Norm-induced entropies for decision forests

Christoph Lassner Rainer Lienhart
Multimedia Computing and Computer Vision Lab, University of Augsburg
Christoph.Lassner@informatik.uni-augsburg.de

Abstract

The entropy measurement function is a central element of decision forest induction. The Shannon entropy and other generalized entropies such as the Rényi and Tsallis entropy are designed to fulfill the Khinchin-Shannon axioms. Whereas these axioms are appropriate for physical systems, they do not necessarily model well the artificial system of decision forest induction.

In this paper, we show that when omitting two of the four axioms, every norm induces an entropy function. The remaining two axioms are sufficient to describe the requirements for an entropy function in the decision forest context. Furthermore, we introduce and analyze the p-norm-induced entropy, show relations to existing entropies and the relation to various heuristics that are commonly used for decision forest training.

In experiments with classification, regression and the recently introduced Hough forests, we show how the discrete and differential form of the new entropy can be used for forest induction and how the functions can simply be fine-tuned. The experiments indicate that the impact of the entropy function is limited, however can be a simple and useful post-processing step for optimizing decision forests for high performance applications.

1. Introduction

While decision trees and forests of arbitrary depth can express any possible concept in their domain, their inductive bias is enforced by preferring less deep trees over deeper trees. This bias finds its manifestation in the optimization of the node split function $h(\mathbf{v}, \theta)$, where $\mathbf{v}$ is a data vector and θ is the vector of split parameters (this is the notation of Criminisi and Shotton [8], which we will use throughout this paper). Whereas many optimization strategies are possible to use, usually an entropy (impurity) measurement is done for a set of training samples to split, and an information gain (loss of entropy) is maximized for possible splits.

The *Shannon entropy*, being an important measure in information theory and physics, has been shown to uniquely fulfill the *Khinchin-Shannon axioms* [15, 21]. Relaxing the additivity axiom, it can be shown that the *Shannon entropy* is a member of the more general family of *Rényi entropies*, which are important for dynamical systems theory (see, *e.g.*, [4]). Dropping the additivity axiom completely, the *Tsallis entropy* [24] becomes a sensible choice of generalized entropies [1], which has important applications in modeling complex and dynamical systems as well.

The two generalized entropies have been developed by carefully analyzing the requirements of the modeled systems and adjusting the axiomatic foundation of the entropies accordingly. In this paper, we show that by relaxing a second of the four axioms, a new meaningful family of entropy functions arises that is adapted to the requirements of decision forest induction. Furthermore, we derive the differential equivalent of this entropy function family and apply it for inducing regression and Hough forests [12].

The new family of entropy functions is continuously parameterized. This allows for simple fine-tuning of the entropy function to the induction task. At the same time, they only need 50% to 70% of the calculation time of the Shannon entropy.

To give the reader an impression of the influence on various machine learning tasks, we performed several experiments: (1) we did five classification experiments on computer vision datasets. (2) We performed two regression experiments and, (3) combining both scenarios, we applied the entropies for training Hough forests on one detection and localization and two human pose estimation datasets.

Our contributions are as follows:

- we show that two of the Khinchin-Shannon axioms are sufficient to model the requirements for an entropy function for decision forest training. This leads to a well-defined family of generalized entropies: the norm-induced entropies.
- we introduce the *induced entropy* and analyze its properties,
- we demonstrate how this new entropy can be optimized and be used to potentially improve scores on various datasets.

2. Related work

There are two kinds of related work that will be discussed: related work with respect to the generalization of entropies and the split evaluation functions for decision forests.

2.1. Entropy generalization

For a system with W states with probabilities $\mathbf{p}$, the classical Shannon entropy is defined as $S(\mathbf{p}) = -\sum_{i=1}^{W} p_i \cdot \log_2 p_i$ [1]. The first generalization attempt is by Alfréd Rényi [20]. His parameterized Rényi entropy preserves the additivity (but only for independent variables in general) and is equivalent to the original Shannon entropy for $q = 1$. It is defined as $R_q(\mathbf{p}) = \frac{1}{1-q} \log_2 \left(\sum_{i=1}^{W} p_i \right)$.

Constantino Tsallis developed the Tsallis entropy [24] for non extensive systems: it drops the additivity axiom (but maintains a so called pseudo additivity). It is defined as $T_q(\mathbf{p}) = \frac{1}{q-1} \left(1 - \sum_{i=1}^{W} p_i^q \right)$ and is equivalent to the Shannon entropy for $q = 1$ as well [24]. Comparisons to the Shannon, Rényi and Tsallis entropies are included in our experiments.

