Indian Journal of Science and Technology, *Vol 10(18), DOI: 10.17485/ijst/2017/v10i18/111768, May 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Comparing the Efficacy of Decision Tree and its Variants using Medical Data

A. Sheik Abdullah[1]*, S. Selvakumar[2], P. Karthikeyan[1] and M. Venkatesh[3]

[1]Department of Information Technology, Thiagarajar College of Engineering, Madurai – 625015, Tamil Nadu, India; asait@tce.edu, karthikit@tce.edu
[2]Department of Computer Science and Engineering, G.K.M. College of Engineering and Technology, Chennai – 600063, Tamil Nadu, India; sselvakumar@yahoo.com
[3]Department of Medicine, Theni Government Medical College and Hospital, Theni – 625531, Tamil Nadu, India; dremvee@gmail.com

Abstract

The objective of this research work focus towards the identification of best variant between decision tree algorithm such as Weighted Decision Trees (WDT), C4.5 Decision Trees and C5.0. Methods: Decision tree has a number of variants such as ID3, Weight based decision tree, C4.5 and C5.0 algorithms. This research work focus towards the predictive performance analysis of weight based decision tree with information gain as the splitting criterion. The algorithm proceeds iteratively with the assignment of weights over the training instances to determine the best among the data attributes. Thereby, the attribute with best weight values can be significantly determined by an improvement over its accuracy. Results: The experimental results proves that among the variants of decision trees the algorithm corresponding to C4.5 provides the highest accuracy of about 71.42% and R^2 value of about 0.265 respectively and for real world data the accuracy is about 48.69%. The effectiveness of the decision tree algorithm can be still improved by adopting certain feature selection techniques with the combination of decision tree algorithm. Conclusion: The determined results show that Decision tree algorithm suits well for medical data problems. The efficiency of the algorithm can still be improved by applying Decision Trees for various real world data problems such as Diabetes, Cancer classification with feature selection paradigms. But still a larger set of real world data has to be investigated.

Keywords: C4.5 Decision Tree Algorithm, Data Classification, Heart Disease, Predictive Analysis, Weighted Decision Tree

1. Introduction

Data mining is to collect the useful patterns from the observed data, the collected information in the form of concepts, law, rules, and so on. It helps the business analyst to analyses the ancient data and present data to find the concealed relationships between the data which is useful to predict the future behavior of the organization. There are several data mining techniques available. They are data classification, association rule mining, data clustering, regression analysis, etc., Data classification and Prediction are the important methods in Data Analytics that can be used to develop data models which describes about the importance of the data. Classification algorithms are Support Vector Machines (SVM), Decision Tree, Naive Bayes, Neural Networks (NN) and Random Forest. Among these algorithms, Decision Trees has been used for predictive analysis over medical data over

*Author for correspondence

the determination of risk factors which signifies to the locality, likelihood and dietary habits of the people over different locations. It is a graphical model which describes decision and their possible outcomes. They allow for intuitive understanding of the problem and can aid in decision making. Thereby experts decision can make a sustainable inference in discerning the probability, cost that incur with the specified disease.

Data classification plays a vital role in predictive analysis. The problem is whether classification is made under the mechanism of binary or multi class classification problem. The development of the predictive model in data analysis focuses majorly towards the correct classification of instances over the set of determined data tuples. Hence the presence and absence of the disease will be determined for each case. Decision tree is one of the decision support tool which is a graphical structure most widely used up for predicting decisions and the possible consequences. This is one of the widely used algorithms for decision analysis. Each of the internal nodes determines the test on the attribute with its preceding branches and the tree terminates with a leaf node. In this research work we have formulated decision tree algorithm with the assignment of weights to the nodes in training the instances also with the evaluation of its variants over medical data[2].

The evaluation implies with the generation of the target attribute with the developed classification model. The data set will be examined in the form of training and testing data tuples over cross validation. The result provides the best split among the class labeled tuple with the input label values. The total number of edges will be equivalent to the set of defined input attributes. The path from the root node to the leaf node contains the target values over the end of each split of the leaf node, which again determines the end of the split branch over the attribute[3].

