Enhancing the Efficiency of Decision Tree C4.5 Using Average Hybrid Entropy

Poonam Rani, Kamaldeep Kaur(✉), and Ranjit Kaur

Lovely Professional University, Phagwara, Punjab, India
ranipoonam405@gmail.com, Kaurdeep230@gmail.com,
reetbansal09@gmail.com

Abstract. Getting the efficient and effective decision tree is important, because of its numerous applications in mining and machine learning. Different modifications have been done on the splitting criteria in the decision tree. Different entropy concepts are introduced by different scholars. Shannon's entropy, Renyi's entropy, and Tsalli's entropy are those entropies which can affect the overall efficiency of decision tree C4.5. This research implemented new average hybrid entropy that has combined statistical properties of Reyni's and Tsalli's entropy, Average Hybrid entropy is the average between the maxima of Reyni's and Tsalli's entropy. The overall idea is, applying Average Hybrid entropy on the basis of instances and integrates those instances after pruning. This makes the pruning process easy and gives better results. Research is done on three standard datasets Credit-g, Diabetes, and Glass dataset taken from UCI repository; it is proved that the average hybrid entropy is having the more efficient results.

Keywords: Shannon's entropy · Reyni's entropy · Tsalli's entropy
Data mining · Machine learning · C4.5 · Decision tree · J48 classifier

1 Introduction

A decision tree is the supervised learning algorithm. Although it is one of the earlier algorithms introduced in machine learning and in data mining process, still it is an important and in-demand area of research. The reason behind it is that it provides better results and it is flexible in nature. It is also easy to understand. Information gain and entropy are the two main execution parts concerned with C4.5 decision tree algorithm. The overall thought is to identify the unspecified or hidden instances using the known rules of formation for the instances. The key plan is to split the tree for getting the resultant leaves which will act like our instances.

In a decision tree, the key inspiration is to quantify the homogeneity for the given dataset. Here to find the information gain of instances through which one can split the dataset accordingly. The algorithm starts in a way by splitting the nodes on the basis of their information gain and the node with highest information gain is selected as parent node or root node of the tree. And the base of calculating the info gain is entropy calculation. Information gain is mainly recorded between entropy before and after the split. Further, the maximum gain point becomes the split value and recursively this

A. K. Luhach et al. (Eds.): ICAICR 2018, CCIS 955, pp. 119–134, 2019.
https://doi.org/10.1007/978-981-13-3140-4_12

function will go on till the minimum leaf child occur or the maximum gain point arises. By the end, we can get the highest purity homogeneity, group.

There are different entropies already introduced like Shannon's entropy, Renyi's entropy, and Tsalli's entropy in decision tree [1]. Shannon's entropy is the earlier one and also useful for the research purpose as well. In [2] the researchers have used Shannon's entropy to know the influence of the landslide causing, they also mentioned the vulnerability. This is default entropy as well. But according to [3] refinement is required in Shannon's entropy for getting the better results. But if do the comparative study on the basis of split criteria [4, 5] Reyni's entropy and Tsalli's entropy gives better efficiency results. But the overall thing is choosing the best approach for the maximum efficiency. For this evaluation, can combine the best statistical properties of the Reyni's and Tsalli's we can generate the average hybrid entropy [6]. This average hybrid entropy contains one more divergence variable named *Dp* which provides more accurate results [7–11].

Zhong [12] have published a paper on the analysis of cases based on the decision tree. The authors have been proposed an improved version of the decision tree classifier on the bases of Taylor method and entropy. The entropy is working as the backbone of the decision tree. So, they have picked this entropy concept and enhance the performance of the ID3 decision tree. Gajowniczek et al. [1] published a paper on Comparison of a decision tree with Reyni and Tsalli's entropy applied for imbalanced churn dataset. In this paper, they modified the decision tree C4.5. They have experimented the α parameter. By changing the α parameter in the algorithm and apply it on the given dataset they analysis the efficiency difference. At the end they have shown the better results of using the Reyni and Tsalli's entropy as compared to the standard Shannon's entropy. Wang et al. [5] presents a paper on unifying the split criteria of decision trees using Tsalli's entropy. They have used the Tsalli's entropy in splitting criteria of the decision tree and introduced new algorithm. They also elaborate the concept of the variable α. Results are clearly shown that the newly proposed algorithm is giving better results as compared to the standard entropy that is Shannon's entropy Mehmet et al. [6] On Statistical Properties of Jizba-Arimitsu Hybrid Entropy. In this paper, the scholar has provided the statistical properties of the Hybrid entropy. They have actually combined the properties of Renyi's and Tsalli's entropy. In this entropy important factor is $\frac{q+1}{2}$, which defines that yes this is average hybrid entropy. This is called average because this provides average between Renyi's and Tsalli's entropy. They have concluded that this entropy can be defined and used in mathematics, physical and statistical application. This will provide the connection bond because of fisher metric and will give the more precise evaluation of the model parameter [9].

