ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Visual Nonlinear Discriminant Analysis for Classifier Design

Tomoharu Iwata, Kazumi Saito and Naonori Ueda

NTT Corporation - NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Keihanna Science City, Kyoto - Japan

Abstract. We present a new method for analyzing classifiers by visualization, which we call *visual nonlinear discriminant analysis*. Classifiers that output posterior probabilities are visualized by embedding samples and classes so as to approximate posterior probabilities using parametric embedding. The visualization provides a better intuitive understanding of such classifier characteristics as separability and generalization ability than conventional methods. We evaluate our method by visualizing classifiers for an artificial data set.

1 Introduction

Designing better classifiers is a major research issue as regards machine learning, and it has been studied for a long time. In terms of classifier design, we must understand classifier characteristics, for example, the degree to which samples can be separated, which samples are difficult to classify, and which classes are closely related. However, this is difficult since the input may consist of high dimensional vectors and there may be many classes. Visualization, which is used to understand complex and high dimensional data in broad applications, can be used to understand classifiers. This visual process is called visual discriminant analysis [2]. Fisher linear discriminant analysis (FLDA)[3] is used for visual discriminant analysis, but we can only observe linear separability even though many nonlinear classifiers have already been proposed.

In this paper, we present an new method for analyzing classifiers by visualization, which we call visual nonlinear discriminant analysis (VNDA). We intend to apply this to classifiers that output posterior probabilities, where samples are classified in terms of the class that has the highest posterior probability. This type of classifier is widely used. One example is a generative classifier that estimates posterior probabilities by the Bayes rule, and another is a logistic regression that directly models the posterior probability as a function[4]. In VNDA, we visualize classifiers by embedding samples and classes into two or three dimensional Euclidean spaces based on posterior probabilities using parametric embedding (PE)[5]. PE seeks an embedding so as to approximate posterior probabilities as closely as possible. The visualization enable us to understand classifiers intuitively and this helps with their design.

2 Visual Nonlinear Discriminant Analysis

Let $\{(P(c_1|\boldsymbol{x}_n), \ldots, P(c_K|\boldsymbol{x}_n))\}_{n=1}^N$ be a set of posterior probabilities estimated by a classifier, where $\boldsymbol{x}_n$ is a sample, c_k is a class, N is the number of samples, and K is the number of classes. $\boldsymbol{x}_n$ can be a vector, and also a sequence or a graph. Given a set of posterior probabilities, we visualize classifiers by embedding samples and classes into two- or three-dimensional space such that the posterior probabilities are approximated using PE. Let $\boldsymbol{r}_n$ be a two- or three-dimensional vector, which represents the coordinates of a sample $\boldsymbol{x}_n$ in the visualization space, and $\boldsymbol{\phi}_k$ be a two- or three-dimensional vector, which represents the coordinates of a class c_k in the visualization space. In the visualization space, PE assumes a unit variance Gaussian mixture. Then, the posterior probability in the visualization space is

$$P(c_k|\mathbf{r}_n) = \frac{P(c_k)\exp(-\frac{1}{2} \parallel \mathbf{r}_n - \boldsymbol{\phi}_k \parallel^2)}{\sum_{l=1}^K P(c_l)\exp(-\frac{1}{2} \parallel \mathbf{r}_n - \boldsymbol{\phi}_l \parallel^2)},$$

where $\| \cdot \|$ is the Euclidean norm. The sum of the KL divergences is a natural measurement for the degree of posterior probability approximation, as follows:

$$E(\{\boldsymbol{r}_n\}, \{\boldsymbol{\phi}_k\}) = \sum_{n=1}^N \sum_{k=1}^K P(c_k|\boldsymbol{x}_n) \log \frac{P(c_k|\boldsymbol{x}_n)}{P(c_k|\boldsymbol{r}_n)}. \quad (1)$$

We can obtain sample and class coordinates in the visualization space, $\{\boldsymbol{r}_n\}$, $\{\boldsymbol{\phi}_k\}$, by minimizing the above objective function using optimization methods such as quasi-Newton methods[6].

With the assumption of a unit variance Gaussian mixture in PE, the posterior probability in the visualization space becomes a function of Euclidean distance between the sample and the class. And if the posterior probability is high, the sample and the class are embedded closely; if it is low, they are embedded far away from each other. Therefore, the visualization provides us with an intuitive understanding of posterior probabilities, which represent classifier characteristics. If few samples are located between classes, it suggests that the discrimination between these classes is easy; otherwise, the discrimination is difficult. The visualization also shows which samples are misclassified, and which class's samples are likely to be misclassified in which class. From the difference between the coordinates of learning and test samples, we can also determine the generalization ability of the classifiers.

