Information Theoretic Learning with Maximizing Tsallis Entropy

Nobuhide Aruga and Masaru Tanaka
Department of Information and Computer Sciences, Faculty of Engineering
Saitama University,
Shimo-Ookubo 255, Saitama-shi, Saitama,338-8570 Japan
Tel. +81-48-858-3957, Fax.: +81-48-858-3957
e-mail : aruga@mi.ics.saitama-u.ac.jp

Abstract:
We present the information theoretic learning based on the Tsallis entropy maximization principle for various q. The Tsallis entropy is one of the generalized entropies and is a canonical entropy in the sense of physics. Further, we consider the dependency of the learning on the parameter σ, which is a standard deviation of an assumed a priori distribution of samples such as Parzen window.

1. Introduction

Learning from examples is intrinsically related with the extraction of information from examples, but in general some of the method such as back propagation with mean squared error (MSE) are used. But in recent years, some of the novel way which use the entropy to estimate the distributions of the input data were sugested, and were given notice. Jose C. Principe and Dongxin Xu proposed information theoretic learning that maximizes entropy with Parzen Window method using a Gaussian kernel as the estimation of probability density function (PDF)[1]. The way they proposed is the unsupervised, directly, sample by sample learning. In their paper, but, only Renyi entropy of information order q fixed 2 is used, so other entropy and information order are not used. Therefore in this paper, we use Tsallis entropy which is one of the generalized entropy, and perform information theoretic learning with Tsallis entropy, with various q. We also consider the dependency of the learning on the parameter σ, which is a standard deviation of an assumed a priori distribution of samples such as Parzen window.

2. Generalized Entropy

Renyi entropy which is used in the Principe's paper, is one of the Genaralized entropy defined as

$$S_q^R = \frac{1}{1-q}\log\int p(x)^q dx \quad ,$$

where q is variable called *information order*, $p(x)$ is PDF of random variable x. The Tsallis entropy includes the Boltzmann - Shannon entropy as the special case $q = 1$,

$$S_1^R = -\int p(x)\log p(x)dx$$

The Tsallis entropy is defined as

$$S_q^T = \frac{1}{q-1}\left(1-\int p(x)^q dx\right) \quad ,$$

where q is called the information order, $p(x)$ is PDF of random variable x. The Renyi entropy and Tsallis entropy is connected by following equation,

$$S_q^R = \frac{1}{1-q}\log\left(1+(1-q)S_q^T\right)$$

3. Information Theoretic Learning with Maximizing Tsallis Entropy

Linsker proposed maximum entropy as a self-organizing principle for neural systems[2]. He showed if the covariance matrix is held constant, the continuous entropy measure is maximized for the normal distibution. If we assume that each element of the random vector is statistically independent from the other elements, we can use this propaty for the learning.
A nonparametric kernel-based method for estimationg the PDF well known is the Parzen window method[4]. Parzen window estimate of the probability distribution, $f_y(a)$, of a random vector $Y \in R^N$ at a point a is defined as

$$f_y(a) = \frac{1}{N}\sum_{i=1}^{N} K(y_i - a) \quad ,$$

where $K(.)$ is a kernel function which itself satisgies the properties of PDFs. Because of the local estimation of the PDF, the kernel function should also be localized, such as the Gaussian function[3]. So, we use the Gaussian function as the kernel of it. Parzen window using a Gaussian kernel is given as

$$f(y,a) = \frac{1}{N}\sum_{i=1}^{N} G(y-a_i,\sigma^2 I) \quad ,$$

where $G(y,\sigma^2 I)$ is Gaussian function, σ^2 is the variance, and $I \in R^{M\times M}$ is *identity matrix*. As mentioned above, we assume that covariance matrix is held constant. Then, if we assume $p(x)$ of Tsallis entropy to $f(y,a)$, The Tsallis entropy becomes:

$$S_q^T = \frac{1}{q-1}\left(1-\int p(x)^q dx\right) = \frac{1}{q-1}\left(1-V(a_i)\right) \quad ,$$

$$V(a_i) = \int\left\{\frac{1}{N}\sum_{i=1}^{N} G(y-a_i,\sigma^2 I)\right\}^q dy$$

When $q > 1$, $V(a_i)$ becomes smaller, entropy becomes larger, and when $q < 1$, $V(a_i)$ becomes larger, entropy becomes

larger. Thus using $V(a_i)$ in the learning of Multilayer perceptrons (MLP), we can adapt it in the back propagation method to minimamize $V(a_i)$, in an unsupervised mode.