Decision trees have numerous variants starting up from ID3, CART, C4.5, C5.0 and weighted decision trees. The main advantage of using decision trees is it can be converted easily to rules for better prediction. The mechanism of discovered patterns can be visualized effectively using decision trees. During the classification process at the initial stage i.e. mapping process decision trees can be efficiently used over the function which best separates the tupes accordingly to the data classes. The attribute values of the unknown estimate can be determined by testing the data tuples accordingly. The development of decision trees does not require any process of parameter estimate over the analysis process[4].

2. Problem Formulation

In general, decision trees assumes that the data instance relates to its crisp class, but this can't be rendered when analysis over real-world problems which utterly falls over noise in the data. Meanwhile, the training data instance can't be always allotted to its exact labeled value. When performing data analysis over medical data the important factor to be considered is its time complexity. In examining analysis over medical data the performance of the algorithm, its accuracy and the execution time are the key factors that have to be considered to be important. Data set with multiple labeled values also to be taken to be the important concern over the performance of the algorithm. If the data records contains categorical data values then ID3 is the most suited algorithm and if the data records contains continuous values algorithms such as WDT, C4.5 and C5.0 can be used up for analysis. Hence for medical data decision trees are considered to be the important predictive algorithm for the determination of risk over the disease specified. This work highlights the performance of decision tree and its variants over the medical data with the risk determination and execution in terms of accuracy.

3. Literature Survey

The proposed work towards the development of improved ID3 algorithm made a novel method for weighting the attribute weighting mechanism. The method provided the formulation of conditional probability measures among the attributes and decision attributes. The experimental results prove that the improved method has higher predictive accuracy than the traditional ID3 algorithm with less number of leaves from the root node. The proposed work maximized the mutual ranking between candidate and decision attribute and minimized the information between candidates to the conditional attribute. The entire mechanism used information gain to be the splitting measure towards the improvement over the generation of attributes[3]. Ensemble methods in data mining play a vital role in position estimation over mobile devices for indoor localization mechanism.

The system works with multiple decision trees parallel with weights assigned over it. Based upon weight values the device location with localization has been developed with sensor combination. The experimental result proves that the proposed work provides a substantial improvement over the existing finger print techniques[4].

The mechanism of discovering the weights over the training instances provides a factual impact on data analytics. The workflow follows the way of emerging patterns in which the measure of probability over one set of class tuples is significantly higher than that of other class labeled tuples. In the first step, scores has been assigned over the instances according to the relationship with other classes. In the second step, assign the class values which have highest score for the specified instance. Generally ignore the relationship between the classes and the instances whose score is found to be lesser. For example, assume that the scores of an instance in a two-class data set are 51 and 49 percent for classes C_1 and C_2, respectively. Then, the experts will assign class C_1 to this instance despite the fact that class C_2 also has a strong relation with this instance[5].

The improvement over weighted classification provides an easier way to perform data classification over medical data. The usage of top down induction of decision trees and association rule mining techniques provides an improved platform to make data classification and prediction. The experimental results shows that the model makes a better understanding over medical data with attribute relevant information. The weighted classification provides different levels of degrees to understand the levels of each class labels which has been assigned over data tuples[6,7].

The work by the authors based on random forest algorithm over medical data provides an improvement over accuracy for heart disease prediction. The evaluation of metrics such as accuracy, precision, recall, kappa statistics shows an improvement over other classification techniques[7]. The combination of decision trees with swarm intelligence techniques provides a significant improvement over accuracy also in the determination of risk factors over the medical data. The results shows that the combination of decision trees and swarm intelligence techniques has been found to be 60.74% over other techniques such as step wise forward selection and step wise backward elemination[8].

4. Dataset Description

The data records of various heart disease patients are taken from the UCI Data Repository. This data set is retrieved from Cleveland Clinic Foundation[17,18]. The dataset contains 298 records with 13 attributes. The Table 1 has all the description about the dataset[8,9].