Further Paper is divided into section organization which contains rest of the detail about this implementation; Sect. 2 contains the materials and methods Sect. 3 presents the improved decision tree. Section 4 describes the algorithm for the average hybrid entropy. Section 5 presents the experiments and a result Sect. 6 provides the conclusion part.

2 Materials and Methods

2.1 Data Sets

To evaluate and identify the implementation progress results, three datasets are used namely Credit-g, Diabetes and Glass datasets. These datasets are extracted from the UCI repository. UCI is the machine learning repository. All the datasets available in the UCI repository are standard datasets. All three datasets are experimented using the four different entropies. And these data sets are evaluated and compared on various parameters such as accuracy, true positive rate, false positive rate, precision and recall and ROC for all entropies.

2.2 Entropy

It is known as the sum of the probability of each class label times the log probability of that respective label [13]. In decision tree, we need to calculate entropy two times. Firstly we have to calculate it for the individual attribute, and then calculate it for the combined attributes.

2.3 Gain Ratio

In decision tree learning, information gain ratio is a ratio of information gain to the intrinsic information. It is used to reduce a bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute.

2.4 Shannon's Entropy

A simple formula to calculate Shannon's entropy is given below-

$$\mathbf{E}(\mathrm{S}) = \sum_{\mathbf{i=1}}^{\mathbf{c}} -\mathbf{pi} \log_2 \mathrm{pi} \tag{1}$$

Here s describes the support whether the set will further classify as yes class or no class. $\sum_{i=1}^{c} -pi \log_2$ pi will provide the summation of all the values of the set. Further logarithm base 2 will be implemented and minus the one set's value from another [14].

To calculate the gain ratio the formula is-

$$\mathrm{Gain}(\mathrm{T}, \mathrm{X}) = \mathrm{Entropy}(\mathrm{T}) - \mathrm{Entropy}(\mathrm{T}, \mathrm{X}) \tag{2}$$

The main fundamental of the decision tree is to identify the attribute having the highest information gain, which will further act as the root node. This is intuition gets from the homogeny of dataset and identifies the best class label at the end. Here T, X are two different attributes from which the attribute having highest value will be considered as root attribute [14].

These two are Shannon-Boltzmann-Gibbs entropy based formulas. But as we see from [15] there are other methods to use these entropies in an efficient manner.

2.5 Renyi's Entropy

The formula to calculate Renyi's entropy-

$$H\alpha(X) = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n} p_i^{\alpha}\right) \tag{3}$$

Here the variable X is a generates the all possible outcomes 1, 2, …, n. Further corresponding probability pi will be calculated for each i = 1, …, n. logarithm base 2 will be called for all the calculation. If the probabilities are for all i = 1, …, n, then all the Renyi's entropies of the distribution are equivalent [16].

2.6 Tsalli's Entropy

Formula to calculate Tsalli's entropy-

$$S_q(p_i) = \frac{k}{q-1}\cdot\left(1 - \sum_{i=1}^{\cdot} p_i^{q}\right) \tag{4}$$

This is same as the Renyi's entropy but here the factor q is introduced. Here q is the real number. This is a helpful factor to calculate the entropy in a different manner [15].

2.7 Average Hybrid Entropy

Entropy is a concept used for the different domains. This is the concept of physics uses for the thermodynamics, statistics and for informational concepts. In decision tree classification entropy leads to finding the best class label. In [6] Jizba and Arimitsu introduced new hybrid entropy. This entropy contains the statistical properties of Reyni's and Tsalli's entropy. Mainly the continuous divergence and continuous average entropy are working differently for measuring from the Reyni's and Tsalli's entropy. In [5] it is clearly mentioned all different entropies and statistical properties of the average hybrid entropy as well. Recently various scholars introducing the hybrid entropy concept, but there are some properties missing as well. In [6] all properties are clearly defined for the hybrid entropy. The average entropy is also known as average hybrid entropy because it is an average between Reyni's and Tsali's entropy [7]. The average hybrid entropy is defined as:

$$A_q(p) = D\frac{q+1}{2}(p). \tag{5}$$

Now, discussion of the different interesting properties has to encounter. First of all, this is clearly described for the parameter q > 0, which is properly matched with Renyi's and Tsalli's entropy. One more thing is that in this entropy fisher metric is used which is also similar to the Reyni's and Tsallis entropy. If the value of q becomes 1 then it will similar to the Shannon entropy. For Reyni's entropy q is defined as $q \leq 1$, similarly in the average hybrid entropy q work for $q \leq 1$.