4. Experimental Results

We show two experimental results. The first experiment is learning results for several distributions of an input data set, in order to see the dependency of the learning on the information order q for a fixed σ. Next, we consider the dependency of the learning on the standard deviation σ of the assumed input data distribution for a fixed q. The learning problem is to classify given samples into two categories.

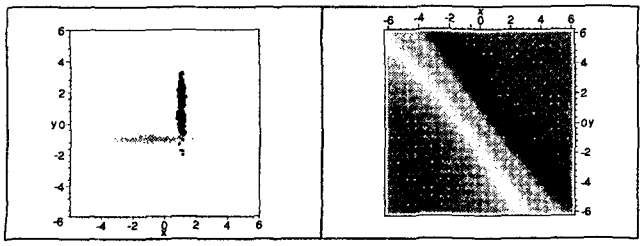
4.1 Dependency on the information order q

In first experiment, we select 2 sets of 250 samples belong to differ distribution each other, intentionally. We mix these samples, use these mixed samples as input of MLP(2-4-1) with information theoretic learning with maximizing Tsallis entropy. When a quantity of entropy increasing becomes under the regulation, learning is over. But this learning is under the control of initializing the learning weight and threshold, so we present the mean result of the 10 times results.

Figure 1 shows the output of MLP for several distributions. All of them, we set $\sigma = 0.1$. Upper 3 graphs (fig.(b)-(d)), we set distributions of samples crossed at a tip (input samples : fig.(a)). Middle 3 graphs (fig.(f)-(h)), we set the distributions of samples crossed at the center (input samples : fig.(e)). We consider these results satisfy the distributions of samples. Notice, difference of q makes small effects to the output result, e.g. curvature of a border line, but it don't make the effect for the result of learning.

Next, we set distributions of samples no-crossed (input samples : fig.(i)), but the result of it doesn't satisfy the distributions of samples. Then, we change σ from 0.1 to 0.35 and do the same process.The result becomes bottom 3 graphs (fig.(b)-(d)), which satisfy the distributions of samples.

Figure 2 shows the iteration times for each distributions. It shows that, though values of each graphs is different, but shape of the line has the similarity. By this graphs, we can also regard that the difference of q makes small effects to the output result.



