Parzen Discriminant Analysis

Youhan Fang[1], Shiguang Shan[2,3], Hong Chang[2,3], Xilin Chen[2,3], Wen Gao[2,4]
[1]*Graduate University of Chinese Academy of Sciences(CAS), Beijing, China*
[2]*Digital Media Research Center, Institute of Computing Tech., CAS, Beijing, China*
[3]*Key Laboratory of Intelligent Information Processing, CAS, Beijing,China*
[4]*Digital Media Institute, Peking University, Beijing, China*
yhfang@jdl.ac.cn; {sgshan, changhong, xlchen}@ict.ac.cn; wgao@pku.edu.cn

Abstract

In this paper, we propose a non-parametric Discriminant Analysis method (no assumption on the distributions of classes), called Parzen Discriminant Analysis (PDA). Through a deep investigation on the non-parametric density estimation, we find that minimizing/maximizing the distances between each data sample and its nearby similar/dissimilar samples is equivalent to minimizing an upper bound of the Bayesian error rate. Based on this theoretical analysis, we define our criterion as maximizing the average local dissimilarity scatter with respect to a fixed average local similarity scatter. All local scatters are calculated in fixed size local regions, resembling the idea of Parzen estimation. Experiments in UCI machine learning database show that our method impressively outperforms other related neighbor based non-parametric methods.

1. Introduction

In statistical pattern recognition, we keep on seeking feature extraction methods in order to describe the original data more efficiently and effectively. One process of obtaining such features from supervised data is Discriminant Analysis (DA). Linear Dimensionality Reduction (LDR) for classification purpose is one of the most important types of DA method for its computational efficiency. These methods actually learn a linear transformation that maps the data to a lower dimensional subspace in which the discriminative information can be maintained as much as possible.

Traditional DA methods seek to optimize criteria which are either related to the upper bound of Bayesian error rate (Chernoff bound [1], for instance), or empirically demonstrating monotonicity with the error rate (e.g. the Kullback-Leibler Divergence [1] or the μ -measure [6]). When using these criteria, the distribution of each class is supposed to be normal. The well-known Linear Discriminant Analysis (LDA) [1] even further assumes that the covariance of all classes be equal. These methods can be regarded as parametric DA methods, with parameters represented as the mean and the covariance of the supposed normal distributions. The normal distribution assumption sometimes approximately holds, considering LDA performs well in many applications. But such methods cannot work well for complex distributions. In contrast, non-parametric methods without the assumption of distributions can deal with this sort of data.

The methods in [2, 3, 8, 9] can be considered as non-parametric DA methods. Their criteria are defined to minimize/maximize the distances between each sample and its nearby similar (with the same class label)/dissimilar (with different class labels) samples. Since non-parametric density estimation methods are also based on the relationship between samples and their neighbors, the idea of this sort of methods should have a statistical explanation, though they did not propose the methods from this view, but only from the geometrical intuition, in their papers.

Other sorts of non-parametric methods include [7, 11], which calculate their error rate based criteria by the non-parametrically estimated distribution other than by the supposed normal distribution; and [4, 10], which directly define the criteria according to the decision boundary. Here, we do not discuss them in detail, but only focus on the neighbor based methods.

Using similar/dissimilar neighbors to define the criteria is an intuitive and rational idea for non-parametric DA. Methods based on this idea therefore may have good performances. At the same time, theoretical analysis on the relationship between this intuitive idea and the error rate is also very important, for it provides the explanation of the idea in the statistical view and reveals the best achievable performance of such methods in ideal situations.

However, this analysis, which provides a theoretical support for using similar/dissimilar neighbors, has not aroused enough attention by related methods before. In this paper, we investigate the non-parametric density estimation and find that minimizing/maximizing the distances between each sample and its similar/dissimilar neighbors is equivalent to minimizing an upper bound of the Bayesian error rate.

From the theoretical analysis, all neighbor based DA criteria can be considered as to minimize this upper bound. Two important things thereby emerge: one is the selection of the neighbors; the other is the formulation of the optimization function. We find that the criteria of previous methods are improperly defined in terms of these two aspects. In this paper, we propose a new neighbor based non-parametric DA method, called Parzen Discriminant Analysis (PDA). Our criterion is defined as maximizing the average local dissimilarity scatter with respect to a fixed average local similarity scatter. All local scatters are calculated in fixed size local regions.

In summary, our contribution lies on two aspects: 1) The theoretical explanation of the basic idea of neighbor based methods; 2) A new non-parametric DA method in which the definition of the criterion is more effective. In the following two sections, we present the theoretical analysis and the proposed method in detail, and also provide a comparison with related methods. Finally, in the experiment we show that our method remarkably outperforms other neighbor based methods as well as the classic LDA.

