Pattern Recognition Letters xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Info-margin maximization for feature extraction ☆

Xipeng Qiu *, Lide Wu

School of Computer Science, Fudan University, Shanghai, China

ARTICLE INFO

Article history:
Received 13 September 2006
Received in revised form 5 December 2008
Available online xxxx

Communicated by W. Pedrycz

Keywords:
Linear feature extraction
Info-margin maximization

ABSTRACT

We propose a novel method of linear feature extraction with info-margin maximization (InfoMargin) from information theoretic viewpoint. It aims to achieve a low generalization error by maximizing the information divergence between the distributions of different classes while minimizing the entropy of the distribution in each single class. We estimate the density of data in each class with Gaussian kernel Parzen window and develop an efficient and fast convergent algorithm to calculate quadratic entropy and divergence measure. Experimental results show that our method outperforms the traditional feature extraction methods in the classification and data visualization tasks.

1. Introduction

In pattern recognition field, the raw input data often have very high dimensionality and the limited number of samples, for example, image or document data. We have to face the "curse of dimensionality" problem. Feature extraction (Fukunaga, 1990; Friedman, 1987) is a dimensionality reduction method to map high-dimensional data to a low-dimensional space for classification or visualization. Principal component analysis (PCA) and linear discriminant analysis (LDA) are two popular linear feature extraction methods among the state-of-art literatures (Fukunaga, 1990).

PCA (Fukunaga, 1990) aims to find a set of mutually orthogonal basis functions that capture the directions of maximum variance in the data, so it is very useful in reducing noise at the data.

LDA (Fukunaga, 1990) is used to derive a discriminative transformation which maximizes between-class scatter while minimizing within-class scatter. However, LDA only makes use of second-order statistical information, covariances, so it is optimal for data where each class has a uni-modal Gaussian density and does not share the same mean (Fukunaga, 1990). When the class conditional densities are multi-modal, LDA does not work well. LDA also fails when the class separability cannot be represented by the between-class scatter matrix S_b. Especially in the case that each class shares the same mean, it fails to find the discriminant direction because $S_b = 0$ (Fukunaga, 1990). Nonparametric discriminant analysis (Fukunaga and Mantock, 1983) is proposed to improve this drawback and calculates S_b in nonparametric nature, but it lacks of the global consideration for the distributions of data. Another drawback of LDA is the so-called "small sample size (SSS)" problem that within-class scatter matrix is singular when dealing with high-dimensional data. PCA + LDA (Belhumeur et al., 1997) and null space-based LDA (NLDA) (Chen et al., 2000) are two effective approaches to solve this problem. Though these methods improve the performance of the basic LDA, they are also solely based on the second order statistical information. Hence, they may not work well especially when the distributions are non-Gaussian in many practical cases.

There are some other works based on higher-order statistics for supervised feature extraction (Linsker, 1988; Bollacker and Ghosh, 1996; Principe et al., 2000; Vasconcelos, 2002; Lyu, 2005; Hild et al., 2006), which are based on maximizing the mutual information (MI) between extracted features and class label.

To clarify these works and lead our motivation, we give some brief definitions firstly. Assuming that a random variable y is drawn from the distribution $p(y)$, and a discrete-valued random variable c representing its class label from $p(c)$, the entropies of c and y, making use of Shannons definition (Cover and Thomas, 1991), are expressed in terms of the prior probabilities $p(y)$ and $p(c)$,

$$H(c) = -\sum_c P(c)\log(P(c)), \tag{1}$$

$$H(p) = -\int_y p(y)\log p(y)dy. \tag{2}$$

After having observed a feature vector y, the uncertainty of the class label c could be defined as the conditional entropy,

$$H(c|y) = -\sum_c \int_y p(c,y)\log p(c|y)dy. \tag{3}$$

The mutual information (MI) $I(y,c)$ between y and c can be written as

☆ This work was (partially) funded by Chinese NSF 60673038, Doctoral Fund of Ministry of Education of China 200802460066, and Shanghai Science and Technology Development Funds 08511500302.

* Corresponding author. Tel.: +86 21 55664405.
E-mail addresses: xpqiu@fudan.edu.cn (X. Qiu), ldwu@fudan.edu.cn (L. Wu).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.08.006

$$I(y,c) = \sum_c \int_y p(c,y) \log \frac{p(c,y)}{P(c)p(y)} dy \quad (4)$$

$$= H(c) - H(c|y) \quad (5)$$

$$= H(y) - H(y|c). \quad (6)$$

Mutual information maximization is a powerful feature extraction criterion, and it is optimal for training samples in the minimum Bayes error (Vasconcelos, 2002). However, it is still not widely used currently due to its computational difficulties, especially for the high-dimensional data. Although histogram-based MI estimation works with two or three variables, it fails in higher dimensions. Torkkola (2003) proposes an effective MI based feature extraction method with replacing Shannon entropy with Renyi entropy, which can improve computational efficiency greatly.

