Intelligent Automation And Soft Computing, 2020

Vol. 26, no. 1, 61–70
DOI: 10.31209/2019.100000153

C5.0 Decision Tree Model Using Tsallis Entropy and Association Function for General and Medical Dataset

Uma K.V[1], Appavu alias Balamurugan S[2]

[1]Department of Information Technology, Thiagarajar College of Engineering, Tiruparankundram, Madurai-625015, Tamilnadu, India.
[2]Research Director and Professor, Department of Computer Science and Engineering, E.G.S.Pillay Engineering College, Nagapattinam, Tamilnadu, India.

ABSTRACT
Real world data consists of lot of impurities. Entropy measure will help to handle impurities in a better way. Here, data selection is done by using Naïve Bayes' theorem. The sample which has posterior probability value greater than that of the threshold value is selected. C5.0 decision tree classifier is taken as base and modified the Gain calculation function using Tsallis entropy and Association function. The proposed classifier model provides more accuracy and smaller tree for general and Medical dataset. Precision value obtained for Medical dataset is more than that of existing method.

KEYWORDS: Data Mining, Association Function, Classification, Decision Tree, Entropy

1 INTRODUCTION

CLASSIFICATION techniques are considered to be the most important Data Mining functionalities. It is called as a supervised learning technique since it contains class label for training the model. There are different Classification techniques. Davis, et al. (2006) proposed a cost sensitive decision tree learning algorithm. Claesan, et al. (2014) developed an Ensemble SVM. Bobadilla, et al. (2013) proposed a recommender system based on k-nearest neighbors. One of the initial machine learning approaches that were successful till now is a Decision tree classification technique. This technique remains as a good method till now in machine learning for its simplicity, interpretability, efficiency and flexibility. Some of the Decision tree algorithms are ID3, CART, C4.5, C5.0 etc. These techniques are widely applied to variety of task. Imai, et al. (2017) used decision tree model for analysis of adverse drug reactions. Hunt (1993) used classification by induction model for control of nonlinear dynamical systems. Attigeri, et al. (2017) used Machine learning algorithms to detect credit risk of loan applicants. Decision tree algorithms uses entropy and Gain measures to determine the important attributes in a dataset .The most important attribute form the root node of the decision tree which is considered as the best predictor. Entropy is a measure that is used by the Decision tree algorithm that is used to identify the homogeneity of a sample. The calculated value of entropy will be zero if all the samples in the dataset are homogeneous and it will be one when the samples are equally classified. The best splitting attribute of tree can be identified using Information Gain measure. The attribute that have highest value of Information gain forms the root node of the Decision tree.

During Decision tree induction, identification of split criterion and tree construction is the two primary issues that need to be handled effectively. Some of the Decision tree algorithms such as Iterative Dichotomiser3 (ID3) algorithm use Shannon entropy and Gain ratio to determine the split of the tree. Similarly, C4.5 algorithm uses Gain Ratio and Gini index is used by Classification and Regression Tree (CART) algorithm as a Split criterion. It's not always the split criteria identified through these measures will suit all datasets. All these measure are based on entropy. C5.0 is an extension of C4.5 algorithm.C5.0 algorithm is easy to understand and more robust (although the dataset is large and has missing value).It requires less training time to build the model. It is a powerful boosting method with improved classification accuracy. Here, a Decision tree induction method is proposed which is based on the C5.0 algorithm and with different types of entropies.

CONTACT K.V.Uma kvuit@tce.edu

2 RELATED WORK

ENTROPY measure is used to calculate the randomness or uncertainty in a given data. There are different types of entropies such as Shannon entropy, Renyi entropy, Tsallis entropy etc. Shannon entropy is greatly associated with random variable X. According to Shannon, entropy of a discrete random variable X with possible values $\{x_1 \ldots x_n\}$ and probability mass function P(X) is defined as given in eqn (1).

$$H(X) = E(I(X)) = E[-ln(P(X))] \quad (1)$$

where E is an operator that defines expected value and in I(X), I defines information content of random variable X. Then the entropy H(X) can be written as given in eqn (2)

$$H(X) = \sum_{i=1}^{n} P(x_i) I(x_i) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) \quad (2)$$

where b defines the base of the logarithm. Murphy (2012) proposed that classification algorithms can be applied to data across variety of domains with heavy tailed distribution .i.e. tends to have very large values with many outliers. For these types of data, probability value will be high. Heavy tailed distribution cannot be handled by maximizing Shannon entropy. Entropy with powers of probability will have such control. Tsallis entropy has the powers of probability. Tsallis entropy $S_q(X)$ is the generalized form of Shannon entropy with adjustable parameter q. Hence Tsallis entropy is defined as given in eqn (3).

