Comparison of Shannon, Renyi and Tsallis Entropy used in Decision Trees

Tomasz Maszczyk and Włodzisław Duch

Department of Informatics, Nicolaus Copernicus University
Grudziądzka 5, 87-100 Toruń, Poland
{tmaszczyk,wduch}@is.umk.pl
http://www.is.umk.pl

Abstract. Shannon entropy used in standard top-down decision trees does not guarantee the best generalization. Split criteria based on generalized entropies offer different compromise between purity of nodes and overall information gain. Modified C4.5 decision trees based on Tsallis and Renyi entropies have been tested on several high-dimensional microarray datasets with interesting results. This approach may be used in any decision tree and information selection algorithm.

Key words: Decision rules, entropy, information theory, information selection, decision trees

1 Introduction

Decision tree algorithms are still the foundation of most large data mining packages, offering easy and computationally efficient way to extract simple decision rules [1]. They should always be used as a reference, with more complex classification models justified only if they give significant improvement. Trees are based on recursive partitioning of data and unlike most learning systems they use different sets of features in different parts of the feature space, automatically performing local feature selection. This is an important and unique property of general divide-and-conquer algorithms that has not been paid much attention. The hidden nodes in the first neural layer weight the inputs in a different way, calculating specific projections of the input data on a line defined by the weights $g_k = \boldsymbol{W}^{(k)} \cdot \boldsymbol{X}$. This is still non-local feature that captures information from a whole sector of the input space, not from the localized region. On the other hand localized radial basis functions capture only the local information around some reference points. Recursive partitioning in decision trees is capable of capturing local information in some dimensions and non-local in others. This is a desirable property that may be used in neural algorithms based on localized projected data [2].

The C4.5 algorithm to generate trees [3] is still the basis of the most popular approach in this field. Tests for partitioning data in C4.5 decision trees are based on the concept of information entropy and applied to each feature $x_1, x_2, ...$ individually. Such tests create two nodes that should on the one hand contain

Lecture Notes in Computer Science, Vol. 5097, 643-651, 2008.

data that are as pure as possible (i.e. belong to a single class), and on the other hand increase overall separability of data. Tests that are based directly on indices measuring accuracy are optimal from Bayesian point of view, but are not so accurate as those based on information theory that may be evaluated with greater precision [4]. Choice of the test is always a hidden compromise in how much weight is put on the purity of samples in one or both nodes and the total gain achieved by partitioning of data. It is therefore worthwhile to test other types of entropies that may be used as tests. Essentially the same reasoning may be used in applications of entropy-based indices in feature selection. For some data features that can help to distinguish rare cases are important, but standard approaches may rank them quite low.

In the next section properties of Shannon, Renyi and Tsallis entropies are described. As an example of application three microarray datasets are analyzed in the third section. This type of application is especially interesting for decision trees because of the high dimensionality of microarray data, the need to identify important genes and find simple decision rules. More sophisticated learning systems do not seem to achieve significantly higher accuracy. Conclusions are given in section four.

2 Theoretical framework

Entropy is the measure of disorder in physical systems, or an amount of information that may be gained by observations of disordered systems. Claude Shannon defined a formal measure of entropy, called Shannon entropy[5]:

$$S = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

where p_i is the probability of occurrence of an event (feature value) x_i being an element of the event (feature) X that can take values $\{x_1 ... x_n\}$. The Shannon entropy is a decreasing function of a scattering of random variable, and is maximal when all the outcomes are equally likely.

Shannon entropy is may be used globally, for the whole data, or locally, to evaluate entropy of probability density distributions around some points. This notion of entropy can be generalized to provide additional information about the importance of specific events, for example outliers or rare events. Comparing entropy of two distributions, corresponding for example to two features, Shannon entropy assumes implicite certain tradeoff between contributions from the tails and the main mass of this distribution. It should be worthwhile to control this tradeoff explicitly, as in many cases it may be important to distinguish weak signal overlapping with much stronger one. Entropy measures that depend on powers of probability, $\sum_{i=1}^{n} p(x_i)^{\alpha}$, provide such control. If α has large positive value this measure is more sensitive to events that occur often, while for large negative α it is more sensitive to the events which happen seldom.