Although the MI based methods achieved some good performances in feature extraction, they also suffer from the following drawbacks: (1) they are based on density estimation in the transformed subspace, which is still computationally expensive and is not robust when the dimensionality of subspace is high; (2) from Eq. (5), we can see MI based methods fail to find the best features among the feature candidates with the same mutual information to class label. Eq. (5) shows that when the classes are separated in extracted feature space ($H(C|y) = 0$), MI based methods will not continue to find the better feature space (see Fig. 1). Thus they cannot guarantee a low generalization error for incoming unknown samples; (3) from Eq. (6), it are not explicit how $H(y|C)$ play role to maximize the mutual information. Since we expect that the samples in same class are close with each other, so $H(y|C)$ should be as low as possible. But MI based methods cannot guarantee this point.

Margin maximization (Vapnik, 1995) is theoretically interesting because it facilitates generalization error analysis, and practically interesting because it presents a clear geometric interpretation of the models being built. Margin is the divergency between classes with different measures indeed, such as Euclidian distance and hypothesis distance (Crammer et al., 2002). A loss function of "margin maximizing" in this sense is useful for generating good prediction models (Rosset et al., 2003). An excellent example for margin maximization is support vector machine (SVM) (Vapnik, 1995), which defines margin as the distance between an instance and the decision boundary.

Fig. 1 shows that the mutual information between the projected samples and class labels are the same in two projection directions $P1$ and $P2$. However the projection direction $P1$ has a larger margin than $P2$ from SVM viewpoint. Once mutual information maximization based methods, such as (Torkkola, 2003), find the projection $P2$, they cannot continue to search the optimal projection $P1$ because $P2$ has reached a local maximum of mutual information.

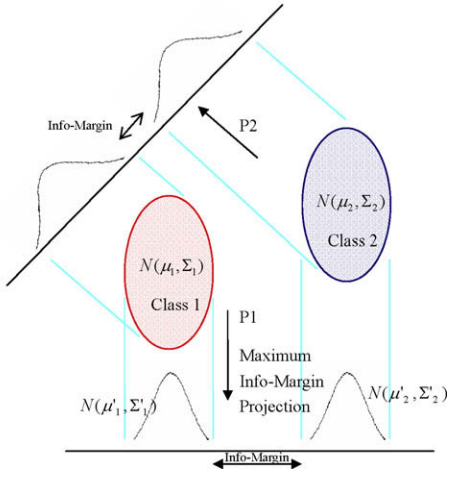


Fig. 1. Illustration of maximize info-margin projection.

In this paper, we define a novel margin (info-margin) from information theoretic view, and propose a linear feature extraction method (InfoMargin) by maximizing the defined info-margin. It aims to achieve a low generalization error by maximizing the information divergence between the distributions which belong to different classes while minimizing the entropy of distribution in each single class. We estimate the density in single class with Gaussian kernel Parzen window and give an efficient algorithm by quadratic entropy and divergence measure, which avoid to use histogram-based methods to estimate the class densities.

The rest of the paper is organized as following. Section 2 gives a brief review for information theory, especially the generalized definitions for entropy and divergence measure. Section 3 proposes the info-margin criterion and info-margin maximization algorithm. Then, we show the experimental results in Section 4. At last, we conclude our works in Section 5.

2. Entropy, divergence measure, and density estimation

In this section, we give a brief introduction to the definitions of entropy, divergence measures and density estimation.

2.1. Shannon's entropy and Kullback–Leibler divergence

Assuming a random variable y from a distribution $p(y)$, its entropy or uncertainty, making use of Shannon's definition (Cover and Thomas, 1991), is expressed as Eq. (2).

Kullback–Leibler divergence measure between two distributions $p(y)$ and $q(y)$ is written as

$$D_{KL}(p||q) = \int p(y) \log \frac{p(y)}{q(y)} dy. \quad (7)$$

To calculate the Shannon's entropy and KL-divergence, the probability densities of the variables are required. Moreover, the calculation involves integration of these probability densities, which leads to a high computational complexity. Histogram-based estimation is a popular method to solve this problem, but it fails for high-dimensional data due to the sparsity of data in spaces (Principe et al., 2000). Therefore, information theory is not widely used for feature extraction on high-dimensional data currently. By relaxing some conditions, other entropy and divergence measures can be used to satisfy requirements such as computational efficiency. Here, we extend them to quadratic forms.

2.2. Quadratic entropy and divergence measure

Principe et al. (2000) have shown that the computational complexity can be reduced by using Renyi's entropy, instead of Shannon's.

Bian and Zhang (1999) gives a definition for generalized entropy measure. We extend it to the generalized differential entropy:

$$H_\alpha(p(y)) = \frac{1}{2^{1-\alpha}-1}\left[\int p(y)^\alpha dy - 1\right], \quad (8)$$

where $\int p(y)dy = 1$, α is a positive parameter and $\alpha \neq 1$.

According to L'Hospital's rule (Weisstein, 2006),

Please cite this article in press as: Qiu, X., Wu, L. Info-margin maximization for feature extraction. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.08.006

$$H_1(p(y)) = \lim_{\alpha\to 1}\frac{1}{2^{1-\alpha}-1}\left[\int p(y)^{\alpha}dy - 1\right] = -\int p(y)\log p(y)dy. \quad (9)$$

So generalized entropy converges to Shannon's entropy when $\alpha \to 1$.