It is difficult to distinguish samples that are located too closely in the visualization space. This difficulty commonly happens in other visualization method, such as multi-dimensional scaling and FLDA. For facilitate visualization, we want to find a set of new coordinates, in which the new coordinates $\tilde{\boldsymbol{r}}_n$ is close to the coordinates $\boldsymbol{r}_n$ as much as possible but $\tilde{\boldsymbol{r}}_n$ does not overlap with others $\tilde{\boldsymbol{r}}_m$, where $m \neq n$. This can be achieved by minimizing the following objective

fiction:

$$J(\{\tilde{\boldsymbol{r}}_n\}) = \frac{1}{2}\sum_{n=1}^{N} \parallel \tilde{\boldsymbol{r}}_n - \boldsymbol{r}_n \parallel^2 + \eta \sum_{n=1}^{N}\sum_{m\neq n} \exp(-\frac{1}{2\sigma^2} \parallel \tilde{\boldsymbol{r}}_n - \tilde{\boldsymbol{r}}_m \parallel^2), \quad (2)$$

where the first term represents the distances between new coordinates and coordinates that are output of PE, the second term represents the degree of overlapping among new coordinates, and $\eta > 0$, $\sigma^2 > 0$.

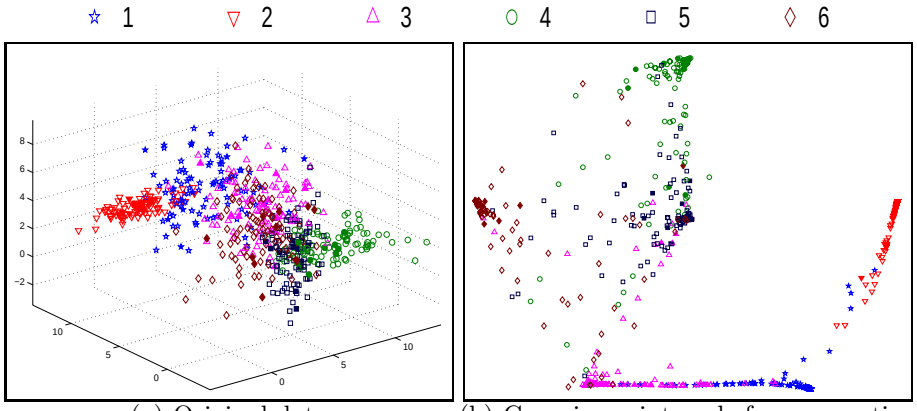
3 Experimental Results

We evaluated VNDA by visualizing four classifiers of an artificial data set. In each class, 10 learning and 90 test samples were generated from a six-class three-dimensional Gaussian mixture with different covariances. Figure 1(a) shows the original data. We used four classifiers: Gaussian mixture (GM), same covariance Gaussian mixture (SGM), probabilistic nearest neighbor (PNN), and quadratic logistic regression (QLR)[4]. VNDA is applicable to linear classifiers, and also nonlinear, non-parametric, or discriminant classifiers that output posterior probabilities.

Figure 1(b) is the visualization result of GM before minimizing Equation 2, in which many samples are overlapping. By separating off the overlapping samples with the minimization of Equation 2, we can get a more facilitate visualization as in Figure 1(c). In Figure 1(c), samples of the same class form a cluster. This indicates that GM classifies samples appropriately. Some samples are located between c_1 and c_3, reflecting the fact that it is difficult to discriminate between them. Conversely, c_2 samples are separated well, reflecting the fact that the discrimination of c_2 samples is easy. In Figure 1(d), the samples are scattered and there are no clear clusters since SGM has an improper assumption. In Figure 1(e), other class samples are located in the c_5 cluster, indicating that these samples are likely to be misclassified in c_5. In Figure 1(f), six clusters are well separated, however, some test samples are located in other class clusters, indicating overfitting. On the other hand, in Figure 1(b), learning and test samples are spread through each cluster since GM has a good generalization ability.

Table 1 shows the precisions and confusion matrices of classifiers. They are commonly used to evaluate the performance of classifiers. Of course, the confusion matrices are more quantitative than the visualizations, however, it is difficult to understand at a glance, especially in the case that many classes exists. In addition, the visualization has more information than the confusion matrix. For example, QLR clearly separated samples, which means that the posterior probabilities are approximately one or zero, and SGM did not separate samples. These characteristics are clear in the visualizations as in Figure 1(d)(f). On the other hand, we cannot understand these characteristics from the confusion matrices, Table1(b)(c).

For comparison, we visualized GM, SGM and QLR classifiers by FLDA and kernel discriminant analysis (KDA)[1][7]. FLDA linearly embeds samples so as to



(a) Original data (b) Gaussian mixture before separation

(c) Gaussian mixture (d) Same covariance Gaussian mixture

(e) Probabilistic nearest neighbor (f) Quadratic logistic regression

Fig. 1: The original three dimensional data (a), the two dimensional visualization of the Gaussian mixtures before separation of overlapping samples (b), and two dimensional visualizations of classifiers (b)∼(f) with VNDA. Filled and not-filled particles represent learning and test samples, respectively, and the shape indicates the class.