2. Parzen Discriminant Analysis

In non-parametric density estimation, the basic idea of estimating the density $p(\mathbf{x})$ at $\mathbf{x}$ is counting the number of samples in the neighborhood of $\mathbf{x}$. For a neighborhood with volume V, $p(\mathbf{x})$ does not change much in the area if V is small. Therefore, the density at $\mathbf{x}$ can be estimated as:

$$\hat{p}(\mathbf{x}) = \frac{k(\mathbf{x})/N}{V} \tag{1}$$

where $k(\mathbf{x})$ is the number of samples in the neighborhood, N is the total number of samples. And if $N \to \infty$, the estimation can become very accurate.

In classification tasks, the posterior probability can be also estimated in the similar way. Note that

$$p(\omega_i \mid \mathbf{x}) = p(\mathbf{x}, \omega_i) / p(\mathbf{x}) \tag{2}$$

where ω_i denotes the i-th class. And we have

$$\hat{p}(\mathbf{x}, \omega_i) = \frac{k_i(\mathbf{x})/N}{V} \tag{3}$$

where $k_i(\mathbf{x})$ is the number of samples which are in the neighborhood of $\mathbf{x}$ and belong to class i. Therefore, the posterior probability of class i at $\mathbf{x}$ is

$$\hat{p}(\omega_i \mid \mathbf{x}) = \hat{p}(\mathbf{x}, \omega_i) / \hat{p}(\mathbf{x}) = k_i(\mathbf{x}) / k(\mathbf{x}) \tag{4}$$

Because the Bayesian error rate of a given $\mathbf{x}$ is

$$p(e \mid \mathbf{x}) = 1 - \max_i (p(\omega_i \mid \mathbf{x})) \tag{5}$$

one can estimate the error rate by counting the numbers of samples belonging to different classes and select the biggest one to divide the total number of samples in $\mathbf{x}$'s neighborhood. But in practice, it might be easy to falsely determine the major class in this way, for we always cannot have unlimited samples.

A much safer strategy to estimate the error rate is using the similar/dissimilar samples. Because of the limitation of samples, this strategy may have a better estimation on the error rate. Specifically, if we know that $\mathbf{x}$ belongs to class j, it is most possible that

$$\max_i (p(\omega_i \mid \mathbf{x})) = p(\omega_j \mid \mathbf{x}) \tag{6}$$

Therefore, if we know that the sample $\mathbf{x}$ belongs to class j (the training data can provide such information), we can select $p(\omega_j \mid \mathbf{x})$ as the maximum posterior prob-ability. Here $p(\omega_j \mid \mathbf{x})$ can be estimated by counting the similar/dissimilar samples in $\mathbf{x}$'s neighborhood.

Consequently, by this strategy, we actually have the probability of $p(\omega_i \mid \mathbf{x})$ to select $p(\omega_i \mid \mathbf{x})$ as the maximum posterior probability for a given $\mathbf{x}$, where $i = 1,2,...c$ and c is class number. Thus, we have the expectation of the estimation of $p(e \mid \mathbf{x})$ as

$$\tilde{p}(e \mid \mathbf{x}) = 1 - \sum_{i=1}^{c} p^2(\omega_i \mid \mathbf{x}) \tag{7}$$

Note that (7) is exactly the same error rate of Nearest Neighbor classifier when $N \to \infty$. It is well-known that this error rate is bounded by two times of Bayesian error rate [1].

From the analysis above, we find the theoretical support for using similar/dissimilar neighbors to define the DA criterion. That is, to minimize the upper bound (7) of Bayesian error rate, we have to maintain in each sample's neighborhood as many similar samples, and as few dissimilar samples as possible.

We want to define our criterion to minimize the upper bound and at the same time expect an analytical solution, which is known as the LDA's advantage and is very important for the convenience of practical uses.

Both these two goals can be attained by minimizing/maximizing the average distance between samples and their similar/dissimilar neighbors (see figure 1 (a)). Similar with LDA, for the purpose of learning a linear subspace, our criterion can be defined as maximizing the average local dissimilarity scatter