In actual scenario the important factor is q + 1/2. This provides average between the Reyni's and Tsalli's entropy [8]. We can examine from here

$$x\frac{q+1}{2}y = x+y+\left(1-\frac{q+1}{2}\right)xy = x+y+\frac{1-q}{2}xy = 2\left(\frac{x}{2}q\frac{y}{2}\right). \tag{6}$$

Here the order of q is non- extensive. According to Kolmogov-Nagumo function the average hybrid entropy defined as [9].

$$ln\frac{q+1}{2}(x) = \frac{x\,1-(q+1)/2-1}{1-\frac{q+1}{2}} = 2\frac{x\,(1-q/2-1)}{1-q} = 2\ lnq(\sqrt{x}) \tag{7}$$

In deformed logarithm, average hybrid entropy can also be defined [10].

$$\begin{aligned} &\ln q\, n \geq \ln q + 1\, 2\, n \geq \ln n \text{ for } q \leq 1, \\ &\ln q\, n \leq \ln q + 1\, 2\, n \leq \ln n \text{ for } q \geq 1 \end{aligned} \tag{8}$$

After defining these forms it is clear that average hybrid entropy lies between maxima of Tsalli's and Reyni's entropy.

3 Improved Decision Tree

There are different classification algorithms but decision tree provides more accurate results and fewer complexes as well. Hereby combing the Reyni's and Tsalli's entropy, we generated the Average Hybrid Entropy [7]. This is defined as:

$$\mathrm{A}_q(p) = D\frac{q+1}{2}(p) \tag{9}$$

Here D is same as α defined in Renyi's and Tsalli's entropy. Here the parameter α is used to adjust the measure depending on the shape of probability distributions [1]. Q is also the same variable as defined by the sales entropy [5]. The actual change is going to be done using the equation $\frac{q+1}{2}$. This defines the Average maxima between Renyi's entropy and Tsalli's entropy.

Pruning is a very important step in decision tree distribution. This helps the decision tree to reduce the overfitting and increase the efficiency. This reduces the complexity of the decision tree and provides the ability to classify instances [17].

Data mining is interdisciplinary field, where the statistical and algorithmic combination can be useful to get the effective outcomes. Entropy is the concept of physics and Average Hybrid Entropy is derived from the statistical properties of the Renyi's and Tsalli's entropy. The modification has been done on the entropy formula. Fisher matrix and Cramer Rao bond depicts that the maxima (calculated in Average Hybrid Entropy) will definitely affect the final calculated results [9, 10], because Cramer Rao

bond works on removing the biasness of the decision tree classifier algorithm. Fisher matrix is again directly linked with the respective outcomes. This is totally depends on the maximum likelihood estimates, calculated by the fisher matrix. This estimation will be reducing the error rate using the Average Hybrid Entropy. By choosing the combined statistical properties new entropy concept has been introduced and further implementation describes that all the factors mentioned here are perfectly satisfied and affect the overall results. Further the flow chart for the Average Hybrid Entropy classifier algorithm is describe, the process of the implementation is inspiration for getting the appropriate results for the research purpose.

Figure 1 describes the flow chart for the Average Hybrid Entropy decision tree algorithm.

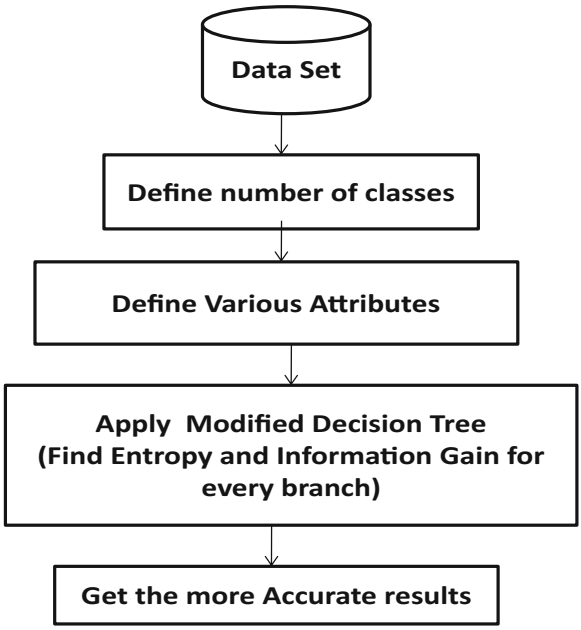

Fig. 1. Flow chart for the implemented algorithm

For implemented improved decision tree the concept of Average Hybrid Entropy is used. In Weka, the C4.5 algorithm is available in J48 classifier Tree. This paper describes the algorithm along with implementation. For achieving the increased efficiency, implemented work used pruning, inside pruning there are instances, these instances also affect the efficiency. So, until decision tree reaches to its leaf node iteration should be there to obtain the average hybrid methodology formula and calculate the overall efficiency.

