UNIFYING ATTRIBUTE SPLITTING CRITERIA OF DECISION TREES BY TSALLIS ENTROPY

Yisen Wang, Shu-Tao Xia

Department of Computer Science and Technology, Tsinghua University, Beijing, China

ABSTRACT

Owing to its simplicity and flexibility, the decision tree remains an important analysis tool in many real-world learning tasks. A lot of decision tree algorithms have been proposed, such as ID3, C4.5 and CART, which represent three most prevalent criteria of attribute splitting, i.e., Shannon entropy, Gain Ratio and Gini index respectively. These splitting criteria seem to be independent and to work in isolation. However, in this paper, we find that these three attribute splitting criteria can be unified in a Tsallis entropy framework. More importantly, theoretically, we reveal the relations between Tsallis entropy and the above three prevalent attribute splitting criteria. In addition, we generalize the splitting criterion of the decision tree, and provide a new simple but efficient approach, Unified Tsallis Criterion Decision Tree algorithm (UTCDT), to enhance the performance of the decision tree. Experimental evidences demonstrate that UTCDT achieves statistically significant improvement over the classical decision tree algorithms, even yields comparable performance to state-of-the-art classification algorithm.

***Index Terms*—** Decision tree, Unified framework, Splitting criteria, Tsallis entropy

1. INTRODUCTION

The decision tree, as one of the first machine learning approaches, has been widely used but still being actively researched in many real-world learning fields, e.g., image interpolation [1], speech synthesis [2], just to name a few. The decision tree is not only simple to understand and interpret, but also offers relatively good results and computational efficiency. The general idea of the decision tree is to predict unknown input instances by learning simple decision rules inferred from several known training instances. A decision tree is mostly often induced in the following top-down manner [3]. A given dataset is partitioned into several subsets by a splitting criterion test on attributes. The highest scoring partition which reduces the average uncertainty mostly is selected to grow the tree, by making the node be the parent of the newly created child nodes. This procedure is applied recursively until some stopping conditions, e.g., maximum tree depth or minimum leaf size, are reached.

Generally speaking, splitting criterion is a fundamental issue in the induction of the decision tree. A series of papers have analyzed the importance of the splitting criterion [4, 5]. They demonstrated that different splitting criteria have substantial influence on the generalization error of the induced decision tree. Thus, a large number of decision tree induction algorithms with different splitting criteria have been proposed. For example, the Iterative Dichotomiser 3 (ID3) algorithm [6] is based on Shannon entropy; the C4.5 algorithm [7] is based on Gain Ratio which is considered as the normalized Shannon entropy; while the Classification And Regression Tree (CART) algorithm [8] is based on Gini index. These algorithms seem to be independent, and have been coexisting for a long time. As we all know, it is difficult to judge which algorithm is usually better than others on all datasets, which reflects their lack of adaptability to different datasets.

In this paper, we propose a unified Tsallis entropy framework, which not only unifies the above three prevalent splitting criteria, i.e., Shannon entropy, Gain Ratio and Gini index, but also can adapt to various datasets through a tunable parameter q. To the best of our knowledge, this is the first time to propose a unified framework combining splitting criteria together. Theoretically, we analyze the corresponding relations between Tsallis entropy with different q and other splitting criteria. Shannon entropy and Gini index are just two specific cases of Tsallis entropy with $q = 1$ and 2, while Gain Ratio can be regarded as the normalized Tsallis entropy with $q = 1$. Based on the unified framework, we propose a Unified Tsallis Criterion Decision Tree algorithm, called UTCDT, which provides a new simple but efficient approach to improve the performance of the decision tree. Empirically, UTCDT significantly outperforms the classical decision tree algorithms, both on classification accuracy and tree size. Compared to state-of-the-art Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel [9], UTCDT also yields comparable performance with a lower algorithm complexity.

2. BACKGROUND

Shannon entropy is a measure of uncertainty in a distribution [10]. However, with respect to the two typical distributions observed in the macroscopic world, i.e., exponential distribution family and power-law heavy-tailed distribution family

[11], we cannot characterize the latter one through maximizing Shannon entropy subject to the normal mean and variance. The reason is that Shannon entropy implicitly assumes a certain trade-off between contributions from the tails and the main mass of the distribution [12]. It should be worthwhile to control this trade-off explicitly to characterize these two distribution family. Entropy measures that depend on powers of probability, e.g., $\sum_{i=1}^{n} p(x_i)^q$, can provide such control. Thus, some parameterized entropies have been proposed. A well-known generalization of this concept is Tsallis entropy [13], which extends its applications to so-called non-extensive systems [14] using an adjustable parameter q.

