Learning from Examples with Quadratic Mutual Information

Dongxin Xu, Jose C. Principe

Computational NeuroEngineering Laboratory,
University of Florida, Gainesville, FL 32611
{xu,principe}@cnel.ufl.edu

Abstract

This paper discusses a novel algorithm to train nonlinear mappers with information theoretic criteria (entropy or mutual information) directly from a training set. The method is based on a Parzen window estimator and uses Renyi's quadratic definition of entropy and a distance measure based on the Cauchy-Schwartz inequality. We apply the algorithm to the difficult problem of vehicle pose estimation in synthetic aperture radar (SAR) with very good results.

1.0 Introduction

Information theory has served a crucial role in communication theory, but its application to pattern recognition and information processing has been sporadic. At the core lies the difficulty that pattern recognition is a discipline based on the learning by example metaphor, while information theory principles require an analytic form for the probability density function (pdf). One possibility is to postulate the form of the pdfs (normally Gaussian) and estimate from the data their parameters (mean and variance for Gaussian). This has been exactly the way Linsker [1] applied his principle of maximum information preservation (InfoMax). The analytic tractability has also restricted most of the work to linear models [1], [2] [3].

Recently, we have shown that these restrictions are no longer necessary [4], [5]. We developed a nonparametric estimator of entropy for a set of data (based on the Parzen window pdf estimator with appropriate entropy measures) and formulated entropy manipulation as seeking extrema of a cost function. Hence any mapper (linear or nonlinear) can be trained with our scheme. We have shown that the method although computationally demanding ($O(N^2)$, N is the number of data points in the training set) is robust and leads to the extraction of more information from the input data than the mean square error criterion (which only captures second order information from the data and can be regarded as a specific case of our scheme). We have applied the technique to blind source separation [6] and pose estimation [5] with very good results.

This paper clarifies and extends the algorithm for entropy estimation to the important case of mutual information. The mutual information of two random vectors is a much more useful principle for information processing as InfoMax clearly

shows. The importance of mutual information for pattern recognition should also be stressed. Fano's inequality clearly shows that maximizing mutual information decreases the lower bound of the probability of classification error. In fact, assume that the goal is to estimate a variable x with a discrete pdf $p(x)$ by calculating an estimate $\hat{x}$ from another random variable y characterized by $p(x|y)$. Fano states that

$$P(x \neq \hat{x}) \geq \frac{H_S(x|y) - 1}{\log(\Theta(x))} \quad (1)$$

where $\Theta(x)$ means the number of possible instances of x . This equation states that the probability of error is lower bounded by the conditional entropy of x given y . Substituting the definition of mutual information $I(x, y)$ we obtain

$$P(x \neq \hat{x}) \geq \frac{H_S(x) - I(x, y) - 1}{\log(\Theta(x))} \quad (2)$$

Notice that we have no control over the entropy of x nor the number of possible instances of x . *Therefore to improve the lower bound on the probability of error, we should maximize the mutual information between x and y .* This expression provides a crisp statement to support the use of a maximization of mutual information principle to improve the classification accuracy by directly minimizing the lower bound of the probability of error.

We will start by briefly reviewing alternative definitions of entropy that we use in our formulation and then present an algorithm that can train arbitrary mappers to maximize (or minimize) mutual information between the input and output of the mapper. We will conclude the paper by presenting an application to pose (aspect angle of a vehicle) estimation of vehicles in synthetic aperture radar (SAR) imagery.

2.0 Quadratic entropy and nonparametric estimators

If we define the amount of information associated with the measurement of an event x which occurs with probability p as $I(p) = \log \frac{1}{p}$, which is Hartley's measure [8], then Shannon's entropy H_S is the expectation of this measure:

$$H_S(x) = \sum_{k=1}^n p_k I(p_k) \quad \sum_{k=1}^n p_k = 1 \quad p_k > 0 \quad (3)$$

Renyi's entropy with order α , which we will denote by $H_{R\alpha}$ is defined as

$$H_{R\alpha} = \frac{1}{1-\alpha} \log \left(\sum_{k=1}^n p_k^\alpha \right) \quad \alpha > 0, \alpha \neq 1 \quad (4)$$

In this paper, $H_{R_2} = -\log \sum_{k=1}^n p_k^2$ is called Quadratic Entropy not only for convenience but also because it totally relies on the quadratic form of the probabilities.