$$S_q(X) = \frac{1}{1-q}\left(\sum_{i=1}^{n} p(x_i)^q - 1\right), q \in R \quad (3)$$

Here X denotes the random variable having value $(x_1, x_2, \ldots\ldots, x_n)$ and $p(x_i)$ is used to define the probability of occurrence of x_i. Renyi Entropy is an another type of entropy defined mathematically as in eqn (4)

$$H_\alpha(X) = \frac{1}{1-\alpha} log\left(\sum_{i=1}^{n} p^{\alpha}{}_i\right) \quad (4)$$

where, the discrete random variable is denoted with X with n number of outcomes. And p_i refers to the probabilities for the values of i from 1 to n. And α is an inherent parameter which can be used to make it more or less sensitive to the shape of probability distributions. The attribute importance can be computed using different methods. Engelbrecht (2001) used sensitivity analysis to train the Neural network.

But it increased the computational complexity. Kwak and Choi (2002) proposed joint information entropy method. But this method does not suit for continuous numerical value. Jin, et al. (2009) proposed a function called as Association function to represent the relations between all elements and their corresponding attributes in a given dataset. If there is an attribute A in dataset D and class C is the category attribute ,association function between attribute A and Class C can be defined in eqn.(5)

$$AF(A) = \frac{\sum_{i=1}^{n} |x_{i1}| - |x_{i2}|}{n} \quad (5)$$

where x_{ij} denotes the attribute A of dataset D that is having the i^{th} value, the category attribute or class attribute C takes the j^{th} value, where n is the number of values that an attribute A can have. In order to do the normalization, for a dataset that is having m number of attributes ,attribute relation degree is defined by AF(1), AF(2),…, AF(m) .By using this relation degree the normalization factor is defined as given in eqn(6).

$$V(k) = \frac{AF(k)}{AF(1)+AF(2)+ \ldots\ldots+AF(m)} \quad (6)$$

where k=1..m. Then this association function can be used during Gain calculation. It provides the information about how far the attribute contribute to the Class label. Wang, et al. (2017) proposed a two term Tsallis entropy Information Metric (TEIM) algorithm. In TEIM algorithm, best split criterion is determined using Tsallis conditional entropy. This algorithm follows two stages for the construction of decision tree. This algorithm reduces the greedy property of decision tree algorithm and it handles noisy data in an effective manner. Farid, et al. (2014) proposed two hybrid mining algorithms Hybrid DT (Decision Tree) and Hybrid NB algorithm. In hybrid DT algorithm, NB classifier is used in order to remove the noisy data present in the dataset before DT induction. And in Hybrid NB classifier, Decision tree algorithm is used for feature selection. Karabatak, (2015) proposed a new NB(Weighted NB) classifier which was used for Breast Cancer detection. Since all attribute cannot contribute equally during the calculation of Posterior probability, weight is assigned to each attribute and is used during posterior probability calculation. Gajowniczek, et al.(2016) proposed modified C4.5 algorithm which that uses Tsallis and Renyi entropy for Telecom churn problem. Since both of the entropies are based on the parameter α that adjust the entropy measure depending on shape of probability distribution. Su, et al. (2014) proposed K-L divergence-based decision tree (KLDDT) for handling the dataset that has class imbalance problem. KLDDT along with SMOTE provides better result in the diagnosis of chronic obstructive pulmonary disease. Dan, et al. (2015) reduced the NIR spectra data of Orange growing locations by PCA and the important features were selected by attribute selection method. Then the subset of features was applied to the different classification algorithm. These proved that NIR spectra data were more suitable to detect the Orange growing location.

3 METHODOLOGY

AFTER the survey of the works related to usage of various entropies instead of Shannon entropy in

algorithms like ID3, C4.5 and CART, classifier model is proposed by replacing the entropies like Renyi, Tsallis entropy in C5.0 algorithm. Along with this, Association function is included with these entropies. The proposed method is developed based on NB and C5.0 classifier. Consider a training dataset D with n instances specified as D= $\{x_1, x_2, \ldots, x_n\}$.Training data in the dataset is represented as $x_i = \{x_{i1}, x_{i2}, \ldots, x_{ih}\}$. Different attributes in the dataset is defined as $\{A_1, A_2 \ldots A_n\}$. Every attribute A_i in the dataset contain values specified as $\{A_{i1}, A_{i2}, \ldots, A_{ih}\}$. The instances in the training data belongs to any one of the class attribute specified in the set $C = \{C_1, C_2, \ldots, C_m\}$.Then the posteriori hypothesis $P(C_i|X)$ of the class label conditioned on X is calculated using the Bayes theorem as given in eqn.(7)

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (7)$$

Here $P(C_i)$ denotes class prior probability. It is calculated as $P(C_i) = |C_i,D|/|D|$ where $|C_i,D|$ represents the total number of instances belonging to class C_i in D. Dataset may contain the attributes that are conditionally independent to one another. The following equations eqn.(8) and eqn.(9) are used to compute $P(X|C_i)$.