For further computational efficiency, we choose $\alpha = 2$ and get quadratic entropy,

$$H_2 = 2\left(1 - \int p(y)^2 dy\right). \quad (10)$$

The quadratic entropy has a relation to Shannon's entropy. Due to $p(y) - 1 \geqslant \log(p(y))$, we have $1 - p(y) \leqslant -\log(p(y))$, $p(y)*(1-p(y)) \leqslant -p(y)*\log(p(y))$, so $H_2 \leqslant 2H_{Shannon}$.

Like Shannon's differential entropy (Cover and Thomas, 1991), H_2 can be negative. However, it still is a good measure of uncertainty for information.

Similarly, KL divergence is also not the only divergence measure of densities. An alternative measure is the quadratic divergence (Kapur, 1994) defined as:

$$D_2(p||q) = \int_y (p(y) - q(y))^2 dy. \quad (11)$$

It is clear that the measure is always nonnegative, and when $p(y) = q(y)$ for all y, it equals to zero.

2.3. Gaussian kernel Parzen density estimation

Assume that the density of y is estimated as a sum of spherical Gaussians each centered at a sample y_n (Parzen density estimator),

$$p(y) = \frac{1}{N}\sum_{n=1}^{N} G(y - y_n, \sigma^2 I), \quad (12)$$

where I denotes a unit matrix and σ is a parameter to control the size of Parzen window. In this paper, we always set σ equal to half of average distance within classes.

The Gaussian kernel in d-dimensional space is defined as

$$G(y, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}y^T\Sigma^{-1}y\right). \quad (13)$$

The Gaussian kernel has an interesting property that the convolution of two Gaussians centered at a_m and a_n is a Gaussian centered at $a_m - a_n$ with covariance matrix equal to the sum of the original covariance matrices.

$$\int_y G(y - a_m, \Sigma_m)G(y - a_n, \Sigma_n)dy = G(a_m - a_n, \Sigma_m + \Sigma_n). \quad (14)$$

This property facilitates evaluating quadratic entropy measure, which is a function of the square of the density function.

Then the quadratic entropy equals

$$\begin{aligned} H_2(Y) &= 2\left(1 - \int_y p(y)^2 dy\right) \\ &= 2 - \frac{2}{N^2}\int_y\left(\sum_{m=1}^{N}\sum_{n=1}^{N} G(y - y_m, \sigma^2 I)G(y - y_n, \sigma^2 I)\right)dy \\ &= 2 - \frac{2}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N} G(y_m - y_n, 2\sigma^2 I). \end{aligned} \quad (15)$$

3. Info-margin maximization

To classify the samples, we need some divergence measure to calculate the divergency between classes. We would hope that a pattern is close to those in the same class but far from those in different classes. Therefore, a good feature extractor should maximize the divergence between classes while minimizing the distances between samples in same class after the transformation. In this section, we reformulate this idea from information theoretic view and give a info-margin maximization criterion to overcome many limitations of the traditional feature extraction methods.

3.1. Definition of info-margin

Instead of maximizing the mutual information between features and class label, we maximize the divergency between classes and minimize the entropy of each class.

For the classes ω_i and ω_j, we define their info-margin as:

$$IM(\omega_i, \omega_j) = D_2(p(y|\omega_i)||p(y|\omega_j)) - H_2(p(y|\omega_i)) - H_2(p(y|\omega_j)), \quad (16)$$

where $p(y|\omega_i)$ is the density of random variable y in given class ω_i.

If we estimate $p(y|\omega_i)$ using a Gaussian kernel Parzen window,

$$p(y|\omega_i) = \frac{1}{N_i}\sum_{k=1}^{N_i} G(y - y_{ik}, \sigma^2 I), \quad (17)$$

where y_{ik} is the k-th sample in class i.

Thus info-margin between ω_i and ω_j can be written as

$$\begin{aligned} IM(\omega_i, \omega_j) &= D_2(p(y|\omega_i), p(y|\omega_j)) - H_2(p(y|\omega_i)) - H_2(p(y|\omega_j)) \\ &= \int_y (p(y|\omega_i) - p(y|\omega_j))^2 dy - 2\left(1 - \int_y p(y|\omega_i)^2 dy\right) \\ &\quad - 2\left(1 - \int_y p(y|\omega_j)^2 dy\right) \\ &= 3\int_y p(y|\omega_i)^2 dy + 3\int_y p(y|\omega_j)^2 dy \\ &\quad - 2\int_y p(y|\omega_i)p(y|\omega_j)dy - 4 \\ &= \frac{3}{N_i^2}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i} G(y_{ik} - y_{il}, 2\sigma^2 I) + \frac{3}{N_j^2}\sum_{k=1}^{N_j}\sum_{l=1}^{N_j} G(y_{jk} - y_{jl}, 2\sigma^2 I) \\ &\quad - \frac{2}{N_i N_j}\sum_{k=1}^{N_i}\sum_{l=1}^{N_j} G(y_{ik} - y_{jl}, 2\sigma^2 I) - 4, \end{aligned} \quad (18)$$

where N_i is the number of samples in class ω_i.