S_E with respect to a fixed average local similarity scatter S_I, namely,

$$J_{NLDR}(W) = tr((WS_IW^t)^{-1}WS_EW^t) \quad (8)$$

$$S_E = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N^E_{R(\mathbf{x}_i)}}\sum_{\substack{\mathbf{x}_j\in R(\mathbf{x}_i),\\ l(\mathbf{x}_j)\neq l(\mathbf{x}_i)}}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^t \quad (9)$$

$$S_I = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N^I_{R(\mathbf{x}_i)}}\sum_{\substack{\mathbf{x}_j\in R(\mathbf{x}_i),\\ l(\mathbf{x}_j)=l(\mathbf{x}_i)}}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^t \quad (10)$$

where $R(\mathbf{x}_i)$ denotes the neighborhood of $\mathbf{x}_i$, and $\mathbf{x}_j \in R(\mathbf{x}_i)$ if $\|\mathbf{x}_i-\mathbf{x}_j\| \le r$. $N^I_{R(\mathbf{x}_i)}/N^E_{R(\mathbf{x}_i)}$ denotes the similar/dissimilar sample numbers in $R(\mathbf{x}_i)$. Because the idea of using fixed size local regions is similar with Parzen density estimation, we then call our method as Parzen Discriminant Analysis.

To optimize the criterion, we can do eigenvalue decomposition of the matrix $S_I^{-1}S_E$ and select d eigenvectors associated with the largest d eigenvalues as the row vectors of the projection matrix W.

The parameter r determines the numbers of samples which we use to calculate local scatters. To set the parameter easier in practice, one can first find a reference value Δ which is equal to the average distance between each sample and its nearest neighbor, and set $r = \varepsilon\Delta$. In this way, one can use ε, which is less sensitive to different data, to replace r.

3. Previous Neighbor Based Methods

In [2], the authors use the nearest similar/dissimilar neighbor of each sample to calculate the local scatters, whereas we use multiple neighbors contained in each sample's local region. In [8], the authors use k nearest similar/dissimilar neighbors to calculate local scatters. The difference is that they use the fixed number of neighbors, whereas we only consider the samples in a local region. So we neglect the dissimilar samples too far away from each sample, in other words, we only focus on the samples near the boundary of different classes when calculating the dissimilarity scatter, whereas they using all samples (see figure 1 (b)). In addition, their criterion is defined as $J(W) = W(S_E - S_I)W^t$ and W is restricted to be orthogonal. In our opinion, however, the subtract form is not proper. The ratio form, as we use, is invariant to the different scales in different dimensions, whereas the subtract form is not.

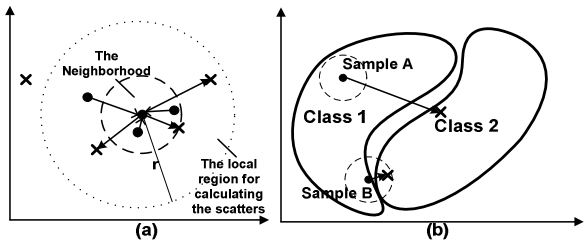

Figure 1. (a) Minimize/maximize the average distance between the sample and its neighbors in the local region (the larger circle) in order to maintain in the neighborhood (the smaller circle) more/less similar/dissimilar samples. (b) The dissimilar sample of sample B (A) should (should not) be included to calculate the local dissimilarity scatter.

In [9], the authors calculate the similarity scatter in the same way as in [8], and calculate the dissimilarity scatter by using sample pairs on the boundaries. But they simply find k nearest pairs of samples with different class labels to describe the boundaries. The parameter k, however, is very difficult to set.

In summary, previous neighbor based methods also use similar/dissimilar neighbors, but not in the most proper ways. They have some problems in either the selection of neighbors or the formulation of optimizing function. In contrast, our method is more rational and effective in terms of the two aspects.

4. Experiment

We test on six datasets of the UCI repository [12]. The experiment covers a large range of original dimensions and class numbers. Except for the 'Vowel' dataset which has its own training/testing sets, we divide each dataset into ten-fold and do cross-validation to estimate the mean classification error rates. In the experiment, we compare our method with NDA [2] and ANMM [8], which are non-parametric neighbor based methods, and with the classic LDA which is parametric. To compare the rationality of subtract form and the ratio form we mentioned in section 3.2, we also give the result obtained by MMC [5] which uses the same within/between class scatters with LDA but the criterion is of the subtract form.

The best classification results of every method in each dataset are listed in table 1, and the results with various dimensions of the 'Wine' dataset are illustrated in figure 2 (a). Here each method's parameters are set to be optimal. We use the NN classifier which is a non-parametric classifier. For fair comparison, all projection matrices learned by different methods are orthogonalized. Because these methods' criteria actually only learn the discriminative subspace, not the affine transformation.

Table 1. The results: lowest mean error rate (%) and the corresponding dimension (in bracket). '***Full***' denotes no dimension reduction. Best results are marked by asterisks.