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (8)$$

$$P(X|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i) \quad (9)$$

The probability $P(x_1 \mid C_i), P(x_2 \mid C_i), \ldots, P(x_n \mid C_i)$ are computed for every instance in the dataset. Here x_k refers to attribute value A_k for instance X. Then the values are substituted in $P(C_i \mid X)$. Posterior probability value is calculated for every instance. If it is greater than the threshold value, then that instance is selected. It depends on the distribution of data. If all the attributes in the dataset contribute equally to the class label, then more number of instances will be selected. C5.0 algorithm is used to build decision tree. As an initial step, choose one of the attribute from given dataset and make it as a root node of the decision tree. Then make branch under the root node to form a decision tree using the remaining attributes in the training set. The tree is grown recursively using the training instances until the stopping criteria are met. Thus the algorithm proceeds. In C5.0 algorithm each attribute in the tree is selected based on information gain values. The attribute with the highest information gain forms the root node of the tree. Entropy is used to obtain this information gain. Consider a dataset S that belong to n different classes. The entropy, a measure of impurity represented by E(S).At this step, the different entropies such as Renyi, Tsallis, are applied and the Information gain is calculated. Information gain is calculated by using the eqn. (10)

$$Gain(A) = E(S) - \sum_{k=1}^{n} E(Sk) \frac{|Sk|}{|S|} \quad (10)$$

Here E(S) is the entropy for the whole dataset. And $E(S_k)$ denotes the average entropy value obtained with each subset of data. $E(S_k)$ denotes the entropy value of each subset computed by multiplying with $|S_k|/|S|$. And Gain(A) is the information quantity measure that the attribute A in dataset offers for classification. In the proposed model, the entropy calculation is done with the different entropies such as Tsallis entropy, Renyi entropy. Association function is used along with the calculation of Information gain. Then normalization relation degree function is calculated. The modified gain Gain'(A) can be defined as given in eqn.(11).

$$Gain'(A) = E(S) - \sum_{k=1}^{n} E(Sk) \frac{|Sk|}{|S|} * V(A) \quad (11)$$

Gain value is computed for every attributes in the dataset.C5.0 algorithm incorporates facilities for providing variable misclassification cost. In C4.5, all errors are treated as equal. But in real, some classification errors are more serious than others. C5.0 allows a separate cost for every predicted/actual class pairs thus it reduces misclassification rate. The property of C5.0 algorithm is that it effectively handles more number of attributes.

3.1 Proposed C5.0 algorithm with different Entropy (Pseudo code)

THE steps involved in the proposed work are given below.

Input: D (Dataset)= $\{x_1, x_2, \ldots, x_n\}$; A(Attributes)= $\{A_{11}, A_{12}, .., A_{nm}\}$;
Output: T, Decision Tree
Method:
1: for each classC_i, of D, **do**
2: Determine the prior probabilities, $P(C_i)$.
3: **end for**
4: **for** each attribute value, A_{ij} of D, **do**
5: Determine the class conditional probabilities, $P(A_{ij}|C_i)$.
6: **end for**
7: **for** every instance x_i of dataset D, do
8: Determine the posterior probability, $P(C_i|x_i)$
9: **if** Posterior Probability >Threshold value**, do**
10. Keep x_i in the D
11. **else**
12 Remove x_i from D
13. **end if**
14. **end for**
15: Tree={}
16: Tree = **Generate_Decisiontree (D, attribute_list)**
17: return Tree

1:**Generate_Decisiontree (D, attribute_list)**
2: Create a node N
3: **if** all the tuples in Dataset D belongs to same class, C **then**

4: **return** N by labelling the leaf node with the class C;
5:**endif**
6: **if** attribute list is empty **then**
7: **return** N by labelling the leaf node with the majority class in D;
8:**endif**
9: a_{best} **= C5.0 (D)**
10: N = Add a_{best} as a root node
11: Eliminate a_{best} from attribute_list
12:D_v = Induced Sub-datasets from D based on a_{best}
13: for all D_v **do**
14: **if** $D_v$is empty **then**
15: Insert a leaf node to the tree labelled with the majority class in D to N;
16: **else**
17: Insert the node returned by **Generate_Decisiontree(**D_v**, attribute list)** to node *N*;
18: end for
19: return N

1: **C5.0(D)**
2: for all attribute A Ɛ D **do**
3: Compute Renyi Entropy/ Shannon Entropy / Tsallis Entropy
4: Compute Information gain criteria with association function
5: e**nd for**
6: a_{best} = maximum (Information Gain' (a))
7: return a_{best}

From the first part of the algorithm, reduced dataset is obtained. Then it is applied to modified C5.0 algorithm which produces a decision tree.