Thus, the average info-margin of all classes can be written as:

$$\begin{aligned} IM(Y) &= \frac{1}{2}\sum_{i,j=1}^{C} p_{\omega_i}p_{\omega_j} IM(\omega_i, \omega_j) = \frac{1}{2}\sum_{i,j=1}^{C}\frac{N_i N_j}{N^2} IM(\omega_i, \omega_j) \\ &= \frac{3}{N}\sum_{i=1}^{C}\frac{1}{N_i}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i} G(y_{ik} - y_{il}, 2\sigma^2 I) \\ &\quad - \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N} G(y_m - y_n, 2\sigma^2 I) - 2, \end{aligned} \quad (19)$$

where C is the number of classes and p_{ω_i} is the probability of class ω_i.

Eq. (19) also gives an intuitive explanation of info-margin, which is equivalent to minimize the within-class distances while maximizing the between-class distances. The different is that distance measure of info-margin is Gaussian-kernel measure rather than Euclidean distance.

3.2. Info-margin maximization criterion

Given a set of training data (x_n, c_n), where $n \in \{1, \ldots, N\}$, $x_n \in R^D$ is random variable and $c_n \in \{\omega_1, \ldots, \omega_C\}$ is class label. We aim to find a transformation $y_n = g(W, x_n)$, $y_n \in R^d$, $d < D$, to maximize

Please cite this article in press as: Qiu, X., Wu, L. Info-margin maximization for feature extraction. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.08.006

average info-margin. Here, we set $y = W^T x$ with constraint $W^T W = I$, W is $D \times d$ matrix.

So, info-margin maximization criterion is:

$$W_{opt} = \arg\max_W IM(W^T x). \tag{20}$$

We maximize Eq. (20) by gradient ascent method.

$$W_{t+1} = W_t + \eta \frac{\partial IM}{\partial W} = W_t + \eta \sum_{n=1}^{N} \frac{\partial y_n}{\partial W} \left(\frac{\partial IM}{\partial y_n} \right)^T. \tag{21}$$

First, we calculate $\partial IM / \partial y_n$,

$$\frac{\partial}{\partial y_{ik}} IM(Y) = \frac{3}{N N_i \sigma^2} \sum_{l=1}^{N_i} G(y_{il} - y_{ik}, 2\sigma^2 I)(y_{il} - y_{ik}) - \frac{1}{N^2 \sigma^2} \sum_{m=1}^{N} G(y_m - y_{ik}, 2\sigma^2 I)(y_m - y_{ik}). \tag{22}$$

Second, $\partial y_n / \partial W = x_n$. We insert it and Eq. (22) in Eq. (21), which results in gradient ascent algorithm to maximize info-margin.

In high dimensional spaces, the Parzen estimator suffers also from the curse of dimensionality. For computation efficiency, we use incremental method to calculate W. In each step, we find a 1D projection direction that maximizes info-margin, then continue to find the other projection directions in the subspace orthogonal to the previously found directions. Algorithm 1 gives a flowchart of our algorithm.

Algorithm 1. Flowchart of maximize info-margin criterion.

- Given samples (x_n, c_n), $n \in \{1, \ldots, N\}$ and $c_n \in \{\omega_1, \ldots, \omega_C\}$ is label;
- For $i = 1, \ldots, d$,
  (1) Initialize W_i randomly and normalize it by $||W_i|| = 1$, where $W_i \in R^{D \times 1}$;
  (2) Update W_i using Eq. (21), where kernel width σ is set to half of average distance within classes;
  (3) Project samples to the orthogonal complement space of W_i by $X = X - W_i W_i^T X$;
  (4) $i = i + 1$.
- The final $W = [W_1, \ldots, W_d]$.

hline

Fig. 2 gives comparisons among LDA, maximize mutual information (MMI) (Torkkola, 2003) and info-margin maximization (InfoMargin) on two synthetic 2D datasets, which shows InfoMargin has better performances than MMI. Fig. 2 also illuminates the drawbacks of MMI that we discussed above. When MMI has found projection direction to separate the classes, it cannot seek a better projection direction to maximize the divergency between classes. In Fig. 2b, LDA fails because that the classes share the same mean.

We also compare the performances of LDA, MMI and InfoMargin for more complicated datasets. The visualization results are shown in Fig. 3. We can see that LDA fail to find the correct class structure. Although MMI gives better projection than LDA, it cannot guarantee a low generalization error since the margin is small between classes. InfoMargin gives larger margin than MMI.

Moreover, InfoMargin has a faster convergent speed than MMI. Table 1 gives the CPU times for two 3D datasets.