Table 1. Description of the dataset

S. no	Features	Description
1.	Age	Age in years
2.	Sex	Sex (1 = male; 0 = female)
3.	Cp	Chest pain type (typical, atypical, non-anginal, asymptomatic)
4.	Trestbps	Resting blood pressure
5.	Chol	Serum cholesterol in mg/dl
6.	Fbs	Fasting blood sugar > 120 mg/dl
7.	Restecg	Resting electrocardiographic results
8.	Thalach	Maximum heart rate achieved
9.	Exang	Exercise induced angina
10.	Oldpeak	ST depression induced by exercise relative to rest
11.	Slope	The slope of the peak exercise ST segment (up sloping, flat, down sloping)
12.	Ca	Number of major vessels (0-3) colored by fluoroscopy
13.	Thal	Normal, Fixed and reversible effects

Similarly data records corresponding to Coronary Heart Disease patients has been collected from a hospital which consists of 306 medical records pertaining to 24 attributes with the distinguishing of labels from initial stage to final level of heart attack.

5. Methodological Workflow

5.1 WDT Flow Chart

The flow of the Weighted Decision Tree (WDT) technique for the attribute selection process is depicted in the Figure 1. The design of the system is depicted in the Figure 2.

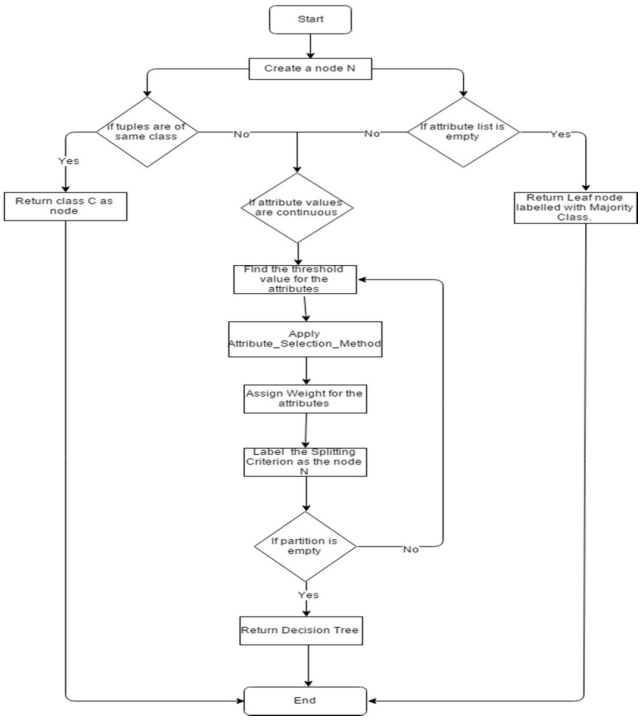

Figure 1. Flow chart for weighted decision tree.

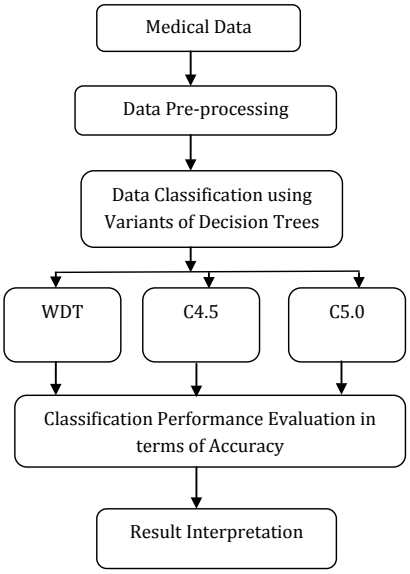

Figure 2. MDesign of the system

5.2 Splitting Measures

5.2.1 Information Gain

Information gain is one of the methods for an attribute selection which is used to find the root node for a dataset. It is a measure which is used to reduce the depth of the tree. For finding the information gain we have to find out a measure called as the entropy. Entropy is used to find out the similarity of the given dataset. The formula for finding the entropy is given as follow:

$$Entropy(s) = \sum_{i=1}^{c} -P_i \log_2 [(P_i])$$

Where,
Entropy(s) = entropy of the overall dataset.
p= probability of an arbitrary tuple.
i= level of the target attribute Han and Kamber [2012].

