RESEARCH ARTICLE OPEN ACCESS

Preprocessing and Classification in WEKA Using Different Classifiers

Payal P.Dhakate[1], Suvarna Patil[2], K. Rajeswari[3], Deepa Abin[4]
[1,2,3,4] Department of Computer Science, Pune University, Aakurdi Pune

ABSTRACT
Data mining is a process of extracting information from a dataset and transform it into understandable structure for further use, also it discovers patterns in large data sets [1]. Data mining has number of important techniques such as preprocessing, classification. Classification is one such technique which is based on supervised learning. It is a technique used for predicting group membership for the data instance. Here in this paper we use preprocessing, classification on diabetes database. Here we apply classifiers on this database and compare the result based on certain parameters using WEKA. 77.2 million people in India are suffering from pre diabetes. ICMR estimates that around 65.1million are diabetes patients. Globally in year 2010, 227 to 285 million people had diabetes, out of that 90% cases are related to type 2 ,this is equal to 3.3% of the population with equal rates in both women and men in 2011 it resulted in 1.4 million deaths worldwide making it the leading cause of death.

Keywords- AD Tree; J48; Random Tree; REP Tree; Simple cart; WEKA;

I. INTRODUCTION

WEKA is introduced by Waikato University , it is open source software written in Java and used for different purposes such as research, education**.** Fig 1 illustrates WEKA Interface. Classification is the process of finding model or function which describes and distinguishes data classes or concepts, for the intension of using the model to predict the class of objects whose class label is unknown. The problem is a comparative study of classification technique such as Random Tree, FT tree, LMT Tree using various parameters using Diabetes Dataset containing 9 attributes and 768 instances. **C**lassification **a**nd **R**egression **t**rees was introduced by Breiman. It is based on binary splitting of the attributes. Gini index is used in selecting the splitting attribute. It uses both numeric and categorical attributes for building the decision tree and also uses in-built features to deal with missing attributes. FT tree is classifier for building functional trees, which are classification trees that could have logistic regression functions at the inner nodes and/or leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values Lmt tree is classifier for building logistic model trees, which are classification trees with logistic regression functions at the leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values.REP Tree is a fast decision tree learner, it builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning. It only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces. Lad tree is class for generating a multi-class alternating decision tree using the Logit Boost strategy. In this experiment we are using diabetes database having 768 attributes and 7 instances, some of the attributes are preg , plas, pedi ,age ,class.

Fig1. WEKA Interface

Here in this paper in Section 2 proposed method is discussed defines problem statement. Section 3 describes the proposed classification method to identify the class of diabetes using data mining classification Random tree. Experimental results and performance evaluation are discussed in Section 4 and in Section 5 conclusion is putforth.

II. PROPOSED METHOD

Classification is the process of finding a model (or function) that describes and distinguishes data

classes or concepts, for the able purpose of being to use the model to predict the class of objects whose class label is unknown [3]. It is a technique which is used to predict group membership for data instances. Classification is a two step process, first, it builds classification model using training data. Every object of the dataset must be pre-classified i.e. its class label must be known, second the model generated in the preceding step is tested by assigning class labels to data objects in a test dataset.Here we are using diabetes dataset because now a days the percentage of diabetes patient is growing very fast. According to Diabetes Atlas published by the International Diabetes Federation (IDF), there were an estimated 40 million persons with diabetes in India in 2007 and this number is predicted to rise to almost 70 million people by 2025. India accounts for the largest number of people 50.8 million suffering from diabetes in the world. India continues to be the "diabetes capital" of the world, and by 2030, nearly 9 per cent of the country's population is likely to be affected from the disease It is estimated that every fifth person with diabetes will be an Indian. This means that India has highest number of diabetes in any one of the country in the world . Here we are using diabetes database having 768 attributes and 7 instances. It consists of attributes like preg ,mass, age, insu .These attributes predict whether a person having diabetes or not.Consider an example we may wish whether a person having type1, type2 diabetes. In classification we apply different classifiers such as simple cart, AD Tree, Random tree etc. Discretization is applied as preprocessing technique, such as preg, plas, pres, insu and it is apply to all attributes. Out of these attributes we preprocess it for preg attribute and it is shown by Fig.2

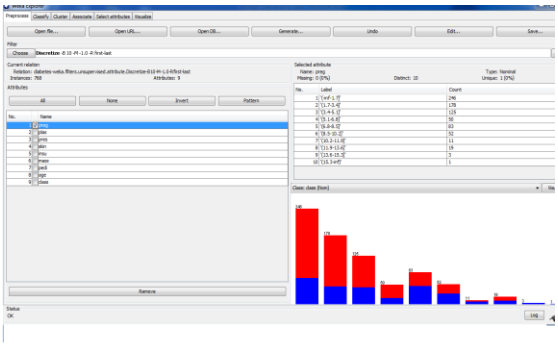

Fig2. Preprocessing of diabetes database

Then we can compare the results of different classification techniques. Fig. 4 can shows Comparison of classifiers using different measures .In this paper out of all other classifiers the Random Tree gives better accuracy and also the time required is less. The below Fig. 3 shows this random tree and Fig .5 illustrates Accuracy and time required for trees different trees. Out of this Random tree gives good accuracy.