4 EXPERIMENT AND RESULTS

THIS section describes the datasets considered, details of experimental environments, results of proposed and existing algorithm. The experiment was conducted in an Intel core(tm) machine of x86_64 architecture with CPU speed of 2.60GHz and 4GB RAM. A net bean IDE8.0 is used to implement the dataset reduction using Naïve Bayes in Java programming language. The experiment is conducted in two different ways.

1. Initially, Naïve Bayes approach is used for reduction of dataset. And the accuracy of different classifier executed with original and reduced dataset are compared.

2. Later, the proposed classifier model that uses different entropies are executed with original dataset and reduced dataset and the accuracy measures are compared with other classifiers.

4.1 Implementation of Naïve Bayes approach to obtain reduced dataset and performance comparison

ABOUT 9 real benchmark datasets are downloaded from UCI repository. Among these four are medical dataset. Class conditional probability of each instance is calculated and the instances are selected that have the value greater than the threshold value. These form the representative or reduced dataset. Following Table 1 shows the different dataset considered with the number of instances in original and reduced dataset.

Table 1. Dataset

Data Set	Number of Instances in Original Dataset	Number of Instances in Reduced Dataset
Iris	150	52
Glass	214	142
Soybean	307	141
Image	210	168
Vote	435	238
Medical Dataset		
Pima Diabetes	768	203
Liver	345	181
Kidney	351	181
Breast Cancer	699	389

Fixing threshold value for different dataset is a tedious process. For, some dataset samples will be reduced for lower threshold. For others high threshold need to be fixed. It's a trial and error method. Once the reduced dataset is obtained, it is applied to different classifier model using R tool with 10 fold cross validation and repeat count of 30. By reiterating the model for more number of times, model learns by itself and reconstructs and an efficient model is built. Thus the error rate is minimized. Here the classifier model C4.5, NB and C5.0 are chosen. For the entire classifier model, the accuracy for the reduced dataset is more than that of the original dataset. Table 2 shows the performance of different classifier model executed with original and reduced dataset.

The following figure 1shows the comparison C4.5, NB and C5.0 classifier accuracy with the original and the reduced dataset.

To prove the results statistically, Wilcoxon Signed rank test is performed. This is called as non-parametric test performed for paired samples. Signed rank test is based on the ranks of the absolute difference in the values of each pair. The null hypothesis is that the original and the reduced dataset are the same. To test the hypothesis, apply the Wilcox. Test function to compare the samples. Table 3 shows the p-value computed.

Table 1 Classifier model accuracy computed for Original (O) and reduced (R) dataset with 10 folds and 30 repeats

Dataset	C4.5 (O)	C4.5 (R)	NB (O)	NB (R)	C5.0 (O)	C5.0 (R)
Iris	96	100	96	100	94.6	96.6
Glass	65.88	75.35	85.10	91.85	77.08	78.9
Soybean	80.85	89.25	49.53	58.45	92.04	95.08
Image	89.04	92.26	78.095	88.09	91.8	94.6
Vote	96.31	96.83	90.11	97.73	96.04	97.16
Medical Dataset						
Pima Diabetes	73.82	76.35	76.30	79.31	80.98	84.67
Liver	63.47	66.85	55.07	69.06	69.9	73.3
Kidney	95.90	98.57	97.541	97.72	97.2	98.4
Breast Cancer	95.56	96.78	96.13	97.95	96.3	98.4

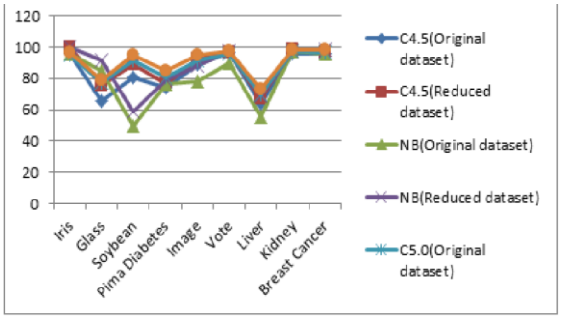

Figure 1 Comparison of accuracy obtained from classifier models on each dataset with 10 folds and 30 repeats

Table 3. p-values for the algorithm comparison

Models	p-value
C4.5 (Original and reduced dataset)	0.003906
Naive Bayes (Original and reduced dataset)	0.003906
C5.0 (Original and reduced dataset)	0.01953

In all cases, classifier model build with original and reduced dataset, the null hypothesis that both the model equivalent is rejected. It shows that the method is statistically significant.