4 Algorithm for Average Hybrid Entropy

To implement the algorithm Weka tool has been used and for the coding purpose Net Beans IDE platform also involved here. In Weka algorithm, the C4.5 source code is available as J48 classification algorithm. For checking the results datasets also used from the Weka. Overall results clearly specify that the proposed work having best accuracy results. In this paper compare the Average Hybrid Entropy with the Shannon's, Renyi's and Tsalli's entropy as well. By using these practical tools finally get the best accuracy with fewer complexities.

```
// Pruning distribution for decision tree ( training instances, data, leaf_node, instances,
//son_nodes, local_instances, local_split_model)
training instances =data;
if(!leaf_node)
        // call local instances and split data
    for (each instance<son_node.lenght)
        // call distribution for local instances
Else
            // check whether there are more instances at the  leaf
              // call hybrid entropy prune the decision tree

        if ( instances empty)
                    //stop pruning the further instances
        end if
end
else
end if
```

5 Experimental Results

This research implemented the average hybrid entropy in the decision tree and modified the decision tree for obtaining the information gain of nodes, After implementing the Average Hybrid Entropy results are clearly showing that newly implemented hybrid entropy is having more accurate results as comparing the Shannon's, Reyni's and Tsalli's entropy. For achieving the output graph three datasets are used, named as Credit-g, Diabetes and Glass dataset. Respective results for each dataset are 72.60%, 75%, 68.22% these are the providing better accuracy results. There are some factors on which these entropies have to be compared that are TP Rate (true positive rate), FP Rate (false positive rate), Precision and ROC Area. 10 fold cross validation is also used for evaluation of the accuracy factor. Graph depicts that Average Hybrid entropy contains better results. The reason is maxima factor of Renyi's and Tsalli's entropy.

Accuracy is overall efficient output results which provide more purity in the dataset results. This will further help in better data mining process. TP is true positive rate; the

instances which are actually positive and accurate for the final results. FP is the false positive rate; the instances which are incorrect and involved to produce the final results. Precision is another factor involved in the final accuracy results. It is something which gives the exactness of the accuracy. The area under ROC curve (AUC) defines the overall accuracy.

Further the table values and their respective line graphs representing the accuracy results and showing that the implemented decision tree having average hybrid entropy gives the better accuracy results.

Table 1. Values for the respective entropy for credit-g data set

Credit-g dataset evaluation					
C4.5	Accuracy	TP rate	FP rate	Precision	ROC area
Shannon's entropy	70.50%	0.705	0.475	0.687	0.639
Reyni's entropy	71.40%	0.714	0.45	0.7	0.662
Tsallis entropy	71%	0.71	0.46	0.695	0.637
Average hybrid entropy	72.60%	0.726	0.453	0.709	0.673

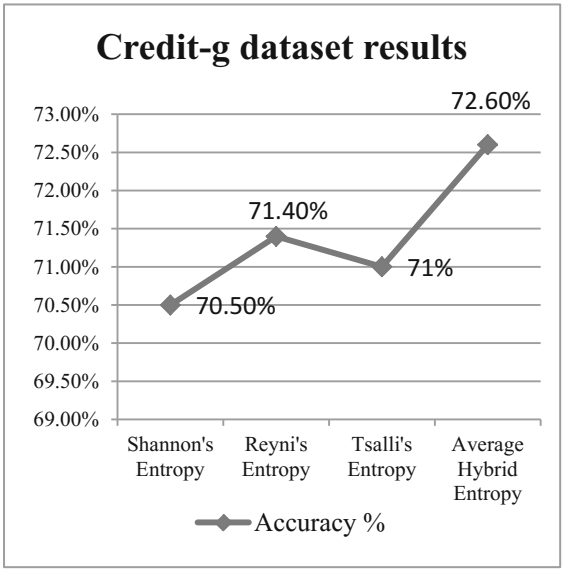

Fig. 2. Line graph representation for credit dataset

Table 1 clearly depicts the overall evolution of decision tree after applying the 4 different entropies on the Credit-g dataset. The affecting factors are overall Accuracy (how many correctly classified instances after the execution), TP rate (True Positive rate depicts the instances that are labeled in the correct class.), FP rate (label classes

which are not involved in the correct instances but it should be), Precision (all the relevant instances) and ROC. ROC is also known as AUR which means the area under ROC curve. This value is directly proportional to the accuracy factor. The respective values are given in Table 1, which clearly presents that Average Hybrid Entropy gives the best results as a statically values that are; Accuracy is 72.60%, TP rate is 0.726, FP rate is 0.453, precision values is 0.709 and ROC Area value is 0.673. Here Fig. 2 presents the graph view for the Credit-g dataset. This clearly shows the differences in the graphical view. 72.60% is the highest accuracy value after applying the Average hybrid Entropy.