Datasets	*Full*	*LDA*	*MMC*	*NDA*	*ANMM*	*PDA*
Pima	33.2	31.6(1)	32.1(7)	34.3(3)	30.7(4)	28.7(3)*
WDBC	8.6	4.3(1)	10.0(29)	4.8(10)	8.4(3)	3.0(3)*
Sonar	18.0	27.5(1)	15.0(55)	12.0(21)	13.0(15)	11.0(45)*
Wine	21.8	1.2(2)	21.2(10)	11.2(2)	21.8(4)	0.0(3)*
Vehicle	35.5	24.9(3)	24.3(13)	22.0(7)	26.3(12)	18.9(6)*
Vowel	45.2	44.6(9)	47.4(9)	44.8(9)	46.3(9)	42.0(8)*

Overall, we can see from table 1 that our method remarkably outperforms other neighbor based methods as well as LDA and MMC. Note that in the 'WDBC' and 'Wine' datasets, for instance, other non-parametric methods perform even much worse than LDA. In contrast, our method has impressive and stable performs among all datasets.

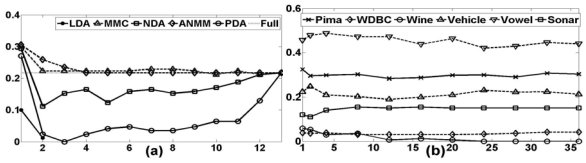

Figure 2. (a) The classification error rates with various dimensions in the 'Wine' dataset. (b) The influence of the parameter ε (the attainable lowest error rate in various ε).

It can be also seen that the ratio form is better than the subtract form from table 1. And note that in figure 2 (a), the performances of ANMM and MMC stop increase even they are still not better than no dimension reduction. They might fail to capture the real discriminative subspace, due to the improperly defined criteria. The influence of the parameter ε is also shown in figure 2 (b). One can see that our method is not very sensitive to the parameter.

We have also tried our method in some face recognition tasks, though the statistical nature of the data is not obvious. Since the data is too sparse (very few samples in each class and very high original dimension), local regions, which have to be large enough to contain sufficient similar samples for an accurate calculation, inevitably contain much more dissimilar samples (due to extreme imbalance of the two types of samples). So, exerting a limitation on the number of dissimilar samples becomes necessary. And the limitation finally makes the criterion of PDA very similar with those criteria using fixed number of similar/dissimilar neighbors. So using fixed number of neighbors may be a good strategy when dealing with intensely sparse data.

5. Concluding Remarks

Through investigating the non-parametric density estimation, we find that the basic idea of neighbor based DA methods is equivalent to minimizing an upper bound of Bayesian error rate. Based on this theoretical analysis, we define a more proper non-parametric DA criterion. Promising experimental results in statistical pattern recognition tasks illustrate the reasonability and effectiveness of our method.

Finally, it is worth noting that our method can be easily extended to a kernelized version, by following the similar steps of kernelizing LDA. With kernel PDA, we can expect even more effective algorithm performance and wider applications.

Acknowledgement

This paper is partially supported by National Natural Science Foundation of China (No.60332010 and No.60772071); Hi-Tech Research and Development Program of China under contract No.2007AA01Z163; ISVISION Technology Co. Ltd.

References

[1] K. Fukunaga. Introduction to statistical pattern recognition. New York: Academic Press, 1990.
[2] M. Bressan and J. Vitrià, Nonparametric discriminant analysis and nearest neighbor classification, Pattern Recognition Lett. 2743-2749, 2003.
[3] K. Fukunaga and S. Ando. Nonparametric discriminant analysis. PAMI, 671-678, 1983.
[4] C. Lee and D. Landgrebe. Feature extraction based on decision boundaries. PAMI, 15 (4), 388-400, 1993.
[5] H. Li, T.Jiang and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. NIPS, 2004.
[6] Z. Nenadic. Information discriminant analysis: feature extraction with an information-theoretic objective. PAMI, vol.29, no.8, 1394-1407, 2007.
[7] K. Torkkola. Feature extraction by non-parametric mutual information maximization. JMLR, 1415-1438, 2003.
[8] F. Wang and C. Zhang. Feature extraction by maximizing the average neighborhood margin. CVPR, 2007.
[9] S. Yan, D. Xu, B. Zhang and H. Zhang. Graph embedding: A general framework for dimensionality reduction. CVPR, 830–837, 2005.
[10] J. Zhang and Y. Liu. SVM decision boundary based discriminative subspace induction. Pattern Recognition, vol.38, 1746-1758, 2005.
[11] M. Zhu and T. Hastie. Feature extraction for nonparametric discriminant analysis, J. of Computational and Graphical Stat. 12, no.1, 101-120, 2003.
[12] P.M. Murphy and D.W. Aha. UCI Repository of Machine Learning Databases. URL:http://www.ics.uci.edu /mlearn/mlrepository.html,2004.