4.2 *Implementation of Proposed Classifier Model using different Entropies*

The evaluation of proposed method is done with the help of testing on 9 real benchmark datasets that are obtained from UCI machine learning repository. After applying the Naïve Bayes method for finding the class conditional probability, the number of instances reduced from the original dataset is given in the Table 4. The reduced dataset contains the instances that are having posterior probability higher than that of the threshold value.

Table 4. Dataset

Name of the Dataset	Number of instances	Reduced number of instances
Car	1728	625
Tic-Tac-Red	723	434
Wine	177	132
Scale	625	568
Lenses	25	14
Blogger	100	73
Vote	435	238
Medical Dataset		
Haberman	306	251
Breast Cancer	700	686

Once the reduced dataset is obtained, it is applied to traditional C5.0 algorithm. Then, it is applied to proposed C5.0 classifier model using Tsallis entropy, Renyi entropy with 10 fold cross validation and repeat count of 30.Table 5 shows the performance of C5.0 classifier for the reduced dataset. Then C5.0 classifier model is modified using Tsallis entropy and Association function for the calculation of Information gain. Table 6 shows the performance of C5.0 classifier using new information gain measure that uses Tsallis entropy and Association function for different dataset. From the table, it is inferred that Precision value for Medical dataset is more than that of other methods. Then C5.0 classifier model is modified using Renyi entropy and Association function for the calculation of Information gain. Tsallis entropy is more efficient in handling the long tail distribution of the dataset. Table 7 shows the performance of C5.0 classifier using new information gain measure that uses Renyi entropy and Association function for different dataset.

Once the results are obtained using different methods, they are compared. Table 8 shows the evaluation result obtained through traditional C5.0 algorithm, and C5.0 classifier that uses Tsallis entropy and Association function, C5.0 classifier that uses Renyi entropy and Association function.

Table 5.Performance of traditional C5.0 Classifier with 10 folds and 30 repeats

Dataset	Accuracy	Precision	Recall	FMeasure	TN Rate	TP Rate	FP Rate	FN Rate
Car	97.6	0.9316	0.9320	0.9905	0.9904	0.9320	0.0094	0.0169
Tic-Tac-Red	97	0.9778	0.9456	0.9603	0.9778	0.9456	0.02213	0.0543
Wine	94.7	0.9472	0.9509	0.9489	0.9726	0.9509	0.02735	0.14961
Scale	83.2	0.57399	0.6035	0.58764	0.9069	0.6035	0.0930	0.39647
Lenses	82.9	0.78576	0.8485	0.8108	0.4059	0.8485	0.1096	0.15143
Blogger	80.5	0.7855	0.7509	0.7635	0.7855	0.7509	0.2145	0.2490
Vote	96	0.9581	0.9592	0.9586	0.9581	0.9592	0.0418	0.0407
Medical Dataset								
Harberman	73.3	0.6379	0.5986	0.6064	0.6379	0.5986	0.36205	0.4013
Breast Cancer	96.2	0.9571	0.9601	0.9586	0.9571	0.9601	0.0428	0.0398

Table 6. Performance of C5.0 Classifier using Tsallis Entropy with 10 fold cross validation and 30 repeats

Dataset	alpha value	Accuracy	Precision	Recall	FMeasure	TN Rate	TP Rate	FP Rate	FN Rate
Car	1.95	99.7	0.9849	0.9834	0.9840	0.9988	0.9834	0.0012	0.0165
Tic-Tac-Red	1.75	96.4	0.9437	0.9349	0.9074	0.9437	0.9349	0.0562	0.0655
Wine	2.15	96.4	0.9632	0.9642	0.9636	0.9823	0.9632	0.0176	0.0357
Scale	1.55	94.5	0.9467	0.9467	0.9468	0.9467	0.9467	0.0532	0.0532
Lenses	1.5	100	1	1	1	1	1	0	0
Blogger	1.50	90.9	0.8944	0.8537	0.8713	0.8944	0.8944	0.1055	0.1462
Vote	1.75	99.94	0.9994	0.9994	0.9994	0.9994	0.9994	0.0005	0.0006
Medical Dataset									
Harberman	1.95	79.1	0.6581	0.5721	0.5798	0.6581	0.6581	0.3418	0.4278
Breast Cancer	1.95	98.0	0.9782	0.9786	0.9784	0.9782	0.9782	0.0217	0.0213