Table 2. Values for the respective entropy for Diabetes dataset

Diabetes dataset evaluation					
C4.5	Accuracy	TP rate	FP rate	Precision	ROC area
Shannon's entropy	73.83%	0.738	0.327	0.735	0.751
Renyi's entropy	74.35%	0.743	0.321	0.74	0.758
Tsalli's entropy	74.48%	0.745	0.331	0.74	0.74
Average hybrid entropy	**75%**	**0.75**	**0.323**	**0.745**	**0.754**

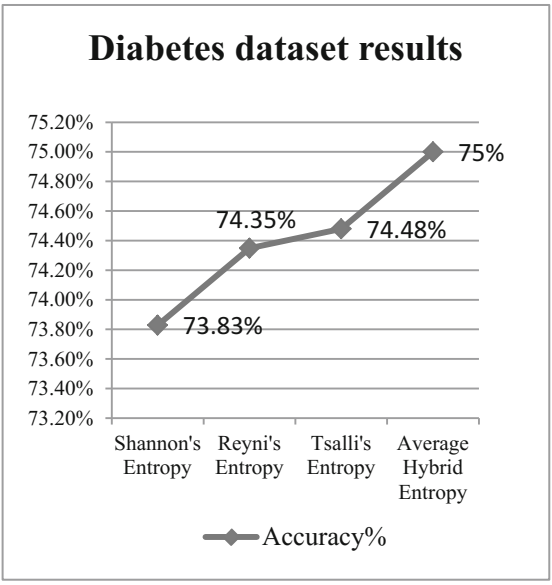

Fig. 3. Line graph representation for Diabetes dataset

In Table 2 all the factors which are involved in the overall results and evaluation purpose have encountered and their respective values are mentioned as well. This table draws all factor values for the different entropies implemented on the Diabetes dataset.

Table 3. Values for the respective entropy for glass data set

Glass dataset evaluation					
C4.5	Accuracy	TP rate	FP rate	Precision	ROC area
Shannon's entropy	66.82%	0.668	0.13	0.67	0.807
Reyni's entropy	67.76%	0.678	0.125	0.677	0.824
Tsallis entropy	67.76%	0.678	0.125	0.678	0.823
Average hybrid entropy	**68.22%**	**0.682**	**0.124**	**0.682**	**0.834**

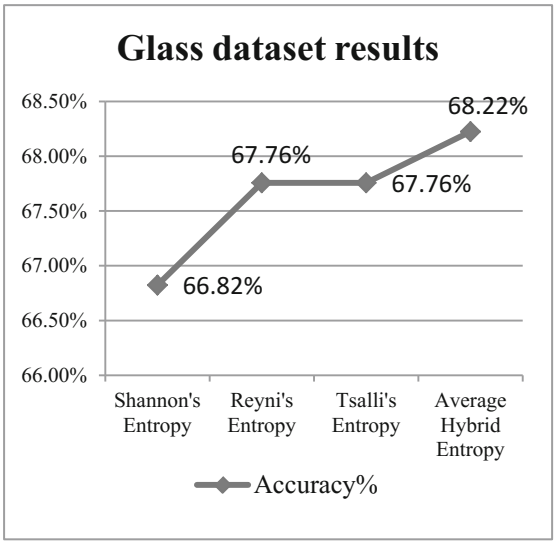

Fig. 4. Line graph representation for glass dataset

After every single evaluation, the results show that the Average Hybrid Entropy gives the best results on this dataset as well. Different factor values for Average Hybrid Entropy are; Accuracy is 75%, TP rate is 0.75, FP rate is 0.323, precision value is 0.745 and AUC or ROC value is 0.754. Line graph representation for the Diabetes dataset is given in Fig. 3. In this Fig, accuracy factor is involved which is 75% is the highest value for the Average Hybrid Entropy after implementation on the Diabetes dataset.

Similarly, Table 3 is showing all the values after evaluation of different entropies on the Glass dataset. This table shows all the affecting factors for the results same as the previous datasets. Consequently shows the Average Hybrid Entropy is giving the best results that are; Accuracy is 68.22%, TP rate is 0.682, FP rate is 0.124, the Precision value is 0.682 and ROC value is 0.834. Figure 4 is drawn after evaluating the factor accuracy. This Line graph representation showing the clear enhancements of the results. 68.22% is the highest accuracy factor measured by involving Average Hybrid Entropy on the Glass dataset.

