Kannan M (2011) "*Analysis of a Population of Cataract Patients Databases in Weka Tool",* International Journal of Scientific & Engineering Research, vol. 2, 10, 2011, pp. 1-5.

[9] Mr. Chintan Shah and Dr. Anjali G. Jivani (2013) "*Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction*", 4th ICCCNT(IEEE), 4-6 July, 2013. Tiruchengode, India.

[10] "Alaa Al Deen" Mustafa Nofal and Sulieman Bani-Ahmad "*Classification Based on Association-Rule Mining Techniques: A General Survey and Empirical Comparative Evaluation",* Ubiquitous Computing and Communication Journal, vol. 5, 3, pp. 9-17.

[11] Sedigheh Khajouei Nejad, Farid Seifi, Hamed Ahmadi and Nima Seifi(2009) "*Applying Data Mining in Prediction and Classification of Urban Traffic",* World Congress on Computer Science and Information Engineering(IEEE), 2009, pp. 674-678.

[12] Rui Wang, Weishan Dong, Yu Wang, Ke Tang and Xin Yao (2013) "*Pipe Failure Prediction: A Data Mining Method*", IEEE- ICDE Conference, 2013. University of Birmingham UK, pp. 1208-1218.

[13] G.Kesavaraj and Dr.S.Sukumaran(2013) "*A Study On Classification Techniques in Data Mining*", 4th ICCCNT(IEEE), 4-6 July, 2013. Tiruchengode, India.

[14] S.Vijiyarani and S.Sudha (2013) "*Disease Prediction in Data Mining Technique – A Survey",* International Journal of Computer Applications & Information Technology, vol. 2, 1, 2013, pp. 17-21.

[15] Nikhil N. Salvithal and Dr. R. B. Kulkarni (2013) "*Evaluating Performance of Data Mining Classification Algorithm in Weka",* International Journal of Application or Innovation in Engineering & Management, vol. 2, 10, 2013, pp. 273-281.

[16] Mahendra Tiwari and Randhir Singh (2013) "*A Benchmark to Select Classification Algorithms for Decision Support System",* International Journal of Scientific and Research Publications, vol. 3, 1, 2013, pp. 1-5.

[17] Yogendra Kumar Jain and Upendra (2012) "*An Efficient Intrusion Detection Based on Decision Tree Classifier Using Feature Reduction",* International Journal of Scientific and Research Publications, vol. 2, 1, 2012, pp. 1-6.

[18] Jerzy Stefanowski (2008) "*Data Mining - Clustering*", Institute of Computing Sciences, Poznan University of Technology Poznan, Poland.

[19] Jiawei Han and Micheline Kamber (2006), "*Data Mining Concepts and Techniques*" 2nd Edition, Morgan Kaufmann Publishers, California.

[20] Sundaram S and Santhanam T (2011) "*A Comparison of Blood Donor Classification Data Mining Models",* Journal of Theoretical and Applied Information Technology, vol. 30, 2, 2011, pp. 98-101.