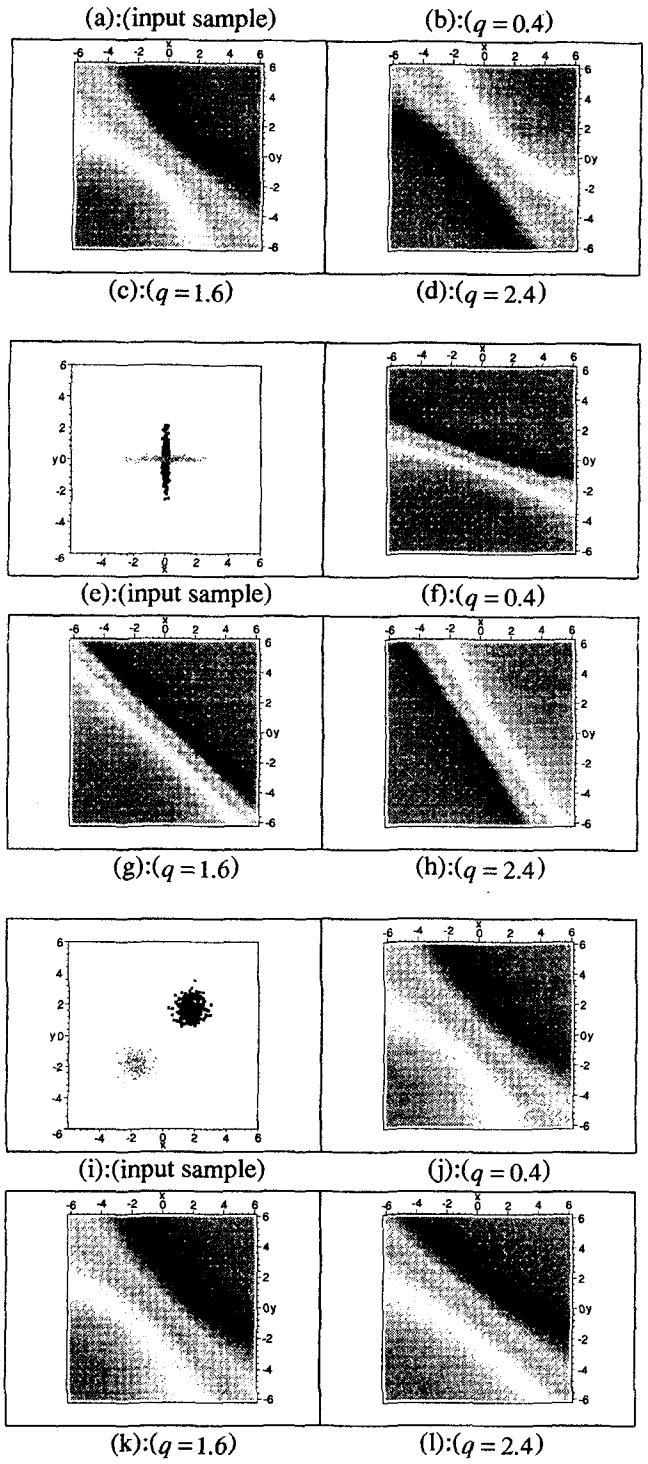


Figure 1:output of MLP for several distribution of input data. Sample datas are ploted on the left-top (symbol of cross and symbol of box). Difference of brightness (from lightness to darkness or its contrary) mean the border of the cluster. In this case, since we suppose two clusters, darkness region is certaion region and others are uncertain region.

4.2 Dependency on the standard deviation σ

In the second experiment, based on the first experiment, we make experiment about σ.

For easily, q is fixed 2, the distributions of samples are set artificially no-cross (like Figure 1-(i)). We select two sets of 100 samples which are given the same process like as first experiment. We set the mean vector and covariance matrix of the distributions of samples as

$$m_1 = \begin{bmatrix} d \\ d \end{bmatrix}, \quad m_2 = \begin{bmatrix} -d \\ -d \end{bmatrix}$$

$$(d = 1.0, 1.5, 2.0, 2.5, 3.0),$$

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$

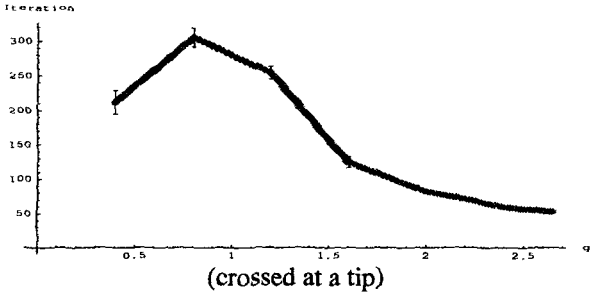
And we do the same process as first experiment.

The light-gray line at the left panels of Figure 3 shows the error between a desired response and the output for σ culculated with MSE. Though this learning is in unsupervised mode, so we can't know a desired response, but to make sample set separated, this error becomes

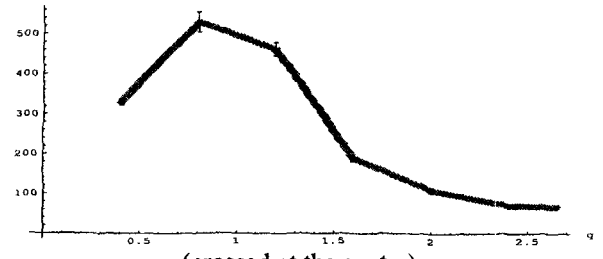
$$\frac{1}{N^2}\sum_{i=1}^{N}(1-|x_i|)^2, \quad -1 \le x_i \le 1.$$

Dark-gray line shows the reliability of learning for σ. Reliability of learning is calculated by fault trials within 100 trials, so if the value of reliability is low, it means *reliabilitiy is high. Fault means that learning* is over by entropy increasing becomes under the regulation before sufficient iterations. Right panels of figure 3 shows the iteration time for σ. In Figure 3, the curves of MSE of and iteration time are the mean by *successful trials of 100 trials.* These *figures show, whatever is* d, when σ is too small, result of learning is fault, and when σ is too large, result of learning becomes better but iterations becomes larger in vain and reliability of learning becomes lower. This mean that the determinant of *suitable* σ *may be related with the distribution of the sample* mostly. We can see that round $\sigma = 0.4$ is the most suitable for these distributions. We notice, when we set $\sigma = 0.1$ like first experiment, output result doesn't become a good thing. It tells us how import to set σ for suitable value.