GM	SGM	PNN	QLR
0.794(0.917)	0.746(0.733)	0.769(0.933)	0.669(1.000)

(a) Precisions

	1	2	3	4	5	6
1	66(7)	13(2)	8(1)	0(0)	3(0)	0(0)
2	6(1)	84(9)	0(0)	0(0)	0(0)	0(0)
3	6(1)	4(0)	55(7)	8(1)	12(0)	5(1)
4	1(0)	0(0)	0(0)	75(6)	14(4)	0(0)
5	0(0)	0(0)	14(0)	20(2)	48(8)	8(0)
6	0(0)	1(0)	5(0)	2(0)	7(3)	75(7)

(b) Same covariance Gaussian mixture

	1	2	3	4	5	6
1	51(10)	26(0)	13(0)	0(0)	0(0)	0(0)
2	3(0)	87(10)	0(0)	0(0)	0(0)	0(0)
3	7(0)	3(0)	67(10)	1(0)	9(0)	3(0)
4	0(0)	1(0)	1(0)	54(10)	29(0)	5(0)
5	2(0)	0(0)	2(0)	16(0)	51(10)	19(0)
6	3(0)	0(0)	13(0)	15(0)	8(0)	51(10)

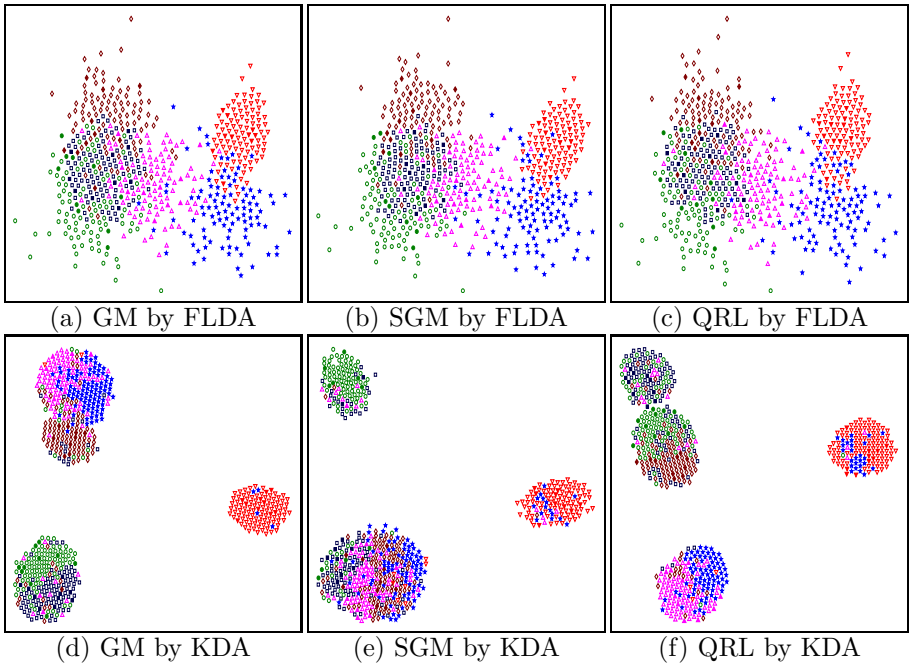
(c) Quadratic logistic regression

Table 1: The precisions (a) and confusion matrices (b)(c) of classifiers. The values in and out of parenthesis are the values of learning and test samples, respectively. In the confusion matrix, each row represents an true class, and each column represents an estimated class.

maximize between-class variance and minimize within-class variance. KDA is an extension of FLDA to nonlinear embedding using the kernel trick. Figure 2 shows the visualizations. We separated overlapping samples by minimizing Equation 2 after we obtained coordinates by FLDA and KDA. As inputs, FLDA and KDA take samples and their estimated classes, but not their posterior probabilities. Therefore, unlike with VNDA, differences between classifiers are not expressed in their visualization. In Figure 2(a)$\sim$(c), the clusters overlap because of the limitation of the linear method. In Figure 2(d)$\sim$(f), some clusters are clearly separated, although some samples cannot be clearly classified by their classifiers. FLDA and KDA need more computational time than VNDA since they lead to generalized eigenvalue problems, whose complexity is cubic in the matrix size.

4 Concluding Remarks

We presented a new method for analyzing classifiers by visualization, namely visual nonlinear discriminant analysis. We showed experimentally that the visualization results represent such classifier characteristics as separability and generalization ability. Our method can assist classifier design with the complementary use of conventional evaluation, such as a precision or confusion matrix. In the experiments, we visualized different classifiers for artificial data. The visualization can