Constantino Tsallis [6] and Alfred Renyi [7] both proposed generalized entropies that for $\alpha = 1$ reduce to the Shannon entropy. The Renyi entropy is defined as [7]:

$$I_\alpha = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} p_i^\alpha \right) \tag{2}$$

It has similar properties as the Shannon entropy:

- it is additive
- it has maximum $= \ln(n)$ for $p_i = 1/n$

but it contains additional parameter α which can be used to make it more or less sensitive to the shape of probability distributions.

Tsallis defined his entropy as:

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

Figures 1-3 shows illustration and comparison of Renyi, Tsallis and Shannon entropies for two probabilities p_1 and p_2 where $p_1 = 1 - p_2$.

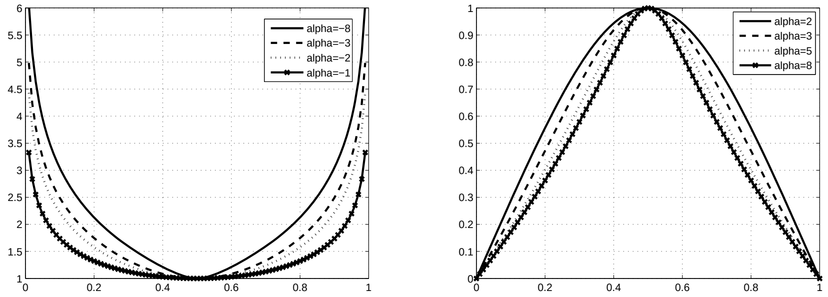

Fig. 1. Plots of the Renyi entropy for several negative and positive values of α.

The modification of the standard C4.5 algorithm has been done by simply replacing the Shannon measure with one of the two other entropies, as the goal here was to evaluate their influence on the properties of decision trees. This means that the final split criterion is based on the gain ratio: a test on attribute A that partitions the data D in two branches with D_t and D_f data, with a set of classes *omega* has gain value:

$$G(\omega, A|D) = H(\omega|D) - \frac{|D_t|}{|D|} H(\omega|D_t) - \frac{|D_f|}{|D|} H(\omega|D_f) \tag{4}$$

where $|D|$ is the number of elements in the D set and $H(\omega|S)$ is one of the 3 entropies considered here: Shannon, Renyi or Tsallis. Parameter α has clearly

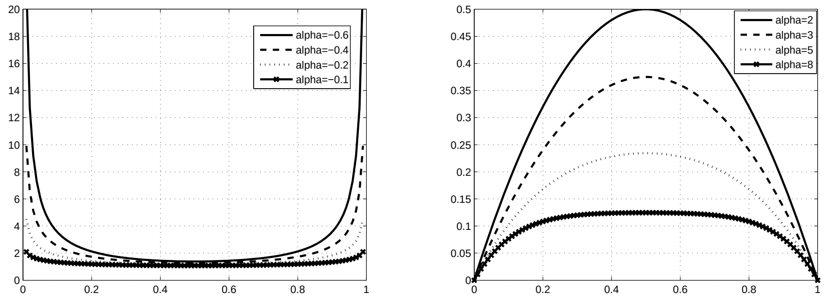

Fig. 2. Plots of the Tsallis entropy for several negative and positive values of α.