Tsallis entropy is defined by:

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

where X is a random variable taking values $\{x_1, ..., x_n\}$ and $p(x_i)$ is the probability of x_i.

For $q < 0$, Tsallis entropy is convex (or convex downward). For $q = 0$, Tsallis entropy is non-convex and non-concave. While for $q > 0$, Tsallis entropy is concave (or convex upward), satisfying similar properties to Shannon entropy [15], e.g., S_q is nonnegative, and obtains maximum when $p(x_i) = 1/n, i = 1, 2, ..., n$.

As a generalization of Shannon entropy, Tsallis entropy has been tested in the decision tree in the prior work [12]. However, Maszczyk and Duch [12] only tested the performance of Tsallis entropy in C4.5 with some given q. The relations between Tsallis entropy and other splitting criteria were not explored, and the unified framework was also not presented. These are exactly what we are doing in our work.

3. THE UNIFIED TSALLIS ENTROPY FRAMEWORK

In this section, the three prevalent attribute splitting criteria, i.e., Shannon entropy, Gain Ratio and Gini index, are unified in the Tsallis entropy framework. We also reveal the relations among them theoretically.

Theorem 1. *Tsallis entropy $S_q(X)$ converges to Shannon entropy $H(X)$ for $q \to 1$.*

Proof.

$$\begin{aligned}\lim_{q\to 1} S_q(X) &= \lim_{q\to 1} \frac{1}{1-q}(\sum_{i=1}^{n} p(x_i)^q - 1) \\ &\overset{(a)}{=} \lim_{q\to 1} \frac{(\sum_{i=1}^{n} p(x_i)^q - 1)'}{(1-q)'} \\ &= -\sum_{i=1}^{n} p(x_i) \ln p(x_i) = H(X), \end{aligned} \tag{2}$$

where the equation (a) is due to the L'Hopital's rule [16]. □

Theorem 2. *Gini index is exactly the specific case of Tsallis entropy with $q = 2$.*

Proof.

$$\begin{aligned}\{S_q(X)\}_{q=2} &= \underbrace{\frac{1}{1-q}(\sum_{i=1}^{n} p(x_i)^q - 1)}_{q=2} = 1 - \sum_{i=1}^{n} p(x_i)^2 \\ &= \sum_{i=1}^{n} (p(x_i) - p(x_i)^2) = \sum_{i=1}^{n} p(x_i)(1 - p(x_i)) \\ &= \text{Gini index}. \end{aligned} \tag{3}$$

□

Theorem 3. *Gain Ratio (GR) adds a normalization to standard Shannon entropy based Information Gain, and if Shannon entropy is replaced by Tsallis entropy, Gain Ratio is generalized to Tsallis Gain Ratio (Tsallis GR). Tsallis Gain Ratio converges to Gain Ratio as $q \to 1$.*

Proof.

$$\begin{aligned}\lim_{q\to 1} \text{Tsallis GR} &= \lim_{q\to 1} \frac{S_q(D) - \frac{|D'|}{|D|} S_q(D') - \frac{|D''|}{|D|} S_q(D'')}{S_q\left(\frac{|D'|}{|D|}, \frac{|D''|}{|D|}\right)} \\ &= \frac{\overbrace{H(D) - \frac{|D'|}{|D|} H(D') - \frac{|D''|}{|D|} H(D'')}^{\text{Information Gain}}}{H\left(\frac{|D'|}{|D|}, \frac{|D''|}{|D|}\right)} \\ &= \text{Gain Ratio (GR)}, \end{aligned} \tag{4}$$

where $|\cdot|$ denotes the number of instances, and D', D'' are two child subsets if dataset D is split in binary. □

With different q, Tsallis entropy degenerates to Shannon entropy, Gini index and Gain Ratio. In other words, we can promote the performance of the decision tree in the unified Tsallis entropy framework via tuning parameter q, which is a new perspective for enhancement.

4. UNIFIED TSALLIS CRITERION DECISION TREE ALGORITHM

As state above, through tuning parameter q, Tsallis entropy can form a variety of attribute splitting criteria. Therefore, we propose the Unified Tsallis Criterion algorithm for Decision Tree induction (UTCDT), which can be adapted to various datasets by choosing an appropriate q.

4.1. Tree Construction

Given a dataset $D_n = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{n}$, $\boldsymbol{x_i} \in \mathbb{R}^d$ with attributes A_j $(j \in \{1, 2, \ldots, d\})$, and class label $y_i \in \{1, 2, \ldots, K\}$.