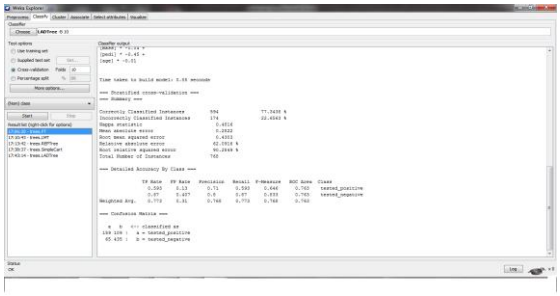

Fig 3.Classification using FT Tree

III. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

a) Measures for performance evaluation:

To measure the performance sensitivity, accuracy and specificity are used, such concepts are readily usable for the evaluation of any binary classifier. TP is true positive, FP is false positive, TN is true negative and FN is false negative. TPR is true positive rate, which is equivalent to Recall [2].

Sensitivity=True Positive Rate=True Positive/(True Positive + False Negative) *(1)*

Specificity=True Negative/(False Positive + True Negative) *(2)*

b) Precision:

Precision is calculated as number of correctly classified instances belongs to X divided by number of instances classified as belonging to class X; that is, it is the proportion of true positives out of all positive results and can be defined as [2].

Precision=True Positive/(True Positive + False Positive) (3)

c) *Accuracy*:

It is a ratio of ((no. of correctly classified instances) / (total no. of instances)) *100) and it can be defined as [2].

Accuracy=(True Positive + True Negative)/(True Positive + False Negative) + (False Negative + True Negative) (4)

***d) False Positive Rate* :**

It is simply the ratio of false positives to false positives plus true negatives.[2] False positive rate is zero in ideal world. It can be defined as [2].

False Positive Rate =False Positive/(False Positive + True Negative) (5)

e) F-Measure: F-measure is nothing but combining recall and precision scores into a single measure of performance. It is given as [2].

F-Measure=2*recall*precision/(recall + precision) (6)

Trees	**J48**	**FT Tree**	**LAD**	**Simple Cart**	**LMT**	**REP Tree**
TP Rate	1	0.593	0.57	**0.534**	0.56	0.58
FP Rate	**0.053**	0.13	0.17	0.132	0.11	0.154
Precision	0.833	0.71	**0.644**	0.684	0.732	0.67
Recall	1	0.593	0.575	**0.534**	0.56	0.582
Time Taken	0.04	0.08	0.08	0.06	0.59	**0.02**
Correctly Classified Instances (%)	**91**	**77**	74	75	**77**	75
Accuracy(%)	**84**	**77**	74	75	**77**	75

Table .1 Comparison of classifiers using different measures

	Kappa statistic	**Mean absolute error**	**Root mean squared error**	**Relative absolute error**	**Root relative squared error**
LAD tree	0.415	**0.0322**	0.4237	70.85	88.88
REP tree	0.4415	**0.3261**	0.4181	75.21	90.7
J48	0.6319	**0.2383**	0.3452	52.4339	72.4207
Simple cart	0.4232	**0.3419**	75.21	76.49	87.47
LMT Tree	0.4756	**0.3175**	0.3963	69.84	83.15

Table .2Error rates of different classifier

IV. CONCLUSION

In this paper we discussed data mining, preprocessing and different classification techniques on diabetes database using WEKA tool . The FT Tree has shown accurate results than other techniques such as LAD Tree , Simple cart,J48 , LMT Tree and REP Tree in terms of accuracy ,time and errors .

REFERENCES

[1] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*", Morgan Kaufmann, 2000.

[2] Varun Kumar and Nisha Rathee," *Knowledge discovery from database Using an integration of clustering and classification", (IJACSA) International Journal of Advanced Computer Science and Applications, 2011.*

[3] Swasti Singhal, Monika Jena, "*A Study on WEKA Tool for Data Preprocessing, Classification and Clustering*", *International Journal of Innovative Technology and Exploring Engineering(IJITEE), 2013*