Sharma and Mittal developed the *Sharma-Mittal entropy* [22]. Their generalization, and to the best of the authors' knowledge the most recent generalization of *supra-extensive entropy* by Marco Masi [18], are the strongest generalizations so far. Both of these entropies generalize Tsallis and Rényi entropy, and hence do not guarantee additivity. They both have two continuous parameters: this deviates from our aim of developing an easy to optimize entropy for decision forest induction.

All of the aforementioned approaches leave the first three axioms untouched and, but the Tsallis entropy, make use of the $\log$ function. We will show in this paper, that the setting in which decision forests work can well be modeled by the first two axioms and does not require to use the compute intensive $\log$.

2.2. Splitting criteria for decision forests

Yu-Shan Shih did a comprehensive review of splitting criteria for classification trees in [23]. A more recent overview is contained in [17].

For regression forests, considerably fewer split optimization criteria are used, depending on the loss function used. For the standard mean-squared error loss, usually the negative sum of variances is maximized (which can be written as $V(\Sigma) = -\text{tr}(\Sigma)$, where Σ is the covariance matrix of the samples). Alternatively, the Least Absolute Deviation (LAD) from the mean can be used, which is less vulnerable to outliers. For a comprehensive review up to the year 2004, see [6]. Criminisi and Shotton use the differential Shannon entropy to estimate the quality of fit of linear models in decision forests [8].

The use of Rényi and Tsallis entropies for decision forest induction have been evaluated briefly in [19]. The evaluation is done on three datasets with less than 80 samples. As a result, the trees reach a depth of about three. We consider this evaluation as not representative for computer vision and extend it to datasets with up to 74000 samples with trees of depth up to 20. Additionally, we explore the use of the differential versions of these entropies for regression and Hough forests.

3. Induced entropies for discrete systems

3.1. The Khinchin-Shannon axioms

The Khinchin-Shannon axioms (KS) for an entropy functional H over the probabilities $\{p_i\}_{i=1,2,...,W}$ are defined as follows:

(KS1) $H(p_1, p_2, \ldots, p_W)$ is continuous with respect to all of its arguments.

(KS2) H takes its maximum for the equiprobability distribution $p_i = \frac{1}{W}$, $i = 1, ..., W$.

(KS3) $H(p_1, p_2, \ldots, p_W, 0) = H(p_1, p_2, \ldots, p_W)$.

(KS4) Given two systems described by two independent probability distributions A and B,

$$H(A \cap B) = H(A) + H(B|A),$$

where $H(B|A) = \sum_{i=1}^{W} p_i(A) H(B|A = A_i)$.

KS4 is the additivity axiom. It is relaxed to hold only for independent distributions in the Rényi entropy and to pseudo-additivity for the Tsallis entropy (and a mix of both relaxations for Sharma-Mittal and supra-extensive entropies).

KS3 states that an additional state with probability 0 does not change the entropy of the system. This is a possible, but not necessary axiom in the context of decision forests: it is known for the entire forest training what states are consistently possible.

Moreover, arguably an entropy violating KS3 might be more appropriate in some contexts than one abiding KS3: a system with three states in a configuration with two equally probable states and one with probability zero might have a lower entropy (be more ordered) than a system with only two equally probable states. The distinction becomes philosophical: is a state with probability zero different from 'no state' by definition?

[1] We denote the various entropy functions with different function names and depart from the classical entropy notion of $H(\mathbf{p})$ for the sake of an easier description. Additionally, we denote the entropy parameter for all generalized entropies with q for a more consistent notation.

3.2. Definition

For the following definition, it is assumed that only KS1 and KS2 must hold. A function fulfilling these two axioms must be continuous in all of its arguments, assuming its maximum at the point of equiprobability.

A particularly interesting family of measures arises, when using the negative sum of absolute distances to the point of equiprobability. It can be parameterized with a power parameter q:

$$\tilde{N_q}(\mathbf{p}) = -\sum_{i=1}^{W} \left| p_i - \frac{1}{W} \right|^q . \tag{1}$$

In particular, all norm-induced metrics that measure the distance to the point of equiprobability can now be used as entropies. KS1 holds in this case, since a norm is by definition uniformly continuous in all of its arguments. KS2 holds as well, because of the positivity and the zero vector property. The p-norms can be used to rephrase Equation 1:

$$\tilde{N_q}(\mathbf{p}) = -\left\| \mathbf{p} - \mathbf{e} \right\|_q^q , \tag{2}$$

where $\mathbf{e}$ is the point of equiprobability with $e_i = \frac{1}{W} \forall i$. KS1 and KS2 still hold in this case, since taking the power retains the zero vector, positivity and continuity properties of the norm.

