International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. 05 Special Issue 01 | 2020

Analysis of Decision Tree Classification Methods Using WEKA

Mr. Dewendra Onkar Bharambe[1], Asst. Prof. Shubhangi Kashinath Patil[2]
[1]Lect. Department of Computer Engineering, J. T. Mahajan Polytechnic, Fazipur[1]
[2]Asst. Prof., Department of Computer Science,Dhanaji Nana Mahavidyalaya, Faizpur[2]

ABSTRACT

Now a day's analyzing large amount of data is a necessity. People have no time to look at the extremely large data like medical data, marketing data, and financial data. So we must have the technique to automatically analyze the data. Data mining is the process of extracting useful information from large amount of data and analyze, classify and summaries it into useful information. Data mining classification technique is used for prediction to diagnose Diabetes. In this paper we have used J48,LAD,REP and LMT decision tree classifiers on diabetes dataset and compare the performance of different classifiers on the basis of accuracy, recall, precision, F-measure, computing time, correctly classified instances, also we observe kappa statistics, MAE, RMSE, RAE, RRSE to find the error rate measurement for different classifiers in WEKA .We include confusion matrices of different classifier to quickly analyze the classifiers. We explored some data mining classification methods to select the suitable methods for efficient classification of Diabetes dataset.

Keywords- Data Mining, WEKA tool, Diabetes Patients dataset, Decision Tree Classification algorithm

1. INTRODUCTION

Data mining turn a large amount of data in Knowledge. It is also called as exploratory data analysis, data driven discovery and deductive learning. Classification is the most popular data mining technique. Classification assigns categories to a collection of data in order to aide in more accurate prediction and analysis. Diabetes affects a large number of the world population and it's a hard disease to diagnose. The main goal of this paper is the classification of Diabetes datasets by using different decision tree classifiers to determine if a person is diabetic or not. We compare different classifiers to select best classifier in order to correctly classify the Diabetes datasets to diagnose the disease more cost effectively.

To diagnose the disease we use the attributes of patient like number of times pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg) , triceps skin fold thickness (mm),2-hour serum insulin (mu U/ml), body mass index (weight in kg/(height in m)^2), diabetes pedigree function and Age (years).Applying various decision tree classifier we classify this dataset and try to find out which is most efficient classifier that can correctly classify maximum amount of instances within small amount of time.

2. WEKA (WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS)

Waikato Environment for Knowledge Analysis (WEKA) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. Weka is a workbench [1] that contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. It is mostly used to load datasets, run algorithms and design and run experiments with results statistically robust enough to publish.

WEKA tool consist of classification methods based on decision trees like the J48 decision tree, some are rule-based like ZeroR and decision tables, and some of them are based on probability and regression, like the Naïve Bye's algorithm. WEKA requires dataset file in ARFF(Attribute Relation file format) format and the file name should have extension dot ARFF (.arff). WEKA is available on the web at www.cs.waikato .ac.nz/ml/weka.

Waikato Environment for Knowledge Analysis
Version 3.5.8
(c) 1999 - 2008
The University of Waikato
Hamilton, New Zealand

3. CLASSIFICATION

Data classification is the process of organizing data into categories for its most effective and efficient use. In data mining there are various classification algorithms such as decision trees, logistic regression, neural networks, etc. In this paper we are using decision tree algorithm for classification. The Classification process involves following steps:

1. Create training data set.
2. Identify class attribute and classes.
3. Identify useful attributes for classification (Relevance analysis).
4. Learn a model using training examples in Training set.
5. Use the model to classify the unknown data samples

3.1 Decision Tree Classification Methods

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

The following decision tree is for the concept buy computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.

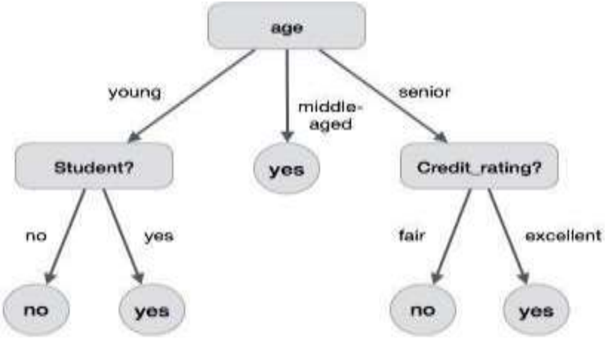

Figure I: Decision Tree

The benefits of having a decision tree are as follows –

1. It does not require any domain knowledge.
2. It is easy to comprehend.
3. The learning and classification steps of a decision tree are simple and fast.

Decision tree classifiers

3.1.1.**J48:** Weka algorithm J48 is the improved version of C4.5. The algorithm uses a greedy technique for decision making.Structure of the output decisiontree having different nodes, such as root node, intermediate nodes and leaf node. Each internal node in the tree denotes different attributes, while the terminal nodes tell us the final value of the dependent variable.

3.1.2. **LAD:** Logical Analysis of Data(LAD) tree is the classifier suggest a way of analyzing data through combinational logic ,Boolean function,and optimization techniques.LAD detect logical combinatory information.

3.1.3.**REP Tree[1] :** Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).

***3.1.4 .LMT[1]* :** A classification model with an associated supervised training algorithm that combines logistic prediction and decision tree learning is logistic model tree (LMT)[7] . Logistic model trees use a decision tree that has linear regression models at its leaves to provide a section wise linear regression model.

4. DATASET

Dataset is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statisticaldata matrix, where every column of the table represents a particular variable, and each rowcorresponds to a given member of the data set in question.