For each tree node, we search for every possible pair of attribute and cutting point to choose the optimal attribute and cutting point to grow the tree in a binary split manner, like CART [8]. For an attribute A_j, we obtain:

$$I(A_j(C_j)) = T(D) - \frac{|D'|}{|D|}T(D') - \frac{|D''|}{|D|}T(D''). \quad (5)$$

Here $A_j(C_j)$ denotes the candidate pair of attribute as well as cutting point, D is the data belonging to one node to be partitioned, and D', D'' are the two child nodes that would be created if D is partitioned at $A_j(C_j)$. The function $T(D)$ is the impurity criterion, i.e., Tsallis entropy, which computes over the labels of the data which fall in the node. The pair of attribute A_j and cutting point C_j is chosen to construct the tree which maximizes $I(A_j(C_j))$.

The above procedure is applied recursively until some stopping conditions are reached. The stopping conditions consist of three principles: (i) The classification is achieved in a subset. (ii) No attributes are left for selection. (iii) The cardinality of a subset is lower than the predefined threshold.

4.2. Prediction

Once the tree has been trained by the data as a classifier g_n, it can be used to predict for the new unlabeled instance $\boldsymbol{x}$.

The decision tree makes prediction in a majority vote manner. The probability of each class $k \in \{1, 2, \ldots, K\}$ is

$$\eta^{(k)}(\boldsymbol{x}) = \frac{1}{|A_n(\boldsymbol{x})|} \sum_{(\boldsymbol{x_i}, y_i) \in A_n(\boldsymbol{x})} \mathbb{I}(y_i = k), \quad (6)$$

where $A_n(\boldsymbol{x})$ denotes the leaf containing $\boldsymbol{x}$, and $\mathbb{I}(e)$ is the indicator function that takes 1 if e is true and 0 for other cases. Then the tree prediction is the class that maximizes this probability:

$$g_n(\boldsymbol{x}) = \arg\max_k \{\eta^{(k)}(\boldsymbol{x})\}. \quad (7)$$

Algorithm 1 UTCDT predicted value at an unlabeled instance $\boldsymbol{x}$

1: **Input**: Data $D_n = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$, Attributes A_j $(j \in \{1, 2, \ldots, d\})$, Class label $y_i \in \{1, 2, \ldots, K\}$
2: **Output**: UTCDT predicted value at $\boldsymbol{x}$
3: **while** not satisfying stopping condition **do**
4: **for** each pair attribute A_j and cutting point C_j **do**
5: Compute $I(A_j(C_j))$ according to Eq. (5)
6: **end for**
7: $A_{best}(C_{best}) = \arg\max I(A_j(C_j))$
8: //A_{best}, C_{best} is the best pair of attribute and cutting point
9: Grow the tree using A_{best}, C_{best}
10: **end while**
11: Compute the predicted value following Eqs. (6) and (7).
12: **Return** Predicted value at $\boldsymbol{x}$.

4.3. Summary of UTCDT

Here, we summarize our proposed Unified Tsallis Criterion Decision Tree algorithm (UTCDT) in a pseudo-code format in Algorithm 1. Compared to the classical decision tree induction algorithms, the difference lies in the Tsallis entropy splitting criterion, which can degenerate to Shannon entropy, Gini index and Gain Ratio through different q. Note that, similar to CART, UTCDT is also applicable to both numerical and categorical attributes. In the following Experiments section, we will see that UTCDT can construct a decision tree with higher classification accuracy and smaller tree size.

5. EXPERIMENTS

In this section, we empirically assess the performance of UTCDT. Firstly, we present the influence of parameter q in the unified Tsallis entropy framework. Secondly, we compare UTCDT with the classical decision tree algorithms and state-of-the-art classification algorithm SVM. The 11 benchmark datasets are from UCI [17], which consists of various number of instances, numeric/categorical attributes and binary/multi classes. As for the measure metric, we choose Accuracy (ACC) and the number of nodes ($Nodes$) to measure the classification effectiveness and the tree size respectively.

5.1. The Influence of Parameter q

To exhibit the influence of parameter q roundly, we traverse q in a step of 0.1 in the range $[0.1, 10.0]$. For each selected q, we choose the Tsallis entropy criterion and perform a 10 times 10-fold cross-validation to evaluate the performance. Besides, the minimum leaf size is set to 5 to avoid overfitting.