However, this entropy is always ≤ 0 and is $= 0$ at the point of equiprobability. This can simply be avoided by adding the minimal value as an offset, so that the entropy is 0 at the points of perfect order, and otherwise > 0. This offset must be the value at the lowest points. Defining $\mathbf{u}$ as $u_1 = 1$, $u_i = 0 \, \forall i > 1$, this can be specified as

$$\left\| \mathbf{u} - \mathbf{e} \right\|_q^q . \tag{3}$$

Combining Equations 2 and 3, the following formula describes the full induced entropy for discrete systems:

$$N_q(\mathbf{p}) = \left\| \mathbf{u} - \mathbf{e} \right\|_q^q - \left\| \mathbf{p} - \mathbf{e} \right\|_q^q . \tag{4}$$

We will refer to this entropy as 'induced entropy'. You can find a comparison of the characteristic plots of the induced entropy and the Shannon, Rényi and Tsallis entropies for a two state system in Figure 1. For the induced entropy, it is impossible to recover the Shannon entropy due to the use of the $\log$ function in its definition. The closest fit (MSE) for a two state system is reached for the value $q \approx 2.60068$.

3.3. Properties

As argued before, KS1 and KS2 hold for all norm induced entropies, and for N_q as well.

3.3.1 Concavity

For values of $q \geq 1$, N_q is a strictly concave function. Remember that, by definition, the inducing norm is Schur convex for $q \geq 1$. Since that norm only occurs negative with an exponent ≥ 1 in N_q, the result is a Schur concave function.

Concavity implies thermodynamic stability [18], hence is a desirable property if the entropy should be applied on a physical system.

For values of $q \in]0; 1[$, the function is still well-defined. For these values, the inducing term loses its property as norm, because the triangle inequality does not hold any more. However, since the term only occurs to the power of q, the resulting function still defines a metric. For non-physical systems, these values might still be of interest.

3.3.2 Lesche stability

Lesche stability is claimed to be a necessary condition for an entropy to be a physical quantity [16]. There are some disputes recently about whether it is really a necessary condition or not [25]. Proving Lesche stability is a non-trivial task and beyond the scope of this work. We note, however, that the Tsallis entropy has been proven to be Lesche stable by Abe [2] for positive values of q. As we will show in the following Section 3.4, the induced entropy is equivalent for $q = 2$ to the Gini measure and T_2.

3.4. Equivalence of N_2, the Gini measure and T_2

The Gini measure of a discrete set of probabilities is defined as $\sum_{i=0}^{W} p_i^2$. Nevertheless, in the context of decision forests, it is used as an entropy-like function as $1 - \sum_{i=1}^{W} p_i^2$. Curiously, in this form it is equivalent to the Tsallis entropy for $q = 2$:

$$T_q(\mathbf{v}) = \frac{1}{q-1} \left(1 - \sum_{i=1}^{W} p_i^q \right) . \tag{5}$$

Similarly, it can be shown that $T_2(\mathbf{v}) = N_2(\mathbf{v})$:

$$\begin{aligned} \left(1 - \frac{1}{W}\right)^2 + (W-1)\left(\frac{1}{W}\right)^2 - \sum_{i=1}^{W} \left| p_i - \frac{1}{W} \right|^2 &= \\ 1 - \frac{1}{W} - \frac{1}{W} + \frac{2}{W} \sum_{i=1}^{W} p_i - \sum_{i=1}^{W} p_i^2 &= \\ 1 - \sum_{i=1}^{W} p_i^2 . \end{aligned} \tag{6}$$

For N_2 KS3 holds (since it holds for the Tsallis entropy), as well as pseudo-additivity. However, it is easy to find counterexamples for other values of q: the properties do not hold in general for norm induced entropies or N_q.

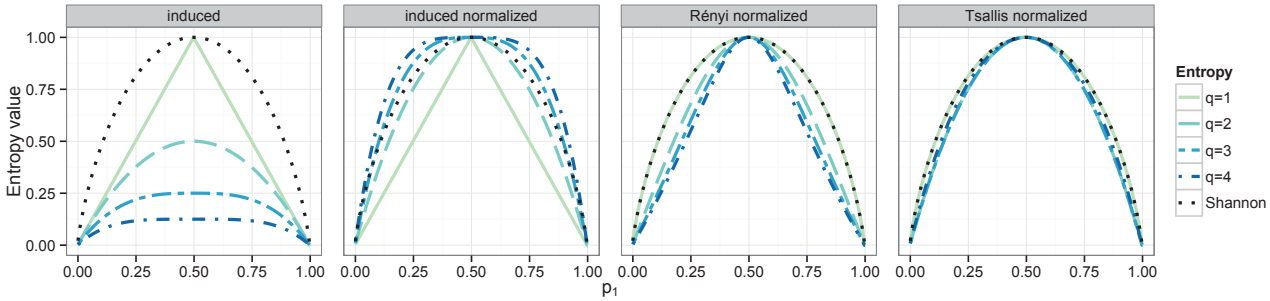


Figure 1: Entropy values for a two state system ($p_2 = 1 - p_1$) for the Shannon entropy and generalized entropies. The generalized entropies have been maximum normalized except for the leftmost plot.