4. Experiments

In this section, we apply info-margin maximization (InfoMargin) to real world data, and compare it with the other popular linear feature extraction methods, such as PCA (Turk and Pentland, 1991), PCA + LDA (Belhumeur et al., 1997), nonparametric discriminant analysis (NDA) (Fukunaga, 1990), nonlinear component analysis (Schölkopf et al., 1998) and mutual information maximization (MMI) (Torkkola, 2003). Since that MMI is not robust when the dimensionality of transformed subspace is high, we use incremental method to calculate W for MMI, like that for InfoMargin.

4.1. Face image datasets

We firstly perform the experiments on two popular face datasets (AT&T database (Samaria and Harter, 1994) and UMIST database (Graham and Allinson, 1998)). All the following experiments are repeated five times independently and the average results are calculated. The classifier is nearest neighbor classifier. The images are formulated in the vector representation. The descriptions of the two datasets are below:

AT&T dataset: This dataset is the AT&T face database (formerly 'The ORL Database of Faces'), which contains 400 images (112×92) of 40 persons, 10 images per person. The set of the 10 images for each person is randomly partitioned into a training subset of five images and a test set of the other five.

UMIST dataset: This dataset is a multi-view database consisting of 575 gray-scale images (112×92) of 20 persons, each covering a wide range of poses from profile to frontal views as well as race, gender and appearance. Five images of each person are randomly chosen for training and the rest for test.

Figs. 4 and 5 show the recognition rates with the different dimension of subspace on the AT&T and UMIST face datasets. It is shown that info-margin maximization criterion (InfoMargin)

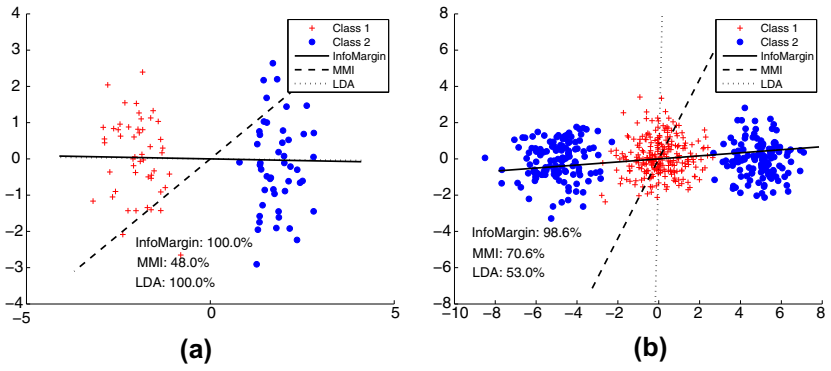
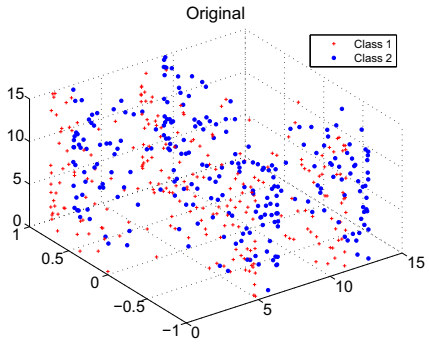
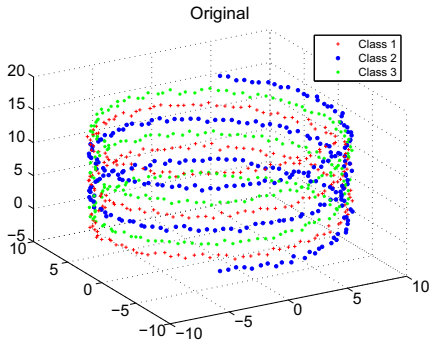
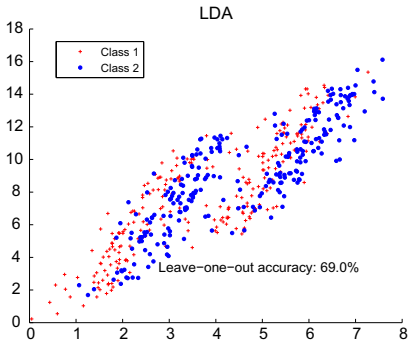


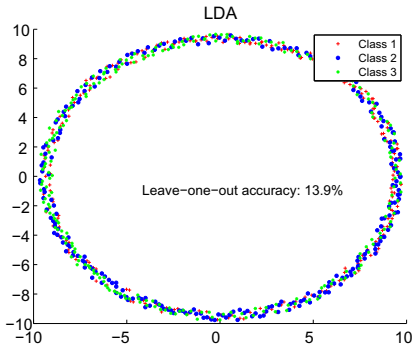
Fig. 2. First projection directions of LDA, MMI and InfoMargin on two synthetic 2D datasets. The leave-one-out accuracy rates are also given in 1D projected subspace by NN classification. The data of each class are generated by sampling from single or multiple Gaussian distributions.