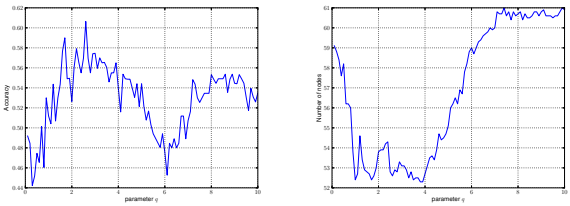
Due to the limited space, we only demonstrate the result on Glass dataset. Actually, other datasets demonstrate similar trends. Figure 1 gives an intuitive exhibition of the influence of different values of parameter q in Tsallis entropy for Glass dataset. Figure 1 (a) illustrates that the accuracy is sensitive to the change of q, and the highest accuracy is obtained at $q = 2.6$; Figure 1 (b) shows that the tree size is also sensitive to q, and the smallest tree size is achieved at $q = 3.9$. More importantly, our proposed UTCDT can obtain acceptable high accuracy and small tree size at the same time (e.g., $q = 1.7$). In summary, experimental results show that the parameter q indeed has an effect on the classification accuracy and the tree size. Moreover, we can achieve different goals through selecting different q, e.g., highest accuracy or smallest tree size or trade-off, which also reflects the adaptability and flexibility of UTCDT.

5.2. Classification Performance Analysis

To evaluate the performance of our proposed UTCDT comprehensively, we not only compare it to the classical decision tree algorithms, i.e., ID3, C4.5 and CART, but also include

Table 1. Comparisons ($ACC\%$, $Nodes$) of different algorithms on various datasets. At the significance level of 0.05, UTCDT significantly outperforms ID3, CART and C4.5 under the Friedman test [18]. Under the Wilcoxon signed rank test [18] with 0.05 significance level, Tsallis entropy significantly outperforms Shannon entropy as well as Gini index, and Tsallis Gain Ratio significantly outperforms Gain Ratio.

Dataset	ID3 (Shannon entropy)		CART (Gini index)		C4.5 (Gain Ratio)		UTCDT (Tsallis entropy)			UTCDT (Tsallis Gain Ratio)		SVM (RBF)
	ACC	$Nodes$	ACC	$Nodes$	ACC	$Nodes$	ACC	$Nodes$	q	ACC	$Nodes$	ACC
Yeast	52.8	199	51.8	196.6	52.1	326.2	**56.9**	**195.8**	1.4	51.2	197.1	59.7
Glass	51.2	52.4	52.6	53.8	44.2	52	**60.6**	52.6	2.6	53.1	**51.5**	66.0
Vehicle	71.7	103	70.2	**100**	72.3	147.2	**73.8**	111.0	0.6	73.4	135.7	83.8
Wine	92.9	12.0	90.0	12.0	92.4	9.4	**95.9**	9.6	3.1	92.9	**9.2**	94.9
Haberman	70.3	32.2	70.3	33.0	72.8	33.0	74.2	33.2	7.1	**74.8**	**32.0**	77.9
Car	98.2	106.4	98.1	106.8	**98.5**	106.5	98.3	**106.2**	0.8	98.4	106.6	93.9
Scale	75.9	97.6	76.1	97.2	74.5	**77.0**	78.2	93.1	3.1	**78.5**	**77.0**	89.8
Hayes	81.5	28.8	80.0	25.3	79.2	19.6	**82.3**	19.5	8.6	81.5	**19.2**	81.2
Monks	51.9	89.0	52.1	88.6	52.9	**88.0**	**57.3**	89.6	8.9	54.9	**88.0**	56.4
Abalone	25.4	89.2	25.0	85.8	20.3	**84.3**	**26.8**	86.2	0.8	25.7	85.1	22.5
Cmc	49.1	267.0	47.4	264.0	45.7	242.8	**52.0**	264.2	1.2	47.8	**242.1**	54.2



(a) Classification Accuracy (b) Tree Size

Fig. 1. Influence of parameter q in ACC and $Nodes$ (Glass)

state-of-the-art classification algorithm [19] SVM with RBF kernel.

We do a grid search to determine $q \in [0.1, 10.0]$ in UTCDT, maybe the optimal q for Tsallis entropy and Tsallis Gain Ratio are different, but for the fair comparison we choose the same q, e.g., optimal q for Tsallis entropy. Other settings are the same to Section 5.1. Besides, the parameters in SVM are optimized following its original paper [9].