3.5. Relation to the classification error

For $q \to 1$ the Tsallis entropy converges to the usual Shannon entropy. The induced entropy converges against a measure similar to the classification error. The classification error measures the 'ratio of misclassification' if it is assumed that the system is in its most probable state. It is defined as

$$C(\mathbf{p}) = 1 - \max_i p_i. \tag{7}$$

The induced entropy is proportional to the classification error for $W = 2$ states: $N_1(\mathbf{p}) = 2 \cdot C(\mathbf{p})$. In general:

$$N_1(\mathbf{p}) = 2\frac{W-1}{W} - \sum_{i \,|\, p_i < \frac{1}{W}} \left(p_i - \frac{1}{W}\right) + \sum_{i \,|\, p_i \geq \frac{1}{W}} \left(p_i - \frac{1}{W}\right). \tag{8}$$

For $W = 2$ classes and equiprobability, $N_1(\mathbf{p}) = 1$. For $W = 2$ classes and all other cases, each of the two sums runs over one element and it follows:

$$\begin{aligned} N_1(\mathbf{p}) = 1 + \sum_{i \,|\, p_i < \frac{1}{W}} p_i - \sum_{i \,|\, p_i \geq \frac{1}{W}} p_i &= \\ 1 + \left(1 - \max_i p_i\right) - \max_i p_i &= \\ 2 \cdot \left(1 - \max_i p_i\right). \end{aligned} \tag{9}$$

For $W > 2$, $N_1(\mathbf{p})$ is equal to the classification error as long as only one state has a higher probability than $\frac{1}{W}$. As an example, the characteristic plot for a three state system with the probability of state $p_3 = 0$ is given in Figure 2. As long as the system entropy gets closer to the point of equiprobability (until $p_1 = \frac{1}{3}$), the entropy is rising. From that point on, the city block distance to $\mathbf{e}$ does not change any more. This results in a constant entropy value. The contrary effect can be observed for $p_1 \geq \frac{2}{3}$.

This is an interesting, and maybe desired property for some systems. In the context of decision forest induction,

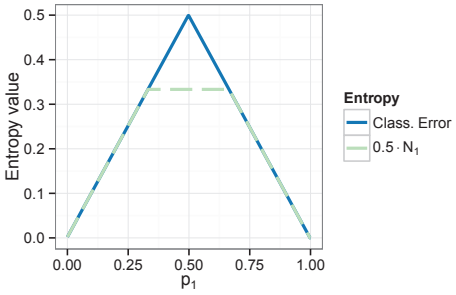


Figure 2: A comparison for classification error and $0.5 \cdot I_1$ for a three state (!) system. p_3 has probability 0 and p_1 and $p_2 = 1 - p_1$ varies.

we recommend to use N_1 only for few classes, but use it in our evaluations for all datasets to give an impression of its performance.

3.6. Implementation remarks

When using entropies for decision forest induction, they are usually used inside of a gain calculation function, which is merely a linear combination. The gain value itself is then used inside an $\arg\max$ function, *e.g.*, for feature or threshold selection.

Since the result of that function is invariant to scaling and shifting, all of the aforementioned entropy families may be used without their scaling and shifting terms. This form of the induced entropy remains particularly compact and efficient to evaluate:

$$\hat{N_q}(\mathbf{p}) = -\left\|\mathbf{p} - \mathbf{e}\right\|_q^q. \tag{10}$$

4. Differential induced entropy

The theory introduced so far applies to systems with discrete states. Criminisi and Shotton have proposed to use the differential Shannon entropy to evaluate splits for regression forests [8]. Extending this idea, we introduce the differential version of the induced entropy.

4.1. Definition

Developing the differential version of the entropy is mathematically expressed as $W \to \infty$. The first notable property arises, when examining the normalization offset:

$$\lim_{W\to\infty} \|\mathbf{u}-\mathbf{e}\|_q^q = \begin{cases} \infty & \text{for } q \in [0;1[, \\ 2 & \text{for } q = 1, \\ 1 & \text{for } q > 1. \end{cases} \tag{11}$$

When examining the rest of the formula, a close correspondence to the Tsallis entropy becomes apparent. With $\lim_{W\to\infty} \frac{1}{W} = 0$, it becomes (compare to Equation 1):

$$\int |p(x)|^q \, dx = \int p(x)^q dx. \tag{12}$$

The full differential induced entropy is thus defined as:

$$N_q[p] = \begin{cases} 2 - \int p(x)dx & \text{for } q = 1, \\ 1 - \int p(x)^q dx & \text{for } q > 1. \end{cases} \tag{13}$$

This is close to the definition of the differential Tsallis entropy:

$$T_q[p] = \frac{1}{q-1}\left(1 - \int p(x)^q dx\right). \tag{14}$$

For $q > 1$, both distributions are equivalent but for the factor $\frac{1}{q-1}$. For $q = 1$, the differential induced entropy becomes uninformative, since $\int p(x)dx = 1$.