Table 7.Performance of C5.0 Classifier using Renyi Entropy with 10 fold cross validation and 30 repeats

Dataset	alpha value	Accuracy	Precision	Recall	F Measure	TN Rate	TP Rate	FP Rate	FN Rate
Car	1.95	89.1	0.2227	0.25	0.2356	0.7209	0.25	0.0290	0.75
Tic-Tac_Red	1.95	81.5	0.4078	0.5	0.4492	0.4078	0.5	0.0921	0.5
Wine	2.15	40.4	0.1348	0.3333	0.3842	0.2824	0.1348	0.2824	0.666
Scale	1.55	49.6	0.4943	0.538	0.4545	0.4943	0.538	0.5056	0.5039
Lenses	1.25	95.2	0.6	0.6666	0.6296	0.9762	0.6666	0.0238	0.3333
Blogger	1.25	75.1	0.3750	0.5	0.4285	0.3750	0.3750	0.1250	0.5
Vote	1.95	53.7	0.2689	0.5	0.3497	0.2689	0.5	0.2311	0.5
Medical Dataset									
Harberman	1.95	79.2	0.3960	0.5	0.4419	0.3960	0.5	0.1040	0.5
Breast Cancer	1.95	65.1	0.325547	0.5	0.394341	0.325547	0.5	0.174453	0.5

Table 8. Comparison of Accuracy different entropies

Dataset	Traditional C5.0 with reduced dataset	C5.0 with TSALLIS for reduced dataset	C5.0 with Renyi for reduced dataset
Car	97.6	99.7	89.1
Tic-Tac_Red	97	96.4	81.5
Wine	94.7	96.5	40.4
Scale	83.2	94.5	49.6
Lenses	82.9	100	95.2
Blogger	80.5	90.9	75.1
Vote	96	99.94	53.7
Medical Dataset			
Harberman	73.3	79.1	79.2
Breast Cancer	96.2	98.0	65.1

From the Table 8, it is inferred that for almost all the dataset considered, C5.0 algorithm that uses Tsallis entropy along with the association function provides more accuracy. But, C5.0 algorithm with Renyi entropy gives lesser accuracy. Figure 2 show that that the proposed method provides more accuracy than other methods.

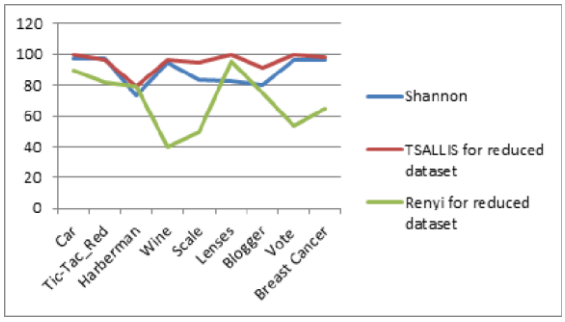

Figure 2 Comparison of Accuracy of Existing and Proposed classifier

To prove the results statistically, Wilcoxon signed rank test and Friedman test is performed. These two tests are conducted. In all the cases, the proposed method that uses Tsallis entropy along with Association function is proved to be significant. The statistical results are shown in Table 9.

Receiver Operating Characteristic curve (or ROC curve) is used to plot the trade-off between the true positive rate (tpr) and the false positive rate (fpr) at different possible cut points. Closer the curve towards top left-hand border, better the model. Otherwise the model is less accurate. Figure 3 shows the ROC for the proposed method with original dataset.

Table 9. p-value

Algorithms	Wilcoxon signed rank test	Friedman rank sum test
Actual Shannon entropy and reduced Tsallis entropy	0.02488	0.033895
Reduced Shannon entropy and Reduced Tsallis entropy	0.03461	0.014306

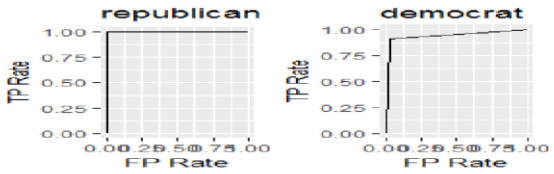

Figure 3. ROC for proposed method with original Vote dataset

Figure 4 shows the ROC for the proposed method with reduced dataset. The curve is very closer towards the left-hand border .Hence the proposed method is more accurate.