Table 1 reports the results of different algorithms on various datasets, where SVM serves as a reference algorithm. Among decision tree algorithms with different splitting criteria, we highlight the highest accuracy and smallest tree size on each dataset in boldface. As expected, UTCDT significantly outperforms ID3, CART and C4.5 due to the fact that Tsallis entropy framework is a unification of Shannon entropy, Gini index and Gain Ratio. To be specific, compared to Shannon entropy and Gini index, Tsallis entropy achieves better performance in accuracy and tree size. Tsallis Gain Ratio also obtains better results compared to Gain Ratio. With respect to the two kinds of splitting criteria in UTCDT, we can see that Tsallis entropy prefers high accuracy while Tsallis Gain Ratio prefers small tree size. The reason lies on the normalized factor in Tsallis Gain Ratio which has influence on the tree structure to a certain extent. Besides, the optimal q for each dataset is usually not equal to 1 or 2, and it is associated with the properties of the dataset, which implies that tuning q enables UTCDT to own adaptability and flexibility.

In addition, note the last column of Table 1, the results where SVM is better than UTCDT are marked with underlining. We can see that UTCDT only achieves lower accuracy on 6 out of 11 datasets, compared to SVM. On these 6 datasets, the gap is not large, even within 3% on half of datasets (i.e., Yeast, Haberman and Cmc). It is worth mentioning that the complexity of SVM is very high $O(dn^3)$, while the decision tree is $O(dn \log n)$. From this view, UTCDT uses the lower complexity but achieves comparable performance to SVM.

In summary, experimental results show that UTCDT indeed provides a new perspective to enhance the performance of the decision tree, and possesses the adaptability to datasets and the low algorithm complexity.

6. CONCLUSIONS

In this paper, we unify the three prevalent splitting criteria into a parametric framework through Tsallis entropy, and theoretically reveal the relations between Tsallis entropy and other splitting criteria. Based on the unified framework, we propose a Unified Tsallis Criterion Decision Tree algorithm (UTCDT) to enhance the performance of the decision tree. Experimental results indicate that, with an appropriate q, UTCDT achieves statistically significant improvement over the classical decision tree algorithms both on classification accuracy and tree size, even comparable performance to SVM. Owning simplicity and adaptability, UTCDT can be easily used to promote the performance of many real-world learning tasks, e.g., image interpolation [1] and speech synthesis [2].

7. REFERENCES

[1] Jun-Jie Huang and Wan-Chi Siu, "Fast image interpolation with decision tree," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1221–1225.

[2] Timo Baumann, "Decision tree usage for incremental parametric speech synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3819–3823.

[3] Lior Rokach and Oded Maimon, "Top-down induction of decision trees classifiers-a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 4, pp. 476–487, 2005.

[4] Wray Buntine and Tim Niblett, "A further comparison of splitting rules for decision-tree induction," *Machine Learning*, vol. 8, no. 1, pp. 75–85, 1992.

[5] Wei Zhong Liu and Allan P. White, "The importance of attribute selection measures in decision tree induction," *Machine Learning*, vol. 15, no. 1, pp. 25–41, 1994.

[6] J. Ross Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[7] J. Ross Quinlan, *C4. 5: programs for machine learning*, Morgan Kaufmann Publishers, 1993.

[8] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen, *Classification and regression trees*, CRC press, 1984.

[9] Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf, "A primer on kernel methods," *Kernel Methods in Computational Biology*, pp. 35–70, 2004.

[10] C.E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[11] Gerhard Jäger, "Power laws and other heavy-tailed distributions in linguistic typology," *Advances in Complex Systems (ACS)*, vol. 15, no. 03, 2012.

[12] Tomasz Maszczyk and Włodzisław Duch, "Comparison of shannon, renyi and tsallis entropy used in decision trees," in *17th International Conference on Artificial Intelligence and Soft Computing (ICAISC)*. 2008, pp. 643–651, Springer.

[13] Constantino Tsallis, "Possible generalization of boltzmann-gibbs statistics," *Journal of Statistical Physics*, vol. 52, no. 1-2, pp. 479–487, 1988.

[14] Constantino Tsallis, *Introduction to nonextensive statistical mechanics*, Springer, 2009.

[15] Constantino Tsallis, "Generalizing what we learnt: Nonextensive statistical mechanics," in *Introduction to Nonextensive Statistical Mechanics*, pp. 37–106. Springer, 2009.

[16] Angus E Taylor, "L'hospital's rule," *The American Mathematical Monthly*, vol. 59, no. 1, pp. 20–24, 1952.

[17] Moshe Lichman, "UCI machine learning repository," `http://archive.ics.uci.edu/ml`, 2013.

[18] Janez Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, no. Jan, pp. 1–30, 2006.

[19] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim, "Do we need hundreds of classifiers to solve real world classification problems," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.