4.2. The normal distribution

The most interesting probability distribution, for which the induced entropy will be derived here, is the normal distribution. Assuming $p(x) = N(x;\mu,\sigma)$, the integral becomes:

$$\int N(x;\mu,\sigma)^q dx = \frac{1}{\sqrt{q}} \cdot \left(\sqrt{2\pi}\sigma\right)^{-(q-1)}. \tag{15}$$

Especially interesting for decision forest induction is the multivariate case. It is required for multivariate regression and for Hough forest induction, since the offset regression is done two-dimensional. The formula for an n-dimensional Gaussian with mean μ and covariance matrix Σ is:

$$\int N(\mathbf{x};\mu,\Sigma)^q dx = \frac{1}{\sqrt{q^n}} \cdot \left(\left(\sqrt{2\pi}\right)^n \sqrt{|\Sigma|}\right)^{-(q-1)}, \tag{16}$$

where $|.|$ denotes the determinant. Summing up, the differential induced entropy for a normal distribution is defined for $q > 1$ as:

$$N_q[p] = 1 - \frac{1}{\sqrt{q^n}} \cdot \left(\left(\sqrt{2\pi}\right)^n \sqrt{|\Sigma|}\right)^{-(q-1)}, \tag{17}$$

where $|\Sigma|$ is σ^2 in the one-dimensional case.

Name	Samples	Classes	Features	Test size
chars74k [11]	74107	62	64	7400 (10%)
g50c [7]	550	2	50	500 (91%)
letter [3]	35000	26	16	8750 (25%)
MNIST [3]	70000	10	784	10000 (14%)
USPS [13]	9298	10	256	2007 (22%)

Table 1: Classification dataset characteristics.

5. Experiments

We conducted several experiments for the main application areas of decision forests to evaluate how to best apply the generalized entropy families[2]. In each experiment we did a grid search using cross-validation with the Shannon entropy to determine the decision forest parameters on the training set. The grid contained the following values for all experiments: depth 15, 20, 25; feature tests per node 7, 10, 15; thresholds tests per feature 4, 7, 10. Each score was then determined by 10 training/testing runs with different random seeds (except for the *g50c* and *Boston housing* datasets, where due to their small size 250 runs were done). The test set was always selected as by convention for the respective dataset, if available. We designed this setup, since we noted that the entropy family has none or hardly any effect on the parameters selected by the grid search, and it is a realistic usage scenario.

5.1. Classification

For the classification setting, we selected five computer vision datasets with varying characteristics. You can find an overview in Table 1. Plots of the resulting scores for the standard Shannon entropy and various parameters for Rényi, Tsallis and induced entropy can be found in Figure 3.

As evaluation measure we used the *F1-score*[3]. It is a reliable measure even for imbalanced datasets, especially when dealing with many classes. Each of the plot facets shows the results for one dataset for each entropy. The Shannon entropy has no parameter, hence is always visible as a straight line. Rényi and Tsallis entropy are equivalent to the Shannon entropy for $q = 1$, so all three entropies have the same value at this position.

We show the parameter range from $q \in [1;5]$. Measurements were taken at each step of 0.5 and are interpolated by applying a Loess smoother, including the 95% confidence interval indicated by a light gray background. We only show results for this parameter range because of limited space, but note that they are representative. Only for the *g50c* dataset we noted a peak for Rényi and Tsallis entropy close to zero that is comparably high to the peak of the induced entropy.

[2]Implementation: `http://www.fertilized-forests.org`.

[3]$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$, the harmonic mean of precision and recall.