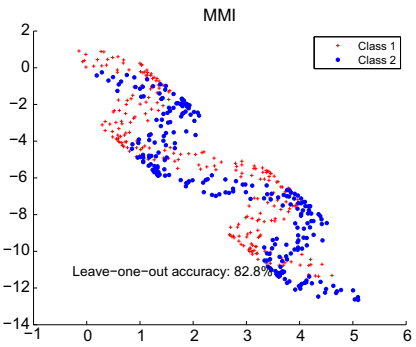
Please cite this article in press as: Qiu, X., Wu, L. Info-margin maximization for feature extraction. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.08.006

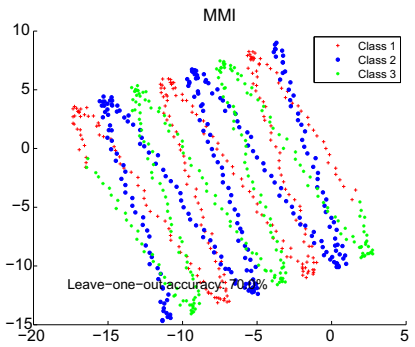

Original
Class 1
Class 2

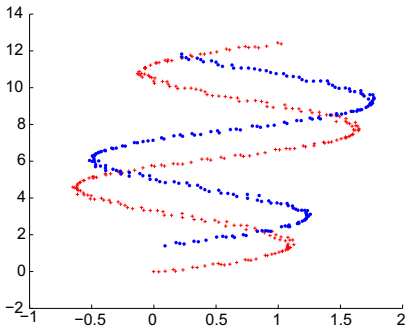

Original
Class 1
Class 2
Class 3


LDA
Class 1
Class 2
Leave−one−out accuracy: 69.0%


LDA
Class 1
Class 2
Class 3
Leave−one−out accuracy: 13.9%


MMI
Class 1
Class 2
Leave−one−out accuracy: 82.8%


MMI
Class 1
Class 2
Class 3



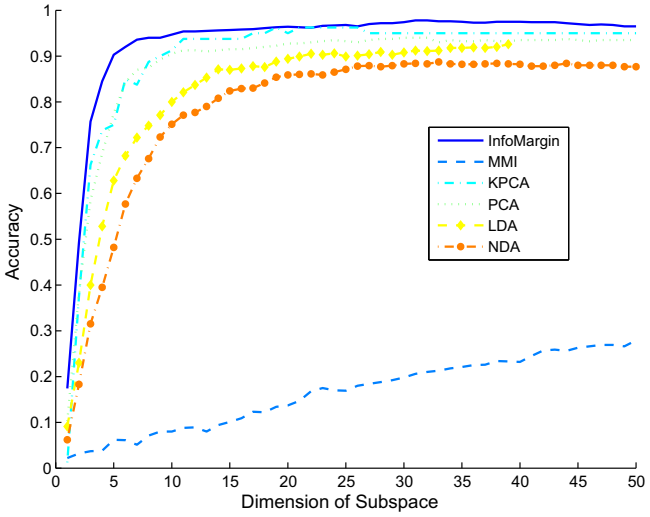


Fig. 4. Accuracy on AT&T face dataset.

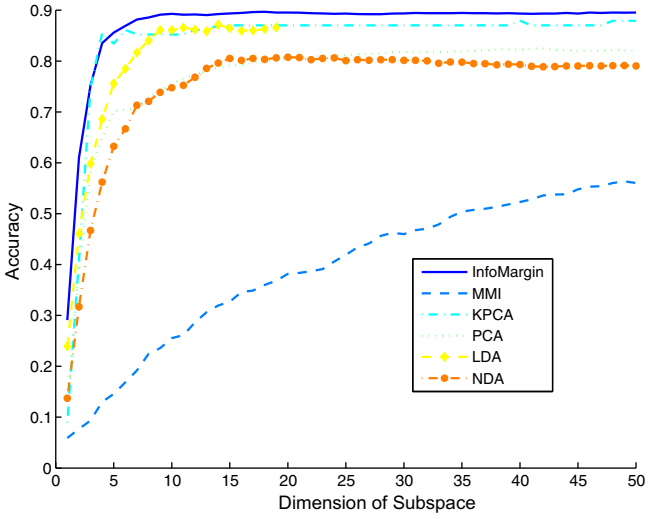


Fig. 5. Accuracy on UMIST face dataset.

directions, so the mutual information between extracted features and class label is not good measure to estimate the separability of classes.

Table 2
Descriptions of the UCI datasets in our experiments.

Dataset	Number of instances	Number of classes	Number of features
wdbc	569	2	30
sonar	208	2	60
isolet	1559	26	617
arrhythmia	452	13	279

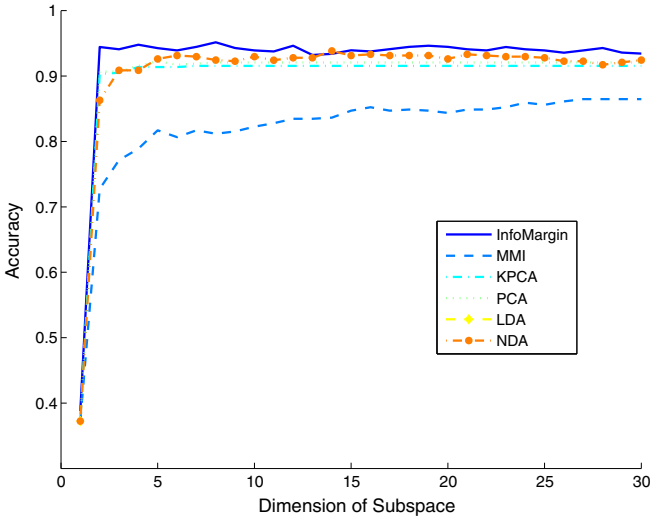


Fig. 7. Accuracy on wdbc dataset.