Information gain for an attribute can be calculated by using the following formula:

$$Gain(S, A) = Entropy(S) - \sum_{Values(A)} \frac{|S_i|}{|S|} Entropy(S_v)$$

Where,
Gain(S,A)= information gain for the defined attribute in the dataset.
v= values that are present in an attribute.
Repeat the process for finding the information gain for each attribute and the attribute which is having highest information gain will be selected as the root node[19,20].

5.3 Weighting the Training Instances

Our approach builds decision trees by weighting the training data instances. Assigning weights to the training instances is the core of weighted decision tree model. Our trivial method to apply weight to the training instances by finding the probability of the occurrence of the class label for each column present in the dataset[10,11]. The formula for finding the weight of the training instance is given as follows:

$$WDT\ (S, A) = P_i * Gain(S, A)$$

WDT(S,A)=Weight of an each attribute.
P_i=Probability of the occurrence of class label.

5.4 Performance Evaluation

Since the observed dataset is of multiclass labeled, with labels ranges from zero to four for all the levels of heart disease specification. The confusion table will be exactly viewed in the form of matrix like representation. The representation for the observed class level is depicted in Table 2.

Table 2. Confusion matrix

		Class Predicted				
		Class zero	**Class one**	**Class two**	**Class three**	**Class four**
Current class	Class zero	133	45	28	25	10
	Class one	4	0	1	2	0
	Class two	20	9	5	8	4
	Class three	1	0	1	1	0
	Class four	0	0	1	0	0

Accuracy - It refers to the number of data records that has been correctly classified. From Table 2 it has been observed that the diagonals correspond to the data records that signify correct classification. The set of misclassified records frames the rest of the cells in the confusion matrix.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Where,

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

Classification Error - refers to the number of data records that are wrongly classified.

$$Classification\ error = \frac{TP + TN}{TP + FP + FN + TN}$$

6. Results and Discussion

6.1 C4.5 Algorithm

In this model when the heart diseases dataset with 13 attributes are given as input at the end it provides an accuracy of 71.42% and we got than as the root node[15,16]. The C4.5 decision tree algorithm performed well with its defined parameters. The splitting measure used is information gain over the defined attributes[12]. The following Figure 3 describes about the generated decision tree. Similarly, for the real dataset used the accuracy was found to be 48.69% but still this is lower than that of the threshold limit in which a greater combination of algorithmic model can enhance the improvement of accuracy over the defined threshold limit. The following Table 3 describes the observed values against the root node with accordance to the labeled attribute. The standard coefficient value with accordance to the root node is depicted in Figure 4 for C4.5 Decision tree algorithm[21].