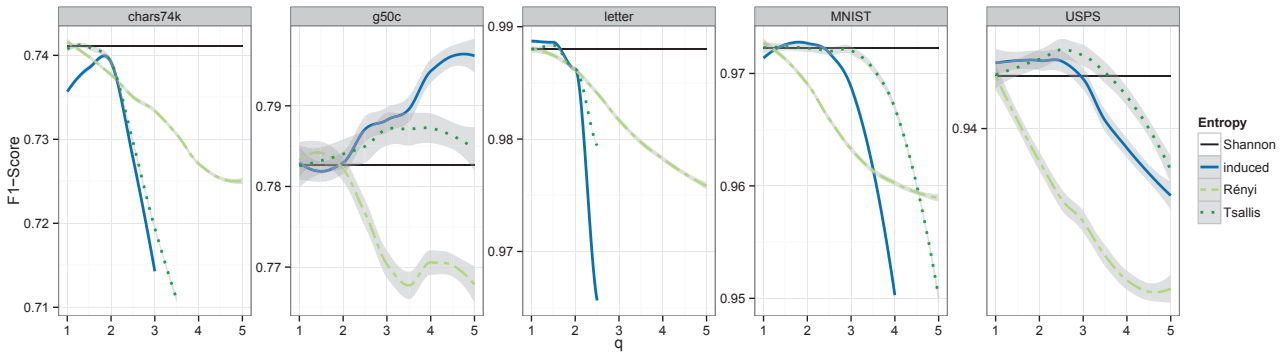


Figure 3: Results of using the various entropies on the classification datasets.

Name	Samples	σ^2	Features	Test size
abalone [3]	4177	10.4	9	1044 (25%)
Boston housing [3]	506	84.4	14	51 (10%)

Table 2: Regression dataset characteristics.

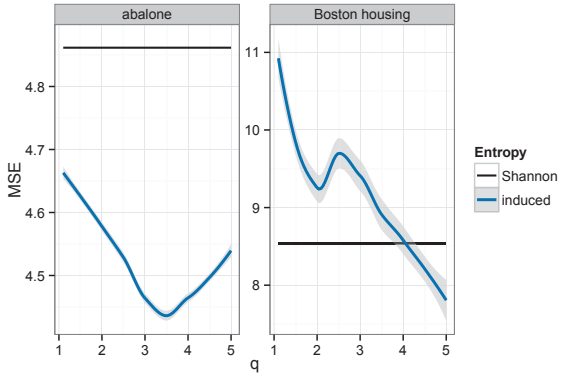


Figure 4: Results on the two regression datasets. Shannon and Rényi entropy results, and induced and Tsallis entropy results are equivalent in the visualized range.

As forest configuration, we used 100 trees, with varying parameters as determined by the grid search. The largest effect is observed on the *g50c* dataset with an improvement of about 1.9%. An improvement can be noted on all datasets.

5.2. Regression

Regression is, especially in computer vision, not as common as classification. Hence, we selected two non-vision datasets to cover the topic (the characteristics can be found in Table 2), and applied the differential entropies in a computer vision setting using Hough forests (see Section 5.3).

For regression, we used the mean squared error (MSE) as evaluation measure. You can find the results in Figure 4. Rényi and Tsallis entropies are omitted in the plot, since their results are equivalent to the ones of Shannon and induced entropies respectively in the plotted range.

As forest configuration, we again used 100 trees with varying parameters as determined by the grid search. On both datasets, the scores could consistently improved by using the induced entropy. On the *abalone* dataset, the performance of the Shannon entropy could consistently be outperformed.

While the induced entropy is only defined for $q \geq 1$, Rényi and Tsallis entropies are also defined for values in $[0; 1]$. Again, we checked the performance in these areas: the Tsallis entropy performs better than the Shannon entropy there, but does not reach the performance of the induced entropy in the plotted range.

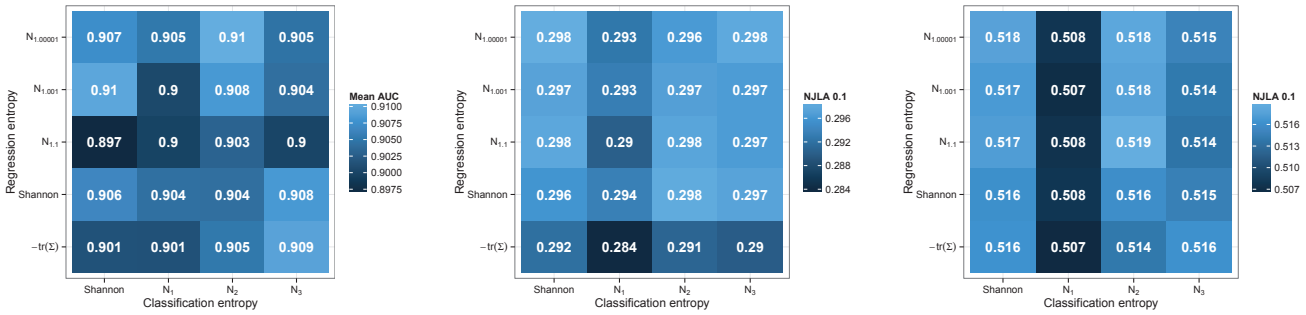
5.3. Hough forests

Hough forests have so far in literature only been used with the regression optimization measure $-\mathrm{tr}(\Sigma)$. We extend the Hough forest approach by assuming a Gaussian distribution of the offset vectors and evaluating its covariance matrix with a 'proper' entropy measure. As evaluation datasets, we used the *Weizmann horse* [5] dataset for detection and localization, and the *Leeds Sports Pose* (LSP) [14] and *FashionPose* [9] datasets for human pose estimation.