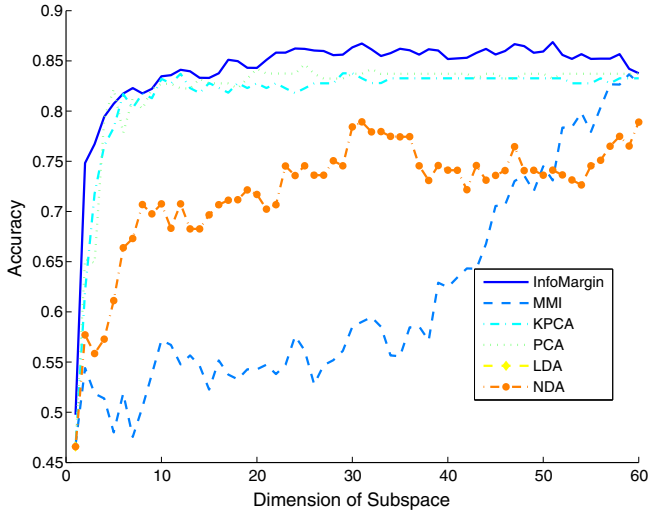


Fig. 8. Accuracy on sonar dataset.

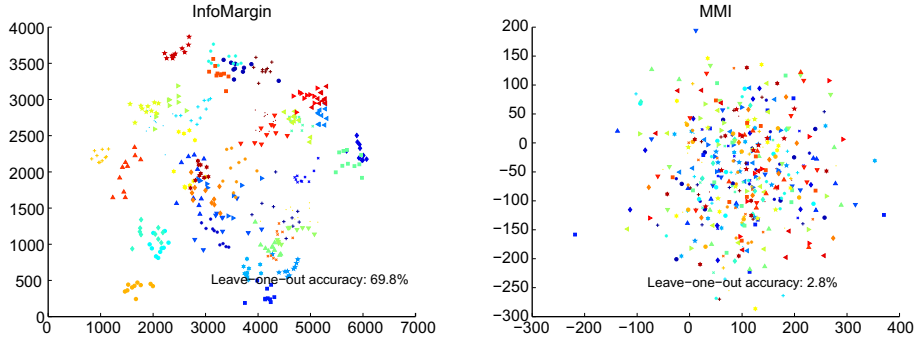


Fig. 6. 2D projections of AT&T dataset with InfoMargin (left) and MMI (right).

Please cite this article in press as: Qiu, X., Wu, L. Info-margin maximization for feature extraction. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.08.006

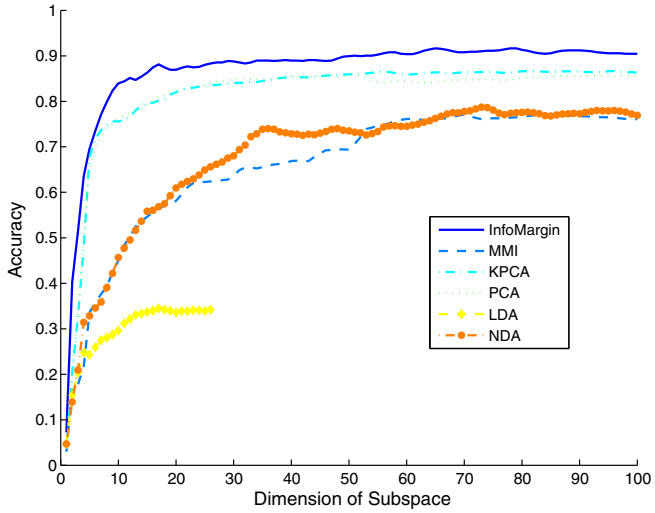


Fig. 9. Accuracy on isolet dataset.

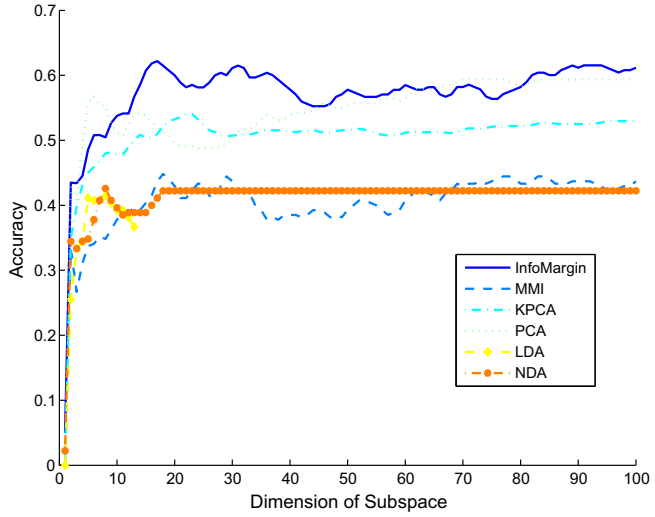


Fig. 10. Accuracy on arrhythmia dataset.