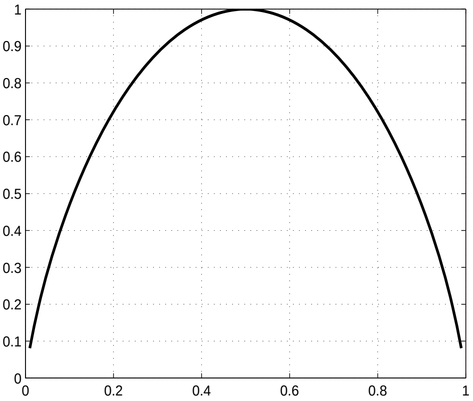

Fig. 3. Plot of the Shannon entropy

an influence on what type of splits are going to be created, with preference for negative values of α given to rare events or longer tails of probability distribution.

3 Empirical study

To evaluate the usefulness of Renyi and Tsallis entropy measures in decision trees the classical C4.5 algorithm has been modified and applied first to artificial data to verify its usefulness. The results were encouraging, therefore experiments on three data sets of gene expression profiles were carried out. Such data are characterized by large number of features and very small number of samples. In such situations several features may by pure statistical chance seem to be quite informative and allow for good generalization. Therefore it is deceivingly simple to reach high accuracy on such data, although it is very difficult to classify these data reliably. Only the simplest models may avoid overfitting, therefore decision trees providing simple rules may have an advantage over other models. A summary of these data sets is presented in Table 1, all data were downloaded from the Kent Ridge Bio-Medical Data Set Repository `http://sdmc.lit.org.sg/GEDatasets/Datasets.html`. Short description of these datasets follows:

1. Leukemia: training dataset consists of 38 bone marrow samples (27 ALL and 11 AML), over 7129 probes from 6817 human genes. Also 34 samples testing data is provided, with 20 ALL and 14 AML cases.
2. Colon Tumor: data contains 62 samples collected from the colon cancer patients. Among them, 40 tumor biopsies are from tumors (labeled as "negative") and 22 normal (labeled as "positive") biopsies are from healthy parts of the colons of the same patients. Two thousand out of around 6500 genes were pre-selected by the authors of the experiment based on the confidence in the measured expression levels.
3. Diffuse Large B-cell Lymphoma (DLBCL) is the most common subtype of non-Hodgkins lymphoma. The data contains gene expression data from distinct types of such cells. There are 47 samples, 24 of them are from "germinal centre B-like" group while 23 are "activated B-like" group. Each sample is described by 4026 genes.

Title	#Genes	#Samples	#Samples per class		Source
Colon cancer	2000	62	40 tumor	22 normal	Alon at all (1999) [8]
DLBCL	4026	47	24 GCB	23 AB	Alizadeh at all (2000) [9]
Leukemia	7129	72	47 ALL	25 AML	Golub at all (1999) [10]

Table 1. Summary of data sets

4 Experiment and results

For each data set the standard C4.5 decision tree is used 10 times, each time averaging the 10-fold crossvalidation mode, followed on the same partitioning of data by the modified algorithms based on Tsallis and Renyi entropies for different values of parameter α. Results are collected in Tables 2-7, with accuracies and standard deviations for each dataset. The best values of α parameter differ in each case and can be easily determined from crossvalidation. Although overall results may not improve accuracy for different classes they may strongly differ in sensitivity and specificity, as can be seen in the tables below.

Entropy	Alpha															
	-1.5	-0.9	-0.5	-0.1	0.1	0.3	0.5	0.7	0.9	1.1	1.3	1.5	2.0	3.0	4.0	5.0
Renyi	64.6±0.2	64.6±0.2	64.6±0.2	77.3±4.1	75.4±2.1	77.7±3.3	77.8±4.7	79.1±2.6	78.8±4.4	82.1±4.2	82.8±4.0	82.9±2.5	84.0±3.9	79.4±3.0	80.8±3.1	78.9±2.2
Tsallis	64.6±0.2	64.6±0.2	64.6±0.2	64.6±0.2	77.3±3.7	75.4±4.0	74.4±4.3	71.3±5.4	73.0±3.4	74.9±1.8	73.4±2.4	71.1±4.0	70.2±3.9	73.9±4.4	72.8±3.6	71.1±4.4
Shannon	81.2±3.7															

Table 2. Accuracy on Colon cancer data set; Shannon and $\alpha = 1$ results are identical.