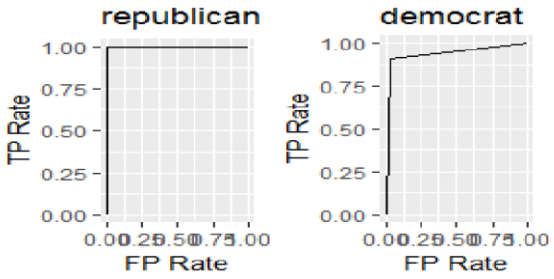

Figure 4. ROC for Proposed method with Reduced Vote dataset

Table 10 shows the comparison of TEIM algorithm proposed by Y. Wang et.al and the proposed method.

Table 10. Comparison of Classification accuracy obtained through SEIM, REIM, TEIM and Proposed method

Dataset	SEIM Y.Wang et.al	REIM Y.Wang et.al	TEIM Y.Wang et.al	Proposed C5.0 with Tsallis entropy	Proposed C5.0 with Renyi entropy
Wine	93.0	94.5	96.0	96.4	40.4
Haberman	74.5	74.7	75.2	79.1	79.2
Scale	78.8	79.6	82.2	94.5	49.6
Car	98.3	98.7	98.8	99.7	89.1

It shows that our proposed classifier algorithm that uses Tsallis entropy along with Association function has better accuracy than that of TEIM algorithm

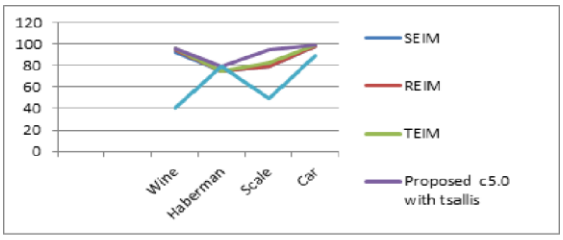

Figure 5 Comparison of SEIM, REIM, TEIM algorithm and proposed algorithm

Table 11 shows the comparison of Hybrid NB, Hybrid DT algorithm proposed by Dewan Md. Farid et.al and our proposed method.

Table 11: Comparison of Classification accuracy obtained through Hybrid NB, Hybrid DT and Proposed method

Dataset	Hybrid NB	Hybrid DT	Proposed C5.0 with Tsallis entropy	Proposed C5.0 with Renyi entropy
Breast Cancer	75.87	81.46	98	65.1
Contact Lens	87.50	91.66	100	95.2
Vote	94.48	97.70	99.94	53.7
Tic-Tac	78.91	88.1	96.4	81.5

It shows that the proposed classifier algorithm that uses Tsallis entropy along with Association function has better accuracy than that of Hybrid NB and Hybrid DT algorithm. Figure 6 shows the comparison of Hybrid NB, Hybrid DT and proposed method that uses Tsallis entropy.

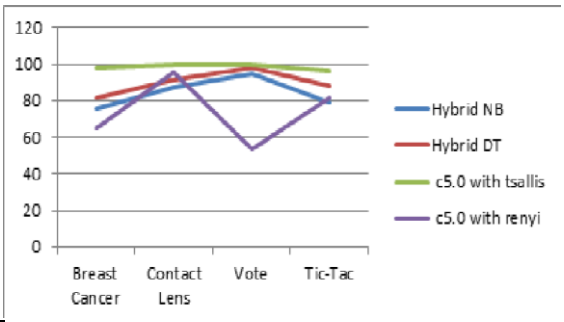

Figure 6 Comparison of Hybrid DT, Hybrid NB and Proposed Method

C5.0 algorithm modified using Tsallis entropy which is used in information gain calculation and association, tells the relation between the individual attributes and class label in an effective manner.

5 CONCLUSION

IN this work, C5.0 algorithm is modified by using different entropies such as Renyi entropy, Tsallis

entropy along with a mathematical function called as Association Function. And it has been proved that Tsallis entropy along with Association function has a better performance compared with Shannon Entropy used in C5.0 algorithm. Also the decision tree obtained by Tsallis Entropy is small in size in comparison with the decision tree obtained by Shannon entropy. Thus this method will help us to construct effective and efficient decision trees and also an effective C5.0 algorithm is being proposed instead of the previously existing algorithm. The proposed method works better for Medical dataset also. The decision trees generated will help to understand the characteristics of data better.

6 REFERENCES

Attigeri, G. V, Pai, M. M. M., & Pai, R. M. (2017). Credit Risk Assessment Using Machine Learning Algorithms. *Advanced Science Letters*, *23*(4), 3649–3653.