As regression entropy, we used the induced entropy: the results are equivalent to the ones of the Tsallis entropy in the given range, $R_{q\geq 1} = S$, and $R_{q<1}, T_{q<1}$ produced worse scores in our former regression experiments. For classification, we used the induced entropy as well. Remember that $N_2 = T_2$, and $R_1 = T_1 = S$. By exploiting these equivalences, we reach the most expressive set of results. Additionally, we evaluated R_2, R_3 and T_3 on the Weizmann horse dataset with worse results than with the corresponding N_q entropies.

5.3.1 The Weizmann horse dataset

The Weizmann horse dataset contains images of horses in slightly varying scales. We chose a 'single detection and localization' setting with a ROC area under curve evaluation criterion. We used a forest configuration similar to [12].



(a) Results on the Weizmann horse dataset. (b) Results on the LSP dataset. (c) Results on the FashionPose dataset.

Figure 5: Results for applying the induced entropy for Hough forest training.

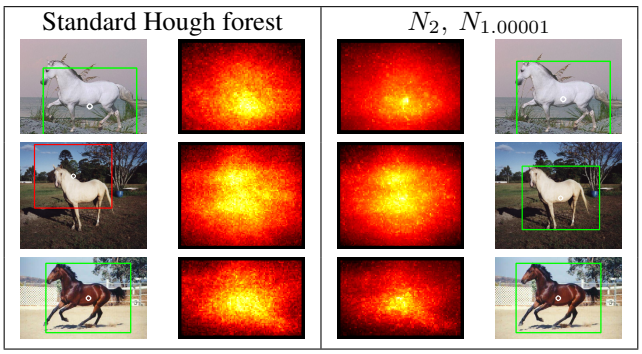


Figure 6: Detections and Hough forest maps on the Weizmann horse dataset [5].

Figure 5a shows the result matrix. The score for the default Hough forest configuration is located in the lower left corner (0.901). The best scores (0.91) can be reached for two entropy combinations with q values very close to 1 for the regression entropy. We found that for the Hough forest task, q values close to 1 produce the best results.

Figure 6 shows a comparison of results with standard training and with training using the induced entropy with the best performing parameters: the resulting maps are denser and reach a higher concentration at their maxima.

5.3.2 The pose estimation datasets

We were able to improve the Hough forest training time of three hours on a 700 CPU cluster reported in [10] to two and a half hours on a 64 CPU cluster. Since the experiments remained time consuming, we did five training/testing runs for each configuration on both datasets.

As evaluation measure we used the *normalized joint localization accuracy*, as introduced in [9] at the threshold 0.1. This value measures the percentage of correctly localized joints with an allowed offset of up to 0.1 times the upper-body size. This roughly corresponds to joints that would be considered correct by a human evaluator.

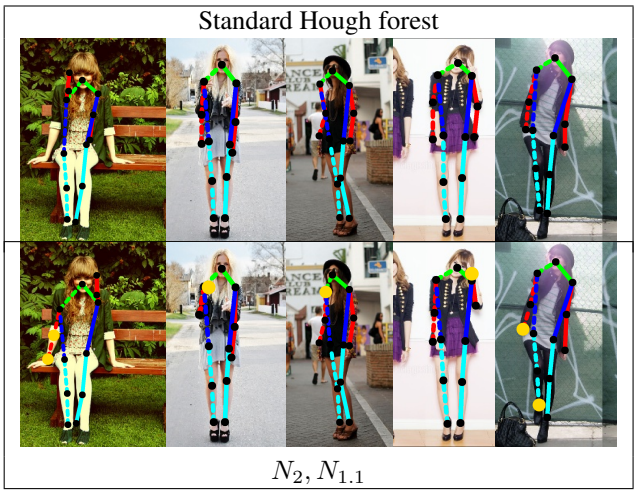


Figure 7: Comparison of Hough forest results on the FashionPose dataset [9]. Improved joint localizations are highlighted with a yellow marker.

The forest configuration was similar to [9]. Of the three presented methods in that paper, we used the *independent joint regression* method to directly show the performance of the Hough forests on the data. We applied a clustered pictorial structure model on the resulting probability maps as described in [9].

The plots of the results for the two datasets can be found in Figures 5b and 5c. On the LSP dataset, the performance could be improved from 0.292 to 0.298 for five configurations. On the FashionPose dataset, the score of 0.516 reached by the standard training method could be improved to 0.519.