In this paper we are using Pima Indians Diabetes Database available on weka. The Original owners of this dataset are National Institute of Diabetes and Digestive Kidney Diseases.

The dataset contains 768 Number of Instances having 8 plus attributes as follows:
1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

For Each Attribute: (all numeric-valued)The dataset contains no missing attribute values.

5. RESULTS AND DISCUSSION

5.1 Evalution Matrics

The result of classification is based on following performance metrics [1]

1. Time: This is referred to as the time required to complete training or modeling of a dataset. It is represented in seconds.
2. Kappa Statistic: A measure of the degree of nonrandom agreement between observers or measurements of the same categorical variable.
3. Mean Absolute Error: Mean absolute error is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error.
4. Mean Squared Error: Mean-squared error is oneof the most commonly used measures of success for numeric rediction. This value is computed by taking theaverage of the squared differences between eachcomputed value and its corresponding correct value. Themean-squared error is simply the square root of themean-squared-error. The mean-squared error gives theerror value the same dimensionality as the actual andpredicted values.
5. Root relative squared error: Relative squarederror is the total squared error made relative to what theerror would have been if the prediction had been theaverage of the absolute value. As with the root mean squared error, the square root of the relative squarederror is taken to give it the same dimensions as thepredicted value.
6. Relative Absolute Error: Relative Absolute Erroris the total absolute error made relative to what the errorwould have been if the prediction simply had been theaverage of the actual values.

Using this metrics the result in Table I is obtained.

A confusion matrix is a useful tool for analyzing classifier accuracy. Structure of confusion matrix is given below.

	a	b
a	True Negative	False Positive
b	False Negative	True Positive

Figure II: confusion matrix

Using this metrics the result in Table II is obtained.

True Positive (TP) refers to positive tuples that were correctly labeled by the classifier. True Negative (TN) refers to negatives tuples that were correctly labeled by the classifier. False Positive (FP) refers to negatives tuples that were incorrectly labeled by the classifier. False Negative (FN) refers to positive tuples that were incorrectly labeled by the classifier.

Accuracy: Accuracy is the percentage of tuples that are correctly classified by the classifier.

$$Accuracy=(TP+TN)/(TP+TN+FP+FN)$$

Recall: Recall is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured.

$$Recall=TP/(TP+FN)$$

Precision: Precision is the proportion of the examples which truly have class x among all those which were classified as class x.

$$Precision=TP/(TP+FP)$$

F-Measure: The harmonic mean of precision and recall. It is an important measure as it gives equal importance to precision and recall.

$$F\text{-}measure=2*recall*precision/precision + recall$$

Receiver Operating Characteristic (ROC) Curve: It is a graphical approach for displaying the tradeoff between true positive rate (TPR) and false positive rate (FPR) of a classifier. TPR is plotted along the y axis and FPR is plotted along the x axis. Performance of each classifier represented as a point on the ROC curve.

5.2 **Result:** The cross validation method used to analysis for the datasets. Various performance measures for all the datasets mentioned in Table I, II, III. Comparative analysis of various decision tree classification results as follows –

Table: 3. Errors Measurement For Different Decision Tree Classifiers In Weka

	J48	LAD	REP	LMT
Time (Seconds)	0.02	0.13	0.02	2.14
Correctly Classified Instances	567 (73.82%)	569 (74.08%)	578 (75.26%)	595 (77.47%)
KAPPA Statistic	0.4164	0.415	0.438	0.4756
MAE	0.3158	0.322	0.3272	0.3175
RMSE	0.4463	0.4237	0.4289	0.3963
RAE %	69.48	70.85	71.98	69.84
RRSE%	93.62	88.88	89.97	83.15

Table: 4. Confusion Metrics For Different Decision Tree Classifiers In Weka

Decision Tree	True Negative	True Positive	Correctly Classified Instances
J48	407	160	567
LAD	415	154	569
REP	423	155	578
LMT	445	150	595

Table: 5. Performance Metrics For Different Decision Tree Classifiers In Weka

Decision Tree	TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE	ROC CURVE AREA
J48	0.738	0.327	0.735	0.738	0.736	0.751
LAD	0.741	0.336	0.736	0.741	0.737	0.788
REP	0.753	0.328	0.747	0.753	0.748	0.766
LMT	0.775	0.325	0.77	0.775	0.766	0.831

6. CONCLUSIONS

In this paper we have studied four different decisions tree classification methods. We analyze J48, LAD, REP, LMT decision tree classification method by applying on diabetes dataset. We conclude that out of these classifiers LMT classified maximum instances but it requires 2.14 seconds where as REP and J48 classifier require same time 0.2 seconds but REP correctly classifies more instances than J48. The performance of J48,LAD,REP,LMT classifier for correctly classified instances is constantly increases.

7. REFERENCES

[1]P.Yasodha, N.R. Ananthanarayanan “Comparative Study of Diabetic Patient Data’s Using Classification Algorithm in WEKA Tool” International Journal of Computer Applications Technology and Research Volume 3– Issue 9, 554 - 558, 2014, ISSN: 2319–8656
[2] PurvaSewaiwar, Kamal Kant Verma “Comparative Study of Various Decision Tree Classification Algorithm Using WEKA”*International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-10)*
[3] K. Rajesh, V. Sangeetha “Application of Data Mining Methods and Techniques for Diabetes Diagnosis” International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012
[4] http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff
[5]Jiawei Han “Data mining Concepts and Techniques” Third Edition
[6] www.cs.waikato