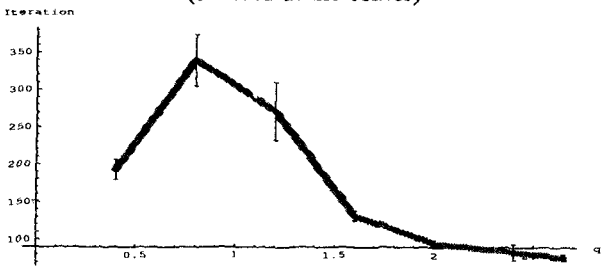
As features about d, when $\sigma = 0.2$, reliability shows the singular wave, and when $\sigma = 0.1$, iteration times shows the singular wave, following the changing of d. This is one of the problems to make clear.



(crossed at a tip)

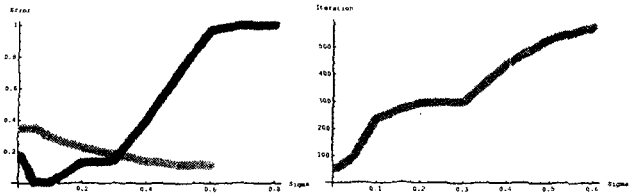


(crossed at the center)

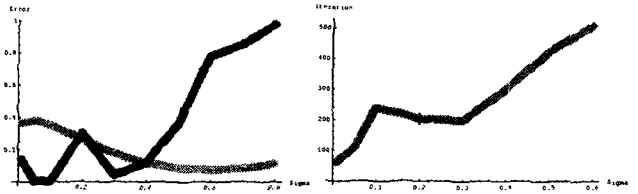


(no-crossed)

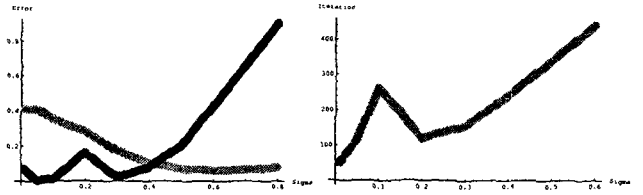
Figure 2:Iteration times for q.



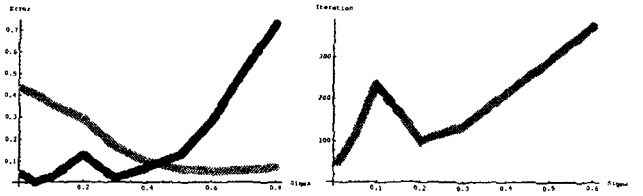
$(d = 1.0)$



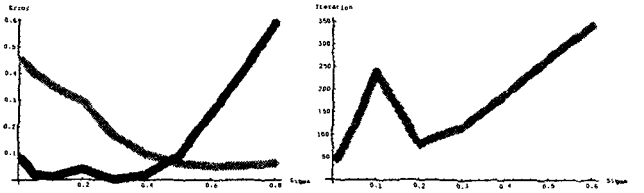
$(d = 1.5)$



$(d = 2.0)$



$(d = 2.5)$



$(d = 3.0)$

Figure 3:The results of the second experiment. Each left figure shows the error between a desired response and the

output (light-gray line), and reliability (dark-gray line) for σ. Right figures show the iteration times we needed to learn.

5. Conclusion

We considered about information theoretic learning with maximizing Tsallis entropy for several information order q, and showed some experiments. The Tsallis entropy is one of the generalized entropy and is a canonical entropy in the sense of physics. From our experiments, the q dependency of output result is not so large, but σ dependence of output result is quite severe. The determination of a suitable σ is related with the distribution of the sample mostly.

References

[1] Jose C.Principe, Dongxin Xu "Information-Theoretic Learning Using Renyi's Quadratic Entropy" ICA 1999

[2] R.Linsker, "Self-organization in a perceptual system.", Computer, vol.21, pp.105-117, 1988

[3] John W.Fisher, Jose C.Principe, "Entropy Manipulation of Arbitrary Nonlinear Mappings", IEEE Workshop NeuralNets for Signal Proc. 1997

[4] E.Parzen, "On the estimation of a probability density function and the mode.", Ann. Math. Stat., 33, pp.1065-1076, 1962