Figure 7 shows five poses estimated with classical and modified Hough forests for a qualitative comparison. The pictorial structure model profits of the denser result maps, which mainly results in improved joint localization for the extremities.

6. Conclusion

Reducing the Khinchin-Shannon axioms to the most necessary ones for decision forest training, we introduced the new 'induced entropy' in its discrete and differential forms. We analyzed its properties and showed various connections to the already established generalized entropies, namely the Rényi and Tsallis entropy.

In three series of experiments on classification, regression and Hough forest tasks, we showed the influence of using the aforementioned three entropies compared with the standard Shannon entropy. In all experiments we achieved an improvement of scores.

While the experimental results do not allow to make a clear recommendation on when to use a specific entropy, we note that by using the induced entropy and exploiting its equivalences to other entropies, a lot of the entropy search space can be covered with few experiments. Applying this method proved to be especially useful for the Hough forest experiments, where the search space contains discrete as well as differential entropies.

Since we noticed that other forest parameters can largely be optimized independently of the entropy type, we suggest to use and optimize the induced entropy as post-processing step after an optimization of forest parameters with the classical Shannon entropy. As we showed in our experiments, significant improvements can be reached even for a low number of samples of $q \in [1; 5]$.

References

[1] S. Abe. Axioms and uniqueness theorem for Tsallis entropy. *Physics Letters A*, 271(1–2):74 – 79, 2000. 1

[2] S. Abe. Stability of Tsallis entropy and instabilities of Rényi and normalized Tsallis entropies: A basis for q-exponential distributions. *Phys. Rev. E*, 66:046134, Oct 2002. 3

[3] K. Bache and M. Lichman. UCI machine learning repository, 2013. 5, 6

[4] R. Badii and A. Politi. *Complexity: Hierarchical Structures and Scaling in Physics*. Number 6 in Cambridge Nonlinear Science Series. Cambridge University Press, 1999. 1

[5] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 109–122, 2002. 6, 7

[6] A. P. Bremner. *Localised splitting criteria for classification and regression trees*. PhD thesis, Murdoch University, 2004. 2

[7] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006. 5

[8] A. Criminisi and J. Shotton, editors. *Decision Forests for Computer Vision and Medical Image Analysis*. Springer-Verlag London, 2013. 1, 2, 4

[9] M. Dantone, J. Gall, C. Leistner, , and L. V. Gool. Human Pose Estimation using Body Parts Dependent Joint Regressors. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 6, 7

[10] M. Dantone, J. Gall, C. Leistner, , and L. V. Gool. Body Parts Dependent Joint Regressors for Human Pose Estimation in Still Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, to appear. 7

[11] T. E. de Campos, B. R. Babu, and M. Varma. Character recognition in natural images. In *Proc. of the International Conference on Computer Vision Theory and Applications (VISAPP)*, February 2009. 5

[12] J. Gall, A. Yao, N. Razavi, L. V. Gool, and V. Lempitsky. Hough Forests for Object Detection, Tracking and Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(11):2188 – 2202, 2011. 1, 6

[13] J. Hull. A database for handwritten text recognition research. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(5):550–554, May 1994. 5

[14] S. Johnson and M. Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *Proc. of the British Machine Vision Conference*, 2010. 6

[15] A. Y. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, 1957. 1

[16] B. Lesche. Instabilities of Rényi entropies. *Journal of Statistical Physics*, 27(2):419–422, 1982. 3

[17] O. Maimon and L. Rokach, editors. *Data Mining and Knowledge Discovery Handbook*. Springer, 2nd edition, 2010. 2

[18] M. Masi. A step beyond Tsallis and Rényi entropies. *Physics Letters A*, 338(3–5):217–224, 2005. 2, 3

[19] T. Maszczyk and W. Duch. Comparison of Shannon, Renyi and Tsallis Entropy Used in Decision Trees. In L. Rutkowski, R. Tadeusiewicz, L. Zadeh, and J. Zurada, editors, *Artificial Intelligence and Soft Computing ICAISC 2008*, volume 5097 of *Lecture Notes in Computer Science*, pages 643–651. Springer Berlin Heidelberg, 2008. 2

[20] A. Rényi. On measures of information and entropy. In *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1961. 2

[21] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1963. 1

[22] B. D. Sharma and D. P. Mittal. New nonadditive measures of entropy for discrete probability distributions. *Journal of Mathematical Sciences (India)*, 10:28–40, 1975. 2

[23] Y.-S. Shih. Families of splitting criteria for classification trees. *Journal on Statistics and Computing*, 9:309–315, 1999. 2

[24] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1-2):479–487, 1988. 1, 2

[25] T. Yamano. Does the Lesche condition for stability validate generalized entropies? *Physics Letters A*, 329(4–5):268 – 276, 2004. 3