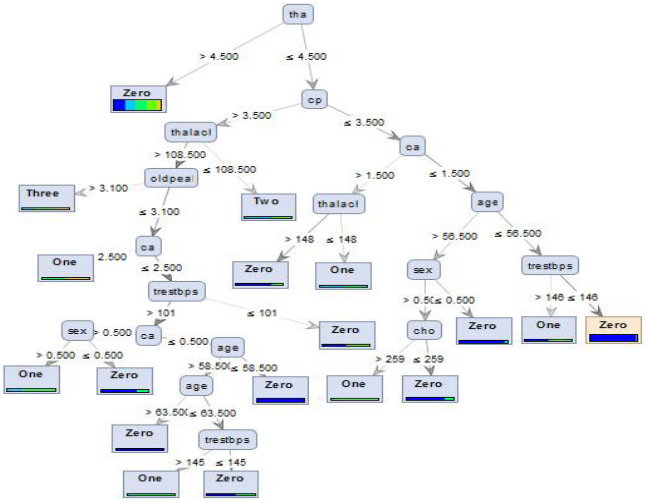

Figure 3. Generated decision tree using C4.5 Algorithm

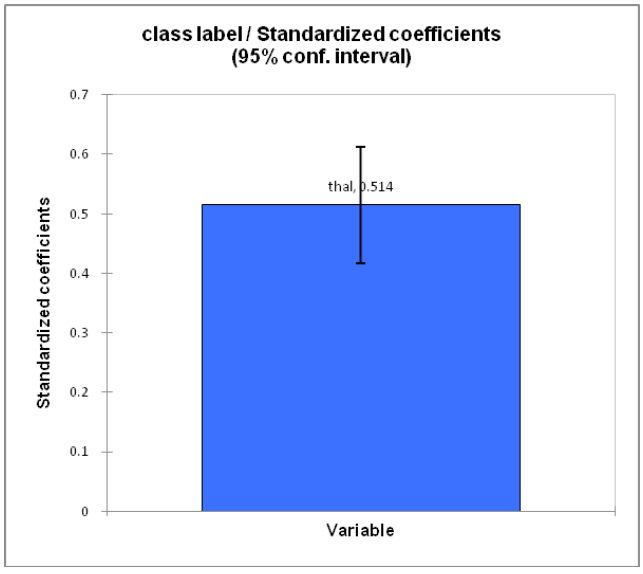

Figure 4. Determination of standardized coefficient value against the attribute thal for C4.5 Algorithm.

Table 3. Observed data values against the root node for UCI dataset

Source	DF	Sum of squares	Mean squares	F	Pr> F
Model	1	119.648	119.648	106.552	< **0.0001**
Error	296	332.382	Mean Sq (error)=1.123		
Corrected Total	297	452.030			

6.2 Weighted Decision Tree Algorithm (WDT)

In this model when the heart diseases dataset with 13 attributes are given as input at the end it provides an improved accuracy of 50.22% and we got rectecg as the root node and for real data 46.20%. The improvement of decision trees with the assignment of weight values is not suitably fitted over the heart disease data. This is because the inferred weights may be abidingly insignificant over each level of the tree generation process. Hence the delivered model may be suitable with some other medical data sets by which prediction and classification can be made adherently better[13,14]. The following Table 4 describes the experimentation of the root node with accordance to the class label. The standard coefficient is depicted in Figure 5 for the attribute restecg.

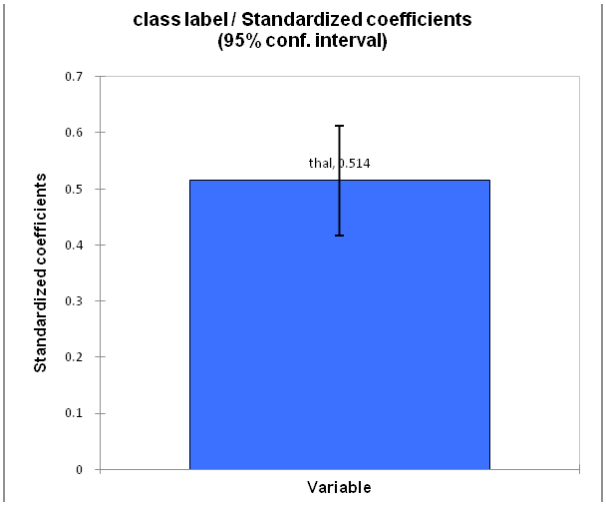

Figure 5. Determination of standardized coefficient value against the attribute Restecg for WDT Algorithm

The observed R^2 value for the root node to be restecg is 0.035 which is smaller than that of the attribute value than which ranges about 0.265 and hence for the case C4.5 can be deployed more probable predictive measures over medical data analysis.

6.3 C5.0 Algorithm

The evaluation has then been determined using C5.0 algorithm the model chooses the same thal attribute as the root node. But the accuracy level is slightly lower than that of the C4.5 decision tree algorithm of about 65.29%. The evaluation results also provided the same set of attributes consideration for all the lower levels for the generated decision tree. When considering the real world data the accuracy level was also still lower than the threshold of about 48.32%

6.4 Accuracy Comparison

The Table 5 and table 6shows accuracy and error rate obtained by both C4.5 and WDT Algorithms. It proves that the proposed Weighted Decision Tree Algorithm has lower rate of accuracy when compared to C4.5 algorithm. The classification also very low for Weighted Decision Tree when compared to the C4.5 algorithm

Table 5. Accuracy comparison of C4.5, C5.0 and WDT for the Benchmark dataset

Algorithm	C4.5	WDT	C5.0
Accuracy	71.42%	50.22%	65.29%
Classification Error	18.58%	49.78%	29.89%

Table 4. Observed data values against the root node for UCI dataset

Source	DF	Sum of squares	Mean squares	F	Pr> F
Model	1	15.676	15.676	10.634	**<0.001**
Error	296	436.354	Mean Sq (error)=1.474		
Corrected Total	297	452.030			