Table 4. Results of the implemented work on the basis TP rate, FP rate, precision, ROC area factors

C4.5	Dataset	Accuracy	TP Rate	FP Rate	Precision	ROC Area
Shannon's Entropy	Credit-g	70.50%	0.705	0.475	0.687	0.639
Renyi's Entropy		71.40%	0.714	0.45	0.7	0.662
Tsalli's Entropy		71%	0.71	0.46	0.695	0.637
Average Hybrid Entropy		**72.60%**	**0.726**	**0.453**	**0.709**	**0.673**
Shannon's Entropy	Diabetes	73.83%	0.738	0.327	0.735	0.751
Renyi's Entropy		74.35%	0.743	0.321	0.74	0.758
Tsalli's Entropy		74.48%	0.745	0.331	0.74	0.74
Average Hybrid Entropy		**75%**	**0.75**	**0.323**	**0.745**	**0.754**
Shannon's Entropy	Glass	66.82%	0.668	0.13	0.67	0.807
Renyi's Entropy		67.76%	0.678	0.125	0.677	0.824
Tsalli's Entropy		67.76%	0.678	0.125	0.678	0.823
Average Hybrid Entropy		**68.22%**	**0.682**	**0.124**	**0.682**	**0.834**

Further Table 4 is showing the overall comparisons of the decision tree after implementing Shannon's entropy, Renyi's entropy, Tsalli's entropy and Average Hybrid Entropy. There are three datasets availed from the UCI repository. UCI repository provides the standard datasets for the research scholars for doing the different analysis. Three datasets are involved for the analysis for the complete results and evaluations. These datasets are Credit-g dataset, Diabetes dataset, and Glass dataset. After all the implementation based on the main factors (Accuracy, TP rate, FP rate, Precision, and ROC) consequently the Average Hybrid Entropy is giving the best results for all the datasets. On the bases of accuracy factor the Credit-g dataset shows 72.60% accuracy, Diabetes dataset presents 75% accuracy and on the Glass dataset, the Average Hybris Entropy gives the 68.22% accuracy. These are showing that this is one of the entropy which can affect the overall decision tree efficiency.

Here Fig. 5 shows the bar graph representation for the comparison of all the datasets on the different mentioned entropies. This is showing the enhancement of the decision tree algorithm's efficiency is achieved by implementing the Average Hybrid Entropy. Accuracy % factor is used for showing the graph results. Three datasets are involved in the for the evaluation purpose their respective results for the different

entropies are shown in Fig. 5. To see the progress of the Diabetes dataset the highest accuracy value is 75% that is using the Average Hybrid Entropy. For the second dataset named Credit-g, the highest accuracy has been achieved using the Average Hybrid Entropy that is 72.60%, last but not least for the third dataset Glass the accuracy value is 68.2%. Here we can clearly see the enhancement of efficieny in decision tree algorithm using Average Hybrid Entropy. For the overall evaluation this is clearly showing that the Average Hybrid Entropy imprving more than 2% as compared to the standard entropy for the same dataset.

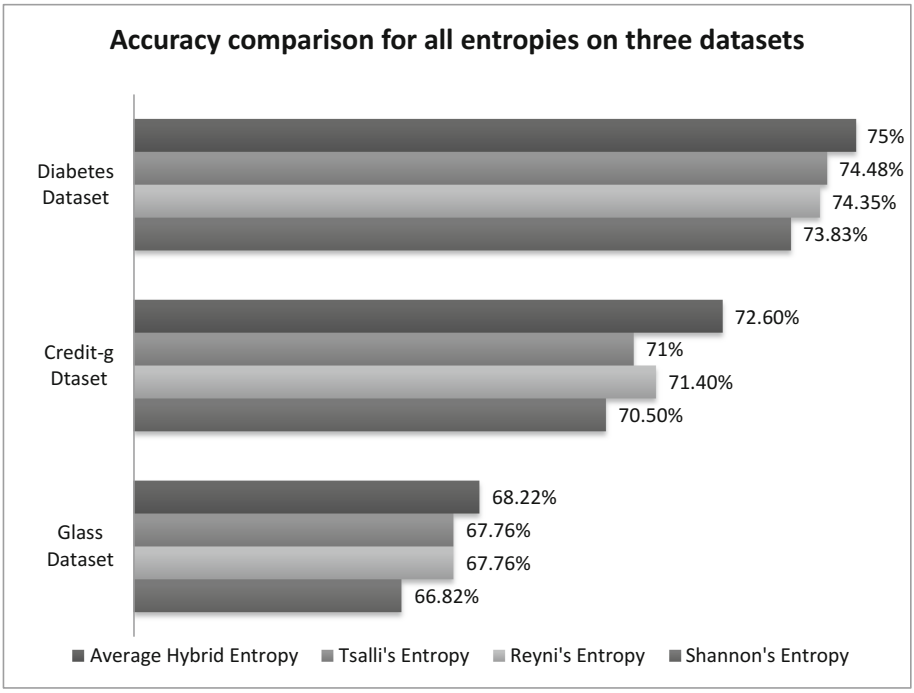

Fig. 5. Respective results for all the entropies on three dataset

The area under ROC curve (AUC) defines the overall accuracy. An area under 0.5 is counted as worthless. The higher area under ROC defines the excellent accuracy results. Here different ROC is shown for separate entropies for the **glass dataset**. The ROC value defines the overall difference. ROC is directly responsible for the final accuracy results. From the following Figs we can easily define what Shannon's entropy is having ROC area **0.807**, similarly Renyi's entropy having ROC area value **0.824**, further Tsalli's entropy having the ROC curve area value **0.823** and finally for the average hybrid entropy the ROC area value is 0.834 Which is the highest value and having excellent ROC area. The ROC area, when it comes near more to the value 1 then we can say that this is excellent value for the accuracy. Here the blue color defines the more excellence part. More the ROC or AUC area involved in the results the more

accurate and precise results that would be. Following Figs are showing the ROC details for different entropies after implementation on the Glass dataset (Figs. 6, 7, 8 and 9).