Entropy	Class	Alpha													
		-1.5	-0.5	-0.1	0.1	0.3	0.5	0.7	0.9	1.1	1.5	2.0	3.0	4.0	5.0
Renyi	1	0.0±0.0	0.0±0.0	59.7±4.7	58.7±6.8	60.8±6.4	63.2±7.3	66.0±6.5	65.8±6.5	70.0±7.2	67.3±5.4	69.2±6.6	58.5±2.9	61.0±3.8	58.7±3.5
	2	100±0.0	100±0.0	87.2±4.1	84.7±2.4	87.2±2.8	85.8±4.0	86.2±3.9	85.8±4.6	88.8±4.5	91.5±2.1	92.1±2.9	90.7±4.3	91.6±3.7	90.1±3.6
Tsallis	1	0.0±0.0	0.0±0.0	0.0±0.0	58.2±5.1	59.8±9.3	59.8±5.2	50.7±10.2	55.7±7.7	58.5±4.9	58.3±8.2	53.2±7.6	67.2±9.2	60.5±7.4	60.0±10.4
	2	100±0.0	100±0.0	100± 0.0	87.6±4.8	83.9±4.3	82.8±4.4	82.6±3.9	82.8±3.4	84.2±2.2	78.9±4.6	80.0±3.7	77.7±6.0	79.7±5.3	77.3±3.4
Shannon	1	69.5±4.2													
	2	87.7±4.8													

Table 3. Accuracy per class on Colon cancer data set; Shannon and $\alpha = 1$ results are identical.

Results depend quite clearly on the α coefficient, with $\alpha = 1$ always reproducing the Shannon entropy results. The optimal coefficient should be deter-

Entropy	Alpha							
	-1.5	-0.9	-0.5	-0.1	0.1	0.3	0.5	0.7
Renyi	46.0±4.2	46.0±4.2	46.0±4.2	69.9±5.2	71.8±5.4	70.7±5.4	70.5±5.0	73.0±4.9
Tsallis	52.4±6.8	52.4±6.8	52.4±6.8	52.4±6.8	71.1±5.6	69.8±5.2	72.4±6.0	79.9±5.0
	Alpha							
	0.9	1.1	1.3	1.5	2.0	3.0	4.0	5.0
Renyi	76.5±6.7	81.0±6.2	81.0±4.8	80.5±5.0	79.3±5.1	79.5±5.6	75.9±7.2	69.7±6.3
Tsallis	81.3±4.7	82.0±4.3	81.8±5.2	80.8±6.5	81.5±5.7	78.8±6.9	81.8±4.1	80.5±4.0
Shannon	78.5±4.8							

Table 4. Accuracy on DLBCL data set

Entropy	Class	Alpha						
		-1.5	-0.5	-0.1	0.1	0.3	0.5	0.7
Renyi	1	90.0±10.5	90.0±10.5	72.3±8.4	75.5±10.6	74.3±11.2	76.3±8.7	79.7±7.1
	2	10.0±10.5	10.0±10.5	65.5±11.8	66.7±10.5	65.7±10.9	62.5±8.5	65.3±5.9
Tsallis	1	64.8±10.6	64.8±10.6	64.8±10.6	74.5±12.1	73.3±10.7	80.2±8.5	85.8±7.2
	2	41.8±11.0	41.8±11.0	41.8±11.0	65.7±10.6	65.2±12.0	64.8±9.1	74.7±9.8
	Class	Alpha						
		0.9	1.1	1.3	1.5	2.0	3.0	5.0
Renyi	1	82.7±8.4	85.5±8.1	86.5±5.8	85.2±5.8	84.8±6.4	84.2±5.3	68.0±12.0
	2	70.2±11.1	77.3±9.0	77.0±7.4	77.3±7.6	74.7±8.1	75.3±9.0	69.3±4.7
Tsallis	1	88.2±6.3	88.2±5.7	86.2±5.3	85.2±6.4	84.7±5.7	83.2±8.8	87.3±5.2
	2	76.0±7.5	77.3±5.3	78.7±6.9	77.8±9.3	80.0±6.9	76.3±6.4	75.3±4.9
Shannon	1	84.8±7.0						
	2	72.7±8.7						