Table 6. Accuracy comparison of C4.5, C5.0 and WDT for the Real world dataset

Algorithm	C4.5	WDT	C5.0
Accuracy	48.69%	46.20%	48.32%
Classification Error	51.31%	53.80%	51.68%

6.5 Statistical Analysis

The statistical data analysis has been inferred using regression analysis in which all the data records corresponding to the class label has been taken for consideration. Regression analysis provides the mechanism of estimating the relationship among the variables and to analyze the importance of each variable. Here the dependent variable class label is examined over the several independent variables. The efficacy of regression analysis over the medical data is given in the following Table 6.

From the observed set of p-values it has been found that chest pain type, old peak, restecg, thalach, exang, thal and ca values are observed to be more significant in which the p-value is less than 0.05 which again states that these variables have dependency among one other with its significance state. Hence the above attributes can be considered for evaluation in the determination of risk that corresponds to heart disease prediction. With the real world dataset the attributes may correlate with its relevant to some other risk factors that corresponds to the disease. The risk of the signified disease may vary with accordance to the likelihood, location and dietary habits of the people living at a particular area. Meanwhile the scalability, the algorithmic execution time and its performance should also be taken into a concern in the determination of risk related to the corresponding disease.

7. Conclusion and Future Work

Thus the proposed work was mainly concerned with the mechanism of comparing the efficiency of decision tree algorithm and its variants. The workflow for weight based decision tree has been elaborated in addition to C4.5 and C5.0 Decision tree algorithm. The assignment of weight in WDT has been done by assigning weight to the attributes which was present in the dataset. Experimental results have been carried out for the dataset taken from UCI machine learning repository and real world data which corresponds to heart disease. The versions of C4.5 algorithms will be compatible corresponding to the application that we choose. The decreased performance of weighted C4.5 algorithm over the selected data is not observed up to the level. Thus the C4.5 algorithm can be used for evaluating and predicting various other kinds of diseases in medical science such as diabetes, cancer, and it can also be applied at any real world problems where the data mining plays a major role in increasing the efficiency. From the observed results it has been found that C4.5 decision tree algorithm provides an improved accuracy of about 71.42% when compared to WDT and C5.0 Decision tree algorithm.

The future work can be preceded with the usage of Real world medical data problem corresponding to vari-

Table 6. Results inferred over regression analysis

Attributes	Coefficients	Standard Error	t Stat	P-value
age	-0.008411996	0.006498	-1.29448	0.196551
sex	0.182125839	0.116945	1.557361	0.120498
cp	0.195389172	0.057326	3.408414	0.000748
trestbps	0.004347845	0.002975	1.46157	0.144965
chol	0.000210985	0.000987	0.213854	0.830814
fps	-0.082867444	0.141762	-0.58456	0.559311
restecg	0.099212779	0.050443	1.966826	0.050177
thalach	-0.005716784	0.002691	-2.12421	0.034517
exang	0.227894166	0.120244	1.89526	0.059073
oldpeak	0.181421043	0.054746	3.313844	0.00104
slope	0.162900708	0.10096	1.61351	0.107744
ca	0.433595338	0.059554	7.280695	3.27E-12
thal	0.140559358	0.030563	4.599075	6.4E-06

ous other diseases such as diabetes, cancer, and kidney disease. Real world problems in data analysis and classification have various forms of challenging and curious problems. To solve and handle those problems different strategically approaches various data analysis approaches has to be identified with accordance to the type of data. Data optimization with classification can provide more improved results with respect to improved classification and prediction accuracy. Data optimization techniques such as Particle Swarm Optimization, Ant Colony Optimization, Harmony Search, Genetic algorithm can be incorporated with data classification techniques.