For the continuous random variable Y with pdf $f_Y(y)$, we can obtain the differential version for these two types of entropy following a similar route to the Shannon differential entropy [8]:

$$\begin{cases} H_{R_\alpha}(Y) = \frac{1}{1-\alpha} \log \left(\int_{-\infty}^{+\infty} f_Y(y)^\alpha dy \right) \\ H_{R_2}(Y) = -\log \left(\int_{-\infty}^{+\infty} f_Y(y)^2 dy \right) \end{cases} \quad (5)$$

We will be working with Quadratic entropy because we have found a way of linking these quantities with potential energy when Parzen window estimator for pdfs is used. Let $y_i \in R^k$, $i = 1, \dots, N$, be a set of samples from a random variable

$Y \in R^k$ in k -dimensional space which can be the output of nonlinear mapper such as a multilayer perceptron (MLP). One interesting question is what will be the entropy associated with this set of data points. One answer lies in the estimation of the data pdf by the Parzen window method using a Gaussian kernel:

$$f_Y(y) = \frac{1}{N} \sum_{i=1}^N G(y - a_i, \sigma^2) \quad (6)$$

where $G(y, \Sigma) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} y^T \Sigma^{-1} y\right)$ is the Gaussian kernel in k

dimensional space, where Σ is the covariance matrix ($\Sigma = \sigma^2 I$). When Shannon's entropy is used along with this pdf estimation, the measure becomes very complex. Fortunately, Renyi's entropy with order 2 leads to a simpler form and we obtain the entropy measure for a set of discrete data points $\{y_i\}$:

$$H(\{y_i\}) = -\log \left(\int_{-\infty}^{+\infty} f_Y(y)^2 dy \right) = -\log V(\{y_i\}) \quad (7)$$

$$V(\{y_i\}) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N G(y_i - y_j, 2\sigma^2)$$

Making the analogy between data points and "physical particles", $V(\{y_i\})$ can be regarded as an overall potential energy of the data set since $G(y_i - y_j, 2\sigma^2)$ can

be taken as the potential energy of data point y_i in the potential field of data point y_j , or vice versa. *We will call this potential energy an information potential, where the data points have a correspondence to physical particles and the information potential to a potential field.* So, maximizing entropy in this case is equivalent to minimizing information potential.

Just like in mechanics, the derivative of the potential energy is a force, in this case an *information force*. The information force moves the data points in the output space to achieve an equilibrium state dictated by our criterion. Therefore,

$$\frac{\partial}{\partial a_i} G(y_i - y_j, 2\sigma^2) = G(y_i - y_j, 2\sigma^2)(y_j - y_i)/(2\sigma^2) \quad (8)$$

can be regarded as the force that the data point y_j impinges upon y_i . If all the data points are free to move in a certain region of the space, then the information forces between each pair of data points $\partial G(y_i - y_j, 2\sigma^2)/\partial y_i$ or

$\partial G(y_i - y_j, 2\sigma^2)/\partial y_j$ will drive all the data points to a state with minimum information potential.

Suppose the data points are the outputs of our parametric mapper MLP: $\{y(n) \in R^k, n=1, \dots, N\}$. If we want to adapt the MLP such that the mapping maximizes the entropy at the output $H(\{y(n)\})$, the problem is equivalent to finding the parameters of the mapper so that the information potential $V(\{y(n)\})$ is minimized. In this case, the data samples are not free but are a function of the MLP parameters. So, the information forces applied to each data sample can be back-propagated to the parameters using the chain rule [9]. As an example, the following gradient can be interpreted as force back-propagation:

$$\frac{\partial}{\partial w_{ij}} V(\{y(n)\}) = \sum_{n=1}^N \sum_{p=1}^k \frac{\partial}{\partial y_p(n)} V(\{y(n)\}) \frac{\partial}{\partial w_{ij}} y_p(n) \quad (9)$$

where $y(n) = (y_1(n), \dots, y_k(n))^T$, and w_{ij} are the MLP weights. The quantity

$$\frac{\partial}{\partial y(n)} V(\{y(n)\}) = \left(\frac{\partial}{\partial y_1(n)} V(\{y(n)\}), \dots, \frac{\partial}{\partial y_k(n)} V(\{y(n)\}) \right)^T \quad (10)$$

is the information force that the data sample $y(n)$ is subject to. Notice that the sensitivity of the output with respect to a MLP parameter $\frac{\partial}{\partial w_{ij}} y_p(n)$ is the transmission mechanism through which information forces are back-propagated to the parameter. From the analogy with the backpropagation formalism we conclude that information forces take the place of the injected error. *So, we obtain a gen-*

eral, nonparametric, and sample-based methodology to adapt arbitrary nonlinear (smooth) mappings for entropy manipulation.

3.0 Quadratic Mutual Information and Cross-Information Potential

For two random variables Y_1 and Y_2 (with marginal pdfs $f_{Y_1}(y_1)$, $f_{Y_2}(y_2)$ and joint pdf $f_{Y_1 Y_2}(y_1, y_2)$), the “mutual information” can be estimated using the Kulback-Leibler divergence between the joint probability and the factored marginals [8]. Inspired by this scheme and constrained by using quadratic forms of pdfs, we propose the following distance based on the Cauchy-Schwartz inequality:

$$C(Y_1, Y_2) = \log \frac{(\iint f_{Y_1 Y_2}(y_1, y_2)^2 dy_1 dy_2)(\iint f_{Y_1}(y_1)^2 f_{Y_2}(y_2)^2 dy_1 dy_2)}{(\iint f_{Y_1 Y_2}(y_1, y_2) f_{Y_1}(y_1) f_{Y_2}(y_2) dy_1 dy_2)^2} \quad (11)$$

It is obvious that $C(Y_1, Y_2) \geq 0$ and the equality holds true if and only if Y_1 and Y_2 are statistically independent, i.e. $f_{Y_1 Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2)$. So, $C(Y_1, Y_2)$ is an appropriate measure for the independence of two variables (minimization of mutual information). We also have experimental evidence that $C(Y_1, Y_2)$ is an appropriate measure for dependence of two variables (maximization of mutual information). Although we are unable to provide yet a strict justification that $C(Y_1, Y_2)$ is appropriate to measure dependence, we still call it “Quadratic Mutual Information” or “Quadratic Cross-Entropy” because of both convenience and the fact that it only relies on the quadratic forms of pdfs.

Now, suppose that we observe a set of data samples $\{y_{i1}, i=1, \dots, N\}$ for the variable Y_1 , $\{y_{i2}, i=1, \dots, N\}$ for the variable Y_2 . Let $y_i = (y_{i1}, y_{i2})^T$. Then $\{y_i, i=1, \dots, N\}$ are data samples for the joint variable $(Y_1, Y_2)^T$. Based on the Parzen window method, the joint pdf and marginal pdf can be estimated as:

$$\begin{cases} f_{Y_1 Y_2}(y_1, y_2) = \frac{1}{N} \sum_{i=1}^N G(y_1 - y_{i1}, \sigma^2) G(y_2 - y_{i2}, \sigma^2) \\ f_{Y_1}(y_1) = \frac{1}{N} \sum_{i=1}^N G(y_1 - y_{i1}, \sigma^2) \\ f_{Y_2}(y_2) = \frac{1}{N} \sum_{i=1}^N G(y_2 - y_{i2}, \sigma^2) \end{cases} \quad (12)$$

Combining (11), (12) and using (6), we obtain the following expressions for the Quadratic Mutual Information based on a set of data samples:

$$\left\{ \begin{array}{l} C((Y_1, Y_2) | \{y_i\}) = \log \frac{V(\{y_i\}) V_1(\{y_{i1}\}) V_2(\{y_{i2}\})}{V_{nc}(\{y_i\})^2} \\ V(\{y_i\}) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N G(y_i - y_j, 2\sigma^2) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left(\prod_{l=1}^2 G(y_{il} - y_{jl}, 2\sigma^2) \right) \\ V_l(y_j, \{y_{il}\}) = \frac{1}{N} \sum_{i=1}^N G(y_{il} - y_{jl}, 2\sigma^2), \quad l = 1, 2 \\ V_l(\{y_{il}\}) = \frac{1}{N} \sum_{j=1}^N V_l(y_j, \{y_{il}\}), \quad l = 1, 2 \\ V_{nc}(\{y_i\}) = \frac{1}{N} \sum_{j=1}^N \left(\prod_{l=1}^2 V_l(y_j, \{y_{il}\}) \right) \end{array} \right. \quad (13)$$

These expressions can be interpreted in terms of information potentials and extended for the case of multiple variables [5], but we do not have space to elaborate on the interpretation.

The cross-information potential is more complex than the information potential. Three different potentials (joint potential $V(\cdot)$, marginal potentials $V_l(\cdot)$, unnormalized cross-potential $V_{nc}(\cdot)$) contribute to the cross-information potential. So, the force applied to each data sample y_p comes from three independent sources (the marginal components). The q marginal force (marginal space indexed by q) that the data point y_p receives can be calculated as:

$$\begin{aligned} \frac{\partial}{\partial y_{pq}} V(\{y_i\}) &= \frac{1}{N^2} \sum_{i=1}^N \left(\prod_{l=1}^k G(y_{il} - y_{pl}, 2\sigma^2) \right) \frac{y_{iq} - y_{pq}}{\sigma^2} \\ \frac{\partial}{\partial y_{pq}} V_q(\{y_{iq}\}) &= \frac{1}{N^2} \sum_{i=1}^N G(y_{iq} - y_{pq}, 2\sigma^2) \frac{y_{iq} - y_{pq}}{\sigma^2} \\ \frac{\partial}{\partial y_{pq}} V_{nc}(\{y_i\}) &= \frac{1}{N^2} \sum_{j=1}^N \frac{1}{2} (B_j) G(y_{jq} - y_{pq}, 2\sigma^2) \frac{y_{jq} - y_{pq}}{\sigma^2} \end{aligned} \quad (14)$$

where $B_j = \prod_{l \neq q} V_l(y_j, \{y_{il}\}) + \prod_{l \neq q} V_l(y_p, \{y_{il}\})$. The overall marginal force that the data point y_p receives is:

$$\frac{\partial}{\partial y_{pq}} C((Y_1, Y_2)|\{y_i\}) = \frac{1}{V(\{y_i\})} \frac{\partial}{\partial y_{pq}} V(\{y_i\}) + \frac{1}{V_q(\{y_{iq}\})} \frac{\partial}{\partial y_{pq}} V_q(\{y_{iq}\}) - 2 \frac{1}{V_{nc}(\{y_i\})} \frac{\partial}{\partial y_{pq}} V_{nc}(\{y_i\})$$

Notice that the force from different sources are normalized by their corresponding information potentials to balance them out. Once the forces that each data point receives are calculated, these forces function as the injected error and can again be back-propagated to all the parameters so that the adaptation take place leading to the extremum of mutual information (cross-information-potential).

4.0 Pose Estimation with Mutual Information

Our problem is to find the aspect angle (also called the pose) of vehicles in synthetic aperture radar (SAR) imagery for a given depression angle. The poor resolution of SAR combined with the image formation properties makes the determination of pose a nontrivial problem. Knowing the pose of a vehicle can simplify tremendously the task of vehicle recognition.

Generally, the pose estimation can be formulated in terms of maximum a posteriori probability (MAP): $\hat{a} = \underset{a}{\operatorname{argmax}} f_{A|X}(a|x) = \underset{a}{\operatorname{argmax}} f_{AX}(a, x)$,

where $\hat{a}$ is the estimation of pose a , $f_{A|X}(a|x)$ is the a posteriori probability density function given the image x , $f_{AX}(a, x)$ is the joint pdf of pose and image.

So, the key issue here is to estimate the joint pdf. The very high dimensionality of the image (size: 80x80), however, makes it very difficult to obtain a reliable estimation. Dimensionality reduction (or feature extraction) becomes necessary. The output of the MLP mapping: $y = \text{MLP}(w, x)$ (w is the parameters of MLP 6400x3x2 or simply 6400x2) will serve as the feature space for pose estimation. So, instead of working directly on the image itself, our pose estimator becomes

$$\hat{a} = \underset{a}{\operatorname{argmax}} f_{A|Y}(a|y) = \underset{a}{\operatorname{argmax}} f_{AY}(a, y) \quad \text{The crucial point}$$

for this pose estimation scheme is how well y , which can be interpreted as a feature, conveys information about the pose. To obtain an effective feature, the mutual information between the feature and the pose $I(y, a)$ is used as the criterion to train the network so that the feature conveys the most information about pose: $w_{\text{optimal}} = \underset{w}{\operatorname{argmax}} I(y = \text{MLP}(w, x), a)$. To implement the idea, the

proposed cross-information-potential method is used to estimate $I(y, a)$ which boils down again to the estimation of $f_{AY}(a, y)$. as we discussed in section 3 The system diagram is shown in Figure 1. In this case, the joint pdf estimation used for the training of the network produces directly the pose estimation. In [5] a more involved pose estimator (discrete angles) was proposed but the results are identical to the ones obtained with this estimator.

The MSTAR public release data set [10] is used here to test our pose estimation algorithm. Each image chip contains one vehicle roughly centered. Data is very scarce, so we train the MLP with only 53 images of the vehicle BMP2_c21 (a personnel carrier), taken approximately 3.5 degrees apart (0 to 180 degrees). The algorithm converges in less than 100 batch iterations to a stable configuration where the outputs for the training set describe a circle in the output space (Figure 2). This solution makes perfect sense from an L2 norm perspective using our knowledge about the problem, but it was obtained directly by the algorithm with-

Figure 1. System Diagram

out imposing a desired response.

Table I shows the performance of the algorithm in other vehicle types (a tank and other personal carriers)

As we can see the estimation of the pose for the two BMP type of vehicles and the BTR (a different vehicle class) are very close to the resolution of the training set data (3.5°). So we conclude that our algorithm generalizes the information contained in the training images. The vehicle T72 is a totally different vehicle class (tanks versus personnel carriers) so we consider the accuracy of 6.65° better than our expectations, which shows that the performance has a smooth roll-off.

Testing results in several vehicles

Vehicle Class_type	Error mean (var) degrees
BMP2_c21	0.54 (0.40) -training set
BMP2_9563	4.25 (3.62)
BMP2_9566	3.81 (3.16)
BTR70_c71	3.18 (2.84)
T72_s7	6.65 (5.04)

Figure 2. System outputs: diamonds training, triangles testing data.

5.0 Conclusion

We present in this paper a novel estimator for mutual information which enables the training of linear or nonlinear mappers by the InfoMax Principle directly from examples. We utilize the concept of Renyi's Quadratic entropy along with the Parzen window pdf estimation to develop an easily implementable entropy estimator based on information potential. With this estimator of the entropy in the output space of a mapper we can adapt parameters using information force back-propagation. Using the Cauchy-Schwartz distance instead of Kulback-Leibler divergence we are able to extend the method to mutual information. We applied the novel algorithm to the estimation of vehicle pose in SAR and found out that the method trains fast and very reliably. The generalization to other vehicle types is much better than we expected in view of the reduced number of exemplars utilized in the training. We are investigating this highly appealing aspect of the method.

Acknowledgments: This work was partially supported by a DARPA-Air Force grant F33615-97-1019.

References

- [1] Linsker R. "An application of the principle of maximum information preservation to linear systems", in *Advances in Neural Information Processing Systems 1*, Morgan-Kaufman, pp 485-494, 1988.
- [2] Deco, G. *An Information-Theoretic Approach to Neural Computing*, New York, Springer, 1996
- [3] Plumbley M., Fallside F., "An information theoretic approach to unsupervised networks", *Int. J. Conf. on Neural Nets*, vol 2, p 598, Washington, DC, 1989.
- [4] Fisher J. W. III "Nonlinear Extensions to the Minimum Average Correlation Energy Filter" Ph.D dissertation, Dept. of ECE, University of Florida, 1997
- [5] Xu Dongxin. "Feature Extraction: From Decorrelation to Statistical Independence" Ph.D research proposal, Dept. of ECE, University of Florida, 1997
- [6] Xu D., Principe J., Fisher J. and Wu H-C. "A Novel Measure for Independent Component Analysis (ICA)" to be presented in ICASSP'98.
- [7] Renyi, A. "Some Fundamental Questions of Information Theory", *Selected Papers of Alfred Renyi*, Vol.2, 1976.
- [8] Kapur, J.N. "Measures of Information and Their Applications". John Wiley & Sons. 1994
- [9] Rumelhart, D.E., Hinton, G.E. and Williams, J.R. "Learning representations by back-propagating errors", *Nature (London)*, 323, pp533-536, 1986.
- [10] MSTAR (public) Targets, CDROM, Veda Inc. Ohio, 1997.