Table 5. Accuracy per class on DLBCL data set

Entropy	Alpha						
	-1.5	-0.5	-0.1	0.1	0.3	0.5	0.7
Renyi	65.4 ±0.4	65.4 ±0.4	88.5 ±2.4	85.6 ±3.9	84.6 ±3.8	82.4 ±4.6	82.0 ±4.6
Tsallis	65.4 ±0.4	65.4 ±0.4	65.4 ±0.4	83.5 ±4.4	84.8 ±4.2	84.3 ±3.5	82.3 ±3.9
	Alpha						
	0.9	1.1	1.3	1.5	2.0	3.0	5.0
Renyi	80.5±3.8	81.5±3.5	82.2±3.5	82.4±2.6	85.3±2.8	86.1±2.8	83.8±2.0
Tsallis	82.5±4.4	81.5±2.9	82.3±1.1	83.3±1.4	82.2±2.5	86.5±2.7	87.5±3.6
Shannon	81.4±4.1						

Table 6. Accuracy on Leukemia data set

Entropy	Class	Alpha						
		-1.5	-0.5	-0.1	0.1	0.3	0.5	0.7
Renyi	1	100±0.0	100±0.0	89.2±1.6	89.4±3.2	88.7±2.5	86.2±3.2	84.8±4.0
	2	0.0±0.0	0.0±0.0	86.6±5.4	77.7±9.9	75.8±10.5	74.3±10.5	76.2±10.6
Tsallis	1	100±0.0	100±0.0	100±0.0	88.3±3.6	88.7±2.9	89.0±3.3	85.4±3.2
	2	0.0±0.0	0.0±0.0	0.0±0.0	73.5±10.1	76.8±10.3	74.8±9.3	76.6±10.3
	Class	Alpha						
		0.9	1.3	1.5	2.0	3.0	4.0	5.0
Renyi	1	85.0±3.9	84.7±4.6	85.2±3.7	88.6±4.5	90.2±3.4	83.9±3.9	86.8±2.0
	2	71.3±11.9	77.4±5.0	76.7±5.7	79.1±4.6	78.3±3.7	80.3±7.1	78.2±5.9
Tsallis	1	84.5±3.9	86.1±3.4	87.4±4.0	85.0±3.4	90.0±3.8	89.1±3.0	91.3±3.5
	2	78.1±11.8	74.4±7.5	75.2±7.8	77.9±7.0	80.2±6.7	84.3±6.2	80.3±6.5
Shannon	1	83.8±5.3						
	2	76.6±5.7						

Table 7. Accuracy per class on Leukemia data set

mined through crossvalidation. For the Colon dataset (Tab. 2-3) peak accuracy is achieved for Renyi entropy with $\alpha = 2$, with specificity (accuracy of the second class) significantly higher than for the Shannon case, and with smaller variance. Tsallis entropy seems to be very sensitive and does not improve over Shannon value around $\alpha = 1$. For DLBCL both Renyi and Tsallis entropies with α in the range $1.1 - 1.3$ give the best results, improving both specificity and sensitivity of the Shannon measure (Tab. 4-5). For the Leukemia data best Renyi result for $\alpha = -0.1$, around 88.5 ± 2.4 is significantly better than Shannon's $81.4 \pm 4.1\%$, improving both the sensitivity and specificity and decreasing variance; results with $\alpha = 3$ are also very good. Tsallis results for quite large $\alpha = 5$ are even better.