8. References

1. Seele P. Predictive sustainability control: A review assessing the potential to transfer big data driven 'predictive policing' to corporate sustainability management. Journal of Cleaner Production. 2016. doi:10.1016/j.jclepro.2016.10.175
2. Thomas H, McCoy, Snapper L, Theodore BS, Stern A, Roy H, Perlis. Underreporting of delirium in statewide claims data: Implications for clinical care and predictive modeling. Psychosomatics. 2016; 57(5):480-8 doi:10.1016/j.psym.2016.06.001
3. Liang X, Qu F, Yang Y. An improved ID3 decision tree algorithm based on attribute weighted. Material and Environmental Sciences. 2015; 8:234-46.doi:10.2991/cmes-15.2015.167
4. David Son, Giri T. Data preparation using data quality matrices for classification mining. European Journal of Operational Research. 2010; 197(2):764-72. doi:10.1016/j.ejor.2008.07.019
5. Alhammady H, Ramamohanarao K. Using emerging patterns to construct weighted decision trees. IEEE Transactions on Knowledge and Data Engineering. 2006; 18(7):865-76. doi:10.1109/TKDE.2006.116
6. Wenguang J, Lijing H. Improved C4.5 decision tree. IEEE Transactions on Evolutionary Computation. 2014; 18(6):4-7. doi:10.1109/ITAPP.2010.5566133
7. Sheik Abdullah A. A data mining model for predicting the coronary heart disease using random forest classifier. International Journal of Computer Applications. 2012; 3:973-93-80867-33-2.
8. Sheik Abdullah A. A data mining model to predict and analyze the events related to coronary heart disease using decision trees with particle swarm optimization for feature selection. International Journal of Computer Applications. 2012; 55(8):973-93-80870-77-4. doi:10.5120/8779-2736
9. Parthiban P, Selvakumar S. Big data architecture for capturing, storing, analyzing and visualizing of web server logs. Indian Journal of Science and Technology. 2016 Jan; 9(4):1–9. doi:10.17485/ijst/2016/v9i4/84173
10. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Journal on Neural Computation. 2013; 10(4):432-38. doi:10.1162/089976698300017197
11. Bhagwatkar P, Parmalik Kumar K. Improved the classification ratio of C4.5 algorithm using attribute correlation and genetic algorithm. International Journal of Advanced Computer Engineering and Communication Technology. 2014; 3(2):2319-526.
12. Alhammad H, Ramamohanarao K. Using emerging patterns and decision trees in rare-class classification. IEEE International Conference on Data Mining; 2004. doi:10.1109/ICDM.2004.10058
13. Selvakumar S, Sheik Abdullah A, Suganya R. Decision support system for type II diabetes and its risk factor prediction using Bee based harmony search and Decision tree Algorithm. International Journal of Biomedical Engineering and Technology. INDERSCIENCE. (In Press).
14. Zhao H, Li X. A cost sensitive decision tree algorithm based on weighted class distribution with batch deleting attribute mechanism. Elsevier. 2017; 378(1):303–16. doi:10.1016/j.ins.2016.09.054
15. Luo H, Chen Y, Zhang W. The application ofemerging patterns for improving the quality of rare-class classification. 2nd International Workshop on Database Technology and Applications. 2010. doi:10.1007/978-3-540-24775-3_27
16. Chen F, Li X, Lixiong. Improved C4.5 decision tree algorithm based on sample selection. 4th IEEE International Conference on Software Engineering and Service Science; 2013. doi:10.1109/ICSESS.2013.6615421
17. Sitanggang IS, Yaakob R, Mustapha N, Nuruddin AAB. An extended ID3 decision tree algorithm for spatial data. IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services; 2011. doi:10.1109/ICSDM.2011.5969003
18. Sheik Abdullah A, Selvakumar S, Karthikeyan P, Mahesh M, Deepchand PK. An efficient prediction model using multi swarm optimization empowered by data classification for Type II diabetes. 3rd International Conference on Business Analytics and Intelligence (ICBAI); Bangalore. 2015.
19. Liu Y, Hu L, Yan F, Zhang B. Information gain with weight based decision tree for the employment forecasting of undergraduates. IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing; 2013. doi:10.1109/GreenCom-iThings-CPSCom.2013.417
20. Han J, Kamber M. Data Mining Concepts and Techniques. 3rd Ed. 2012.
21. Karaolis M, Moutiris JA, Pattichs L. Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Transactions on IT in Biomedicine. 2010; 14(3). doi:10.1109 TITB.2009.2038906