4.2. UCI machine learning repository

To evaluate the robustness of our method for the high-dimensional data, we also compare InfoMargin with the other methods on four high-dimensional datasets from UCI Machine Learning Repository (Asuncion and Newman, 2007): Wisconsin Diagnostic Breast Cancer dataset (wdbc), the Connectionist Bench dataset (sonar), the part of Isolated Letter Speech Recognition (isolet) and Cardiac Arrhythmia Database (arrhythmia). We use 10-fold cross-validation for nearest neighbor classification. A description of the datasets is presented in Table 2.

Figs. 7–10 show the accuracy with the different dimension of subspace on the four UCI datasets. As shown in the above experimental results, a major character is that info-margin maximization criterion always has a stable and high performance on the different datasets, while the other methods have unstable performances. For example, NDA does better on the UMIST dataset than on the wdbc dataset, but works badly on the sonar dataset. In addition, we found InfoMargin converges faster than MMI in our experiments.

5. Conclusions

In this paper, we proposed a new feature extraction method, info-margin maximization criterion (InfoMargin), which finds the important discriminant directions from information theoretic viewpoint and deals well with non-Gaussian class distribution. We use quadratic entropy and divergence measure with Gaussian kernel Parzen density estimator, which enhances the computational efficiency greatly. Our experiments show that our method is very efficient and robust. In the further works, we will extend info-margin maximization criterion to nonlinear feature extraction.

References

Asuncion, A., Newman, D., 2007. UCI machine learning repository. URL: <http://www.ics.uci.edu/~mlearn/MLRepository.html>.
Belhumeur, P., Hespanda, J., Kiregeman, D., 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Machine Intell. 19 (7), 711–720.
Bian, Z., Zhang, X., 1999. Pattern Recognition, second ed. Tsinghua University, Beijing, China.
Bollacker, K., Ghosh, J., 1996. Linear feature extractors based on mutual information. In: Proc. 13th ICPR, pp. 720–724.
Chen, L., Liao, H., Ko, M., Lin, J., Yu, G., 2000. A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recogn. 33 (10), 1713–1726.
Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley-Interscience, New York, NY, USA.
Crammer, K., Gilad-Bachrach, R., Navot, A., Tishby, N., 2002. Margin analysis of the lvq algorithm. In: Proc. Neural Information Processing System (NIPS).
Friedman, J.H., 1987. Exploratory projection pursuit. J. Amer. Statist. Assoc. 82 (397), 249–266.
Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, second ed. Academic Press, Boston.
Fukunaga, K., Mantock, J., 1983. Nonparametric discriminant analysis. IEEE Trans. Pattern Anal. Machine Intell. 5, 671–678.
Graham, D.B., Allinson, N.M., 1998. Characterizing virtual eigensignatures for general purpose face recognition. (in) Face Recognition: From Theory to Applications, NATO ASI Series F, Computer and Systems Sciences 163, 446–456.
Hild, K.E., Erdogmus, D., Torkkola, K., Principe, J.C., 2006. Feature extraction using information-theoretic learning. IEEE Trans. Pattern Anal. Machine Intell. 28 (9), 1385–1392.
Kapur, J., 1994. Measures of Information and their Applications. Wiley, New Delhi, India.
Linsker, R., 1988. Self-organization in a perceptual network. IEEE Computer 21 (3), 105–117.
Lyu, S., 2005. Infomax boosting. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).
Principe, J.C., Xu, D., Fisher III, J.W., 2000. Information theoretic learning. In: Unsupervised Adaptive Filtering. Wiley, New York, NY, USA, pp. 265–319.
Rosset, S., Zhu, J., Hastie, T., 2003. Margin maximizing loss functions. In: Proc. Neural Information Processing System (NIPS).
Samaria, F., Harter, A., 1994. Parameterisation of a stochastic model for human face identification. In: Proc. 2nd IEEE Workshop on Applications of Computer Vision.
Schölkopf, B., Smola, A., Müller, K.-R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 10, 1299–1319.
Torkkola, K., 2003. Feature extraction by non-parametric mutual information maximization. J. Mach. Learn. Res. 3, 1415–1438.
Turk, M., Pentland, A., 1991. Eigenfaces for recognition. J. Cognitive Neurosci. 3 (1), 71–86.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., New York, NY, USA.
Vasconcelos, N., 2002. Feature selection by maximum marginal diversity. In: Proc. Neural Information Processing System (NIPS).
Weisstein, E.W., 2006. L'hospital's rule. From MathWorld – A Wolfram Web Resource. <http://mathworld.wolfram.com/LHospitalsRule.html>.

Please cite this article in press as: Qiu, X., Wu, L. Info-margin maximization for feature extraction. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.08.006