Avenue, N. (n.d.). Rule Extraction Based on Data Dimensionality Reduction Using RBF Neural Networks. *Electronic Engineering*.

Bobadilla, J., Ortega, F., Hernando, A., & Glez-De-Rivera, G. (2013). A similarity metric designed to speed up, using hardware, the recommender systems k-nearest neighbors algorithm. *Knowledge-Based Systems*, *51*, 27–34.

Breiman, L., Friedman, J., Stone, C. ., & Olshen, R. . (1984). *Classification and Regression Trees*.

Chen Jin, Luo De-lin, & Mu Fen-xiang. (2009). An improved ID3 decision tree algorithm. *2009 4th International Conference on Computer Science & Education*, 127–130.

Claesen, M., De Smet, F., Suykens, J., & De Moor, B. (2014). EnsembleSVM: A Library for Ensemble Learning Using Support Vector Machines, *15*, 141–145. Retrieved from http://arxiv.org/abs/1403.0745

Dan, S., Yang, S. X., Tian, F., & Den, L. (2015). Classification of Orange Growing Locations Based on the Near-infrared Spectroscopy Using Data Mining. Intelligent Automation & Soft Computing, 22(2), 229–236.

Davis, J., Ha, J., & Rossbach, C. (2006). Cost-sensitive decision tree learning for forensic classification. *Proceedings of the European Conference on Machine Learning*, 622–629. Retrieved from http://link.springer.com/chapter/10.1007/11871842_60

Engelbrecht, A. P. (2001). A new pruning heuristic based on variance analysis of sensitivity information. *IEEE Transactions on Neural Networks*, *12*(6), 1386–1389.

Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, *41*(4 PART 2), 1937–1946.

Gajowniczek, K., Orłowski, A., & Zabkowski, T. (2016). Entropy based trees to support decision making for customer churn management. *Acta Physica Polonica A*, *129*(5), 971–979.

Imai, S., Yamada, T., Kasashi, K., Kobayashi, M., & Iseki, K. (2017). Usefulness of a decision tree model for the analysis of adverse drug reactions: Evaluation of a risk prediction model of vancomycin-associated nephrotoxicity constructed using a data mining procedure. *Journal of Evaluation in Clinical Practice*.

K. J. HUNT. (1993). Classification by induction: Applications to modelling and control of non-linear dynamic systems. *Intelligent Systems Engineering*, *2*(4), 231--245.

Karabatak, M. (2015). A new classifier for breast cancer detection based on Naïve Bayesian. *Measurement: Journal of the International Measurement Confederation*, *72*, 32–36.

Kuhn, M., & Johnson, K. (2013). Classification Trees and Rule-based Models. In *Applied Predictive Modeling* (pp. 369–413).

Kwak, N., & Choi, C.-H. (2002). Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, *13*(1), 143–159.

Murphy, K. (2012). *Machine Learning:A Probabilistic Perspective*.

Park, S. Y., & Bera, A. K. (2009). Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics*, *150*(2), 219–230.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Quinlan, J. R., & Ross, J. (1993). C 4.5: Programs for machine learning. *The Morgan Kaufmann Series in Machine Learning, San Mateo, CA: Morgan Kaufmann, |c1993*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(July 1928), 379–423.

Such, F. (1972). Probability Theory, (1967), 9–32.

Su, C. ,Ju,S.,Liu,Y., and Yu,Z. (2014). An empirical study of Skew-Insensitive Splitting Criteria and Application in Traditional Chinese Medicine. *Intelligent Automation & Soft Computing*, 535-554.

Wang, Y., Xia, S. T., and Wu, J. (2017). A less-greedy two-term Tsallis Entropy Information Metric approach for decision tree classification. *Knowledge-Based Systems*, *120*, 34–42.

Zhang, S., and Zhu, Z. (2005). Study on decision tree algorithm based on autocorrelation function. *Systems Engineering and Electronics*, *27*.

7 NOTES ON CONTRIBUTORS

K.V. Uma is working as Assistant Professor in the Department of Information Technology, Thiagarajar College of Engineering, Madurai, Tamil Nadu, India. She is doing research in the area of Data Mining. She has published more than 20 Journal and conference paper in the area of Data mining specifically in Classification.
Email kvuit@tce.edu

Dr S. Appavu alias Balamurugan working as a Research Director and Professor, Department of Computer Science Engineering, E.G.S.Pillay Engineering College, Nagapattinam, Tamilnadu, India. He has published more than 123 Journal and Conference publication in the area of Data mining and Big Data Analytics with Elsevier Science Direct, Springer and IEEE publishers.
Email: datasciencebala@gmail.com