Below one of the decision trees generated on the Leukemia data set with Renyi's entropy and $\alpha = 3$ is presented:

```
g760 > 588 : AML
g760 <= 588 :
|   c1926 <= 134 : ALL
|   c1926 > 134 : AML
```

Rules extracted from this tree are:

```
Rule 1:
     g760 > 588
-->  class AML  [93.0%]

Rule 2:
     g1926 > 134
-->  class AML  [89.1%]
```

```
Rule 3:
     g760 <= 588
     g1926 <= 134
-->  class ALL  [93.9%]
```

The trees are very small and should provide good generalization for small datasets, avoiding overfitting that other methods may suffer from. However, for gene expression data this may not be sufficient as the results are not stable against small perturbation of data (all learning methods suffer from this problem, see [?]). Therefore in practical applications a better approach is to define approximate coverings of redundant features and replace such groups of features with their linear combinations to reduce their dimensionality, aggregating information about similar genes (Biesiada and Duch, in preparation). For the demonstration of efficiency of non-standard entropy measures this is not so important.

5 Conclusions

Information Theoretic Learning (ITL) has been used to train neural networks for blind source separation [11], definition of new error functions in neural networks [12], classification with labeled and unlabeled data, feature extraction and other applications [13]. The quadratic Renyi's error entropy has been used to minimize the average information content of the error signal for supervised adaptive system training. However, the use of non-Shannon entropies in extraction of logical rules using decision trees has not yet been attempted. The importance of decision trees in large-scale data mining knowledge extraction applications and the simplicity of this approach encouraged us to explore this possibility.

First experiments with modified split criterion of the C4.5 decision trees were presented here. Theoretical results and tests on artificial data show that this can be useful particularly for datasets with one or more small classes. Additional parameter α that may be easily adapted using crossvalidation tests gives possibility to tune the tree to the discrimination of classes of different size. This algorithm makes it more attractive than standard approach based on Shannon entropy that does not allow for exploration of the tradeoff between the probability of different classes and the overall information gain. This opens a way to applications in many types of decision trees, encouraging also modification of split criteria that are not based on entropy to account for the same tradeoff. An explicit formula for such tradeoff may be defined, aimed at separation of nodes of high purity at the expense of lower overall information gain. This as well as applications of non-standard entropies to the text classification data, where small classes are very important, remain to be explored. Bearing in mind the results and the simiplicity of the approach presented here the use of non-standard entropies may be highly recommended.

References

1. Duch, W., Setiono, R., Zurada, J.: Computational intelligence methods for understanding of data. Proceedings of the IEEE **92**(5) (2004) 771–805
2. Grochowski, M., Duch, W.: Learning highly non-separable boolean functions using constructive feedforward neural network. Lecture Notes in Computer Science **4668** (2007) 180–189
3. Quinlan, J.: C 4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA (1993)
4. Duch, W.: Filter methods. In Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., eds.: Feature extraction, foundations and applications. Physica Verlag, Springer, Berlin, Heidelberg, New York (2006) 89–118
5. Shannon, C., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana, Ill. (1964)
6. Tsallis, C., Mendes, R., Plastino, A.: The role of constraints within generalized nonextensive statistics. Physica **261A** (1998) 534–554
7. Renyi, A.: Probability Theory. North-Holland, Amsterdam
8. Alon, U.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS **96** (1999) 745–750
9. Alizadeh, A.: Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature **403** (2000) 503–511
10. Golub, T.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531–537
11. Hild, K., Erdogmus, D., Principe, J.: Blind source separation using renyi's mutual information. IEEE Signal Processing Letters **8** (2001) 174–176
12. Erdogmus, D., Principe, J.: Generalized information potential criterion for adaptive system training. IEEE Trans. on Neural Networks **13** (2002) 1035–1044
13. Hild, K., Erdogmus, D., Torkkola, K., Principe, J.: Feature extraction using information-theoretic learning. IEEE Trans. on Pattern Analysis and Machine Intelligence **28** (2006) 1385–1392