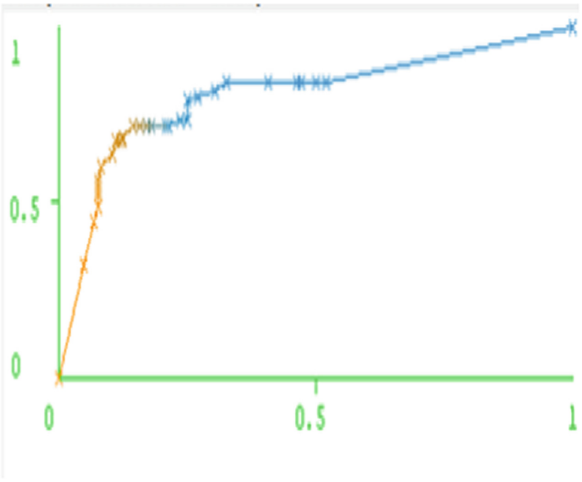

Fig. 6. ROC curve of Shannon's entropy having for glass dataset having ROC area 0.807

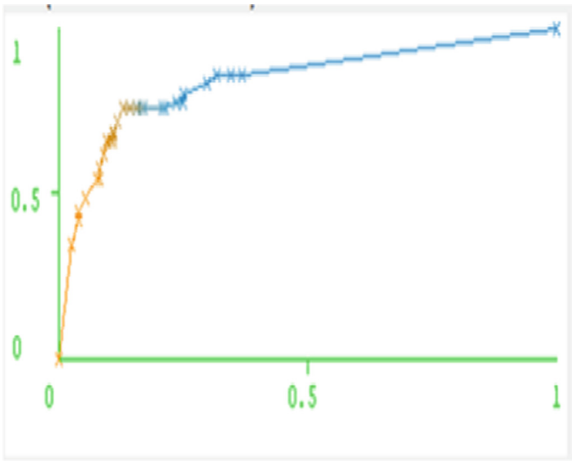

Fig. 7. ROC curve of Renyi's entropy having for glass dataset having ROC area 0.824

TP and TN here are the same because both are the sum of all true classified examples, regardless their classes. False positives, which are items incorrectly labeled as belonging to the class. False negatives items are which were not labeled as belonging to the positive class but should have been. Precision is another factor involved in the

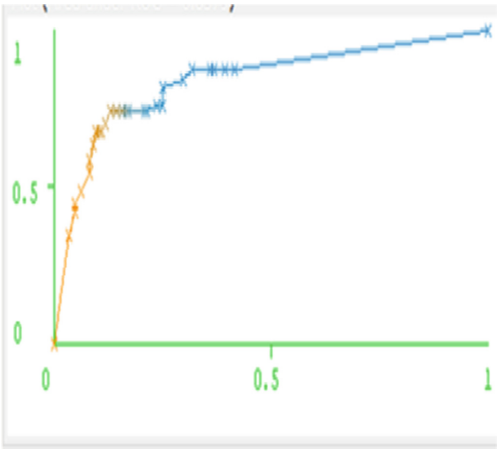

Fig. 8. ROC curve of Tsalli's entropy having for glass dataset having ROC area 0.823

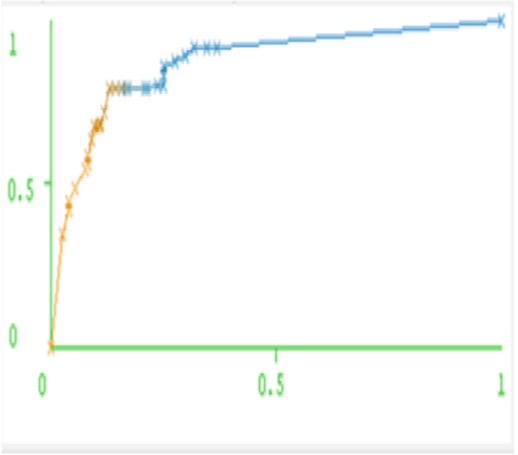

Fig. 9. ROC curve of average hybrid entropy having for glass dataset having ROC area 0.834

final accuracy results. It is something which gives the exactness of the accuracy. Here different formulas for the calculation of all these important factors are given:

Sensitivity = TP / TP + FN

Specificity = TN / TN + FP

Precision = TP / TP + FP

True-Positive Rate = TP / TP + FN

False-Positive Rate = FP / FP + TN

True-Negative Rate = TN / TN + FP

False-Negative Rate = FN / FN + TP

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

Table 5. Analysis of average hybrid entropy on three datasets

Analysis for the average hybrid entropy			
Accuracy parameters	Datasets		
	Credit-g dataset	Diabetes dataset	Glass dataset
True positive	**0.75**	**0.75**	**0.682**
False positive	**0.453**	**0.323**	**0.124**
Accuracy	**72.60%**	**75%**	**68.22%**
Precision	**0.709**	**0.745**	**0.682**

In the Table 5 here mentioned about the analysis results for Average Hybrid Entropy after performing on different datasets. All these factors calculated according to the class type. Here mentioned table represents the weighted average values for all the datasets.

6 Conclusions

In the era of machine learning and data mining, there is need of more effective and accurate decision tree algorithm, that's why enhancement of research on this topic is preferable. Decision tree uses the concept of split criteria and pruning the instances on the basis of information gain and entropy. The C4.5 algorithm is available as a source code in Weka software. To increase the efficiency there are different entropy introduced. Shannon's entropy is the standard one. Reyni's entropy and Tsalli's entropy are redefined through statistical properties. Average Hybrid Entropy is a collective form of Reyni's and Tsalli's entropy. By using the pruning properties of instances this paper implemented this Average Hybrid entropy. Consequently, it gives the better results from the previous one. It provides the more precise bond with the nodes has to be pruned in the decision tree. Different datasets from UCI repository have been picked for evaluating the results that are Credit-g, Glass and Diabetes datasets. Overall it is clear that 72.60% accuracy gives for the Credit-g, 75% and 68.22% accuracy provides for the Diabetes and Glass datasets respectively. Overall evaluation is showing that Average Hybrid Entropy enhanced the more than 2%accuracy as compared to the standard entropy. In future, more elaboration has to be performed on the basis of statistical properties of the different entropies and should be tested against the continuous as well as discrete values as a future work.

References

1. Gajowniczek, K., Ząbkowski, T., Orłowski, A.: Comparison of decision trees with Rényi and Tsallis entropy applied for imbalanced churn dataset. In: Computer Science and Information Systems, pp. 39–44 (2015)
2. Sharma, L.P., Patel, N., Ghose, M.K., Debnath, P.: Influence of Shannon's entropy on landslide-causing parameters for vulnerability study and zonation—a case study in Sikkim. India. Arab. J. Geosci **5**, 421–431 (2012)

3. Truffet, L.: Shannon Entropy Reinterpreted (2017)
4. Lima, C.F.L., De Assis, F.M., De Souza, C.P.: Decision tree based on Shannon, Rényi and Tsallis entropies for intrusion tolerant systems. In: 5th International Conference Internet Monitoring and Protection, ICIMP 2010, pp. 117–122 (2010)
5. Wang, Y., Song, C., Xia, S.-T.: Unifying Decision Trees Split Criteria Using Tsallis Entropy (2015)
6. Dogra, A.K.: A review paper on data mining techniques and algorithms. Int. Res. J. Eng. Technol. **4**, 1976–1979 (2015)
7. Ilić, V.M., Stanković, M.S.: Comments on " On q-non-extensive statistics with non-Tsallisian entropy". Phys. A Stat. Mech. Appl. **466**, 160–165 (2017)
8. Contreras-Reyes, J.E., Arellano-Valle, R.B.: Kullback-Leibler divergence measure for multivariate skew-normal distributions. Entropy **14**, 1606–1626 (2012)
9. Bercher, J.F.: Some properties of generalized Fisher information in the context of nonextensive thermostatistics. Phys. A Stat. Mech. Appl **392**, 3140–3154 (2013)
10. Bercher, J.-F.: On generalized Cramér-Rao inequalities, generalized Fisher information, and characterizations of generalized q-Gaussian distributions. J. Phys. A: Math. Theor. **45**, 255303 (2012)
11. Teimouri, M., Nadarajah, S., Hsing Shih, S.: EM algorithms for beta kernel distributions. J. Stat. Comput. Simul. **84**, 451–467 (2014)
12. Zhong, Y.: The analysis of cases based on decision tree. In: 2016 7th IEEE International Conference on Software Engineering and Service Science, pp. 142–147 (2016)
13. tackOverflow. https://stackoverflow.com/questions/1859554/what-is-entropy-and-information-gain
14. An Introduction to Data Science. http://saedsayad.com/decision_tree.htm
15. Wikipedia. https://en.wikipedia.org/wiki/Tsallis_entropy
16. Wikipedia. https://en.wikipedia.org/wiki/Rényi_entropy
17. Wikipedia. https://en.wikipedia.org/wiki/Pruning_%28decision_trees%29