Learning Discriminative Feature Transforms to Low Dimensions in Low Dimensions

Kari Torkkola
Motorola Labs, 7700 South River Parkway, MD ML28, Tempe AZ 85284, USA
Kari.Torkkola@motorola.com  http://members.home.net/torkkola

Abstract
The marriage of Renyi entropy with Parzen density estimation has been shown to be a viable tool in learning discriminative feature transforms. However, it suffers from computational complexity proportional to the square of the number of samples in the training data. This sets a practical limit to using large databases. We suggest immediate divorce of the two methods and remarriage of Renyi entropy with a semi-parametric density estimation method, such as a Gaussian Mixture Models (GMM). This allows all of the computation to take place in the low dimensional target space, and it reduces computational complexity proportional to square of the number of components in the mixtures. Furthermore, a convenient extension to Hidden Markov Models as commonly used in speech recognition becomes possible.

1 Introduction
Feature selection or feature transforms are important aspects of any pattern recognition system. Optimal feature selection coupled with a particular classifier can be done by actually training and evaluating the classifier using all combinations of available features. Obviously this wrapper strategy does not allow learning feature transforms, because all possible transforms cannot be enumerated. Both feature selection and feature transforms can be learned by evaluating some criterion that reflects the “importance” of a feature or a number of features jointly. This is called the filter configuration in feature selection. An optimal criterion for this purpose would naturally reflect the Bayes error rate. Approximations can be used, for example, based on Bhattacharyya bound or on an interclass divergence criterion. These are usually accompanied by a parametric estimation, such as Gaussian, of the densities at hand [6, 12]. The classical Linear Discriminant Analysis (LDA) assumes all classes to be Gaussian with a shared single covariance matrix [5]. Heteroscedastic Discriminant Analysis (HDA) extends this by allowing each of the classes have their own covariances [9].

Maximizing a particular criterion, the joint mutual information (MI) between the features and the class labels [1, 17, 16, 13], can be shown to minimize the lower bound of the classification error [3, 10, 15]. However, MI according to the popular definition of Shannon can be computationally expensive. Evaluation of the joint MI of a number of variables is plausible through histograms, but only for a few variables [17]. As a remedy, Principe et al showed in [4, 11, 10] that using Renyi’s entropy instead of Shannon’s, combined with Parzen density estimation, leads to expressions of mutual information with computational complexity of $O(N^2)$, where $N$ is the number of samples in the training set. This method can be formulated to express the mutual information between continuous variables and discrete class labels in order to learn dimension-reducing feature transforms, both linear
and non-linear [14], for pattern recognition. One must note that regarding finding the extrema, both definitions of entropy are equivalent (see [7] pages 118, 406, and [8] page 325).

This formulation of MI evaluates the effect of each sample to every other sample in the transformed space through the Parzen density estimation kernel. This effect can also be called as the “information force”. Thus large/huge databases are hard to use due to the $O(N^2)$ complexity.

To remedy this problem, and also to alleviate the difficulties in Parzen density estimation in high-dimensional spaces ($d > 8$), we present a formulation combining the mutual information criterion based on Renyi entropy with a semi-parametric density estimation method using Gaussian Mixture Models (GMM). In essence, Parzen density estimation is replaced by GMMs. In order to evaluate the MI, evaluating mutual interactions between mixture components of the GMMs suffices, instead of having to evaluate interactions between all pairs of samples. An approach that maps an output space GMM back to input space and again to output space through the adaptive feature transform is taken. This allows all of the computation to take place in the target low dimensional space. Computational complexity is reduced proportional to the square of the number of components in the mixtures.

This paper is structured as follows. An introduction is given to the maximum mutual information (MMI) formulation for discriminative feature transforms using Renyi entropy and Parzen density estimation. We discuss different strategies to reduce its computational complexity, and we present a formulation based on GMMs. Empirical results are presented using a few well known databases, and we conclude by discussing a connection to Hidden Markov Models.

2 MMI for Discriminative Feature Transforms

Given a set of training data $\{x_i, c_i\}$ as samples of a continuous-valued random variable $X$, $x_i \in \mathbb{R}^D$, and class labels as samples of a discrete-valued random variable $C$, $c_i \in \{1, 2, \ldots, N_c\}$, $i \in [1, N]$, the objective is to find a transformation (or its parameters $w$) to $y_i \in \mathbb{R}^d$, $d < D$ such that $y_i = g(w, x_i)$ that maximizes $I(C, Y)$, the mutual information (MI) between transformed data $Y$ and class labels $C$. The procedure is depicted in Fig. 1. To this end, we need to express $I$ as a function of the data set, $I(\{y_i, c_i\})$, in a differentiable form. Once that is done, we can perform gradient ascent on $I$ as follows

$$w_{t+1} = w_t + \eta \frac{\partial I}{\partial w} = w_t + \eta \sum_{i=1}^{N} \frac{\partial I}{\partial y_i} \frac{\partial y_i}{\partial w}. \quad (1)$$

To derive an expression for MI using a non-parametric density estimation method we apply Renyi’s quadratic entropy instead of Shannon’s entropy as described in [10, 15] because of its computational advantages. Estimating the density $p(y)$ of $Y$ as a sum of spherical Gaussians each centered at a sample $y_i$, the expression of Renyi’s quadratic entropy of $Y$ is

$$H_R(Y) = -\log \int_y p(y)^2 dy = -\log \frac{1}{N^2} \int_y \left( \sum_{k=1}^{N} \sum_{j=1}^{N} G(y - y_k, \sigma^2 I) G(y - y_j, \sigma^2 I) \right) dy$$

$$= -\log \frac{1}{N^2} \sum_{k=1}^{N} \sum_{j=1}^{N} G(y_k - y_j, 2\sigma^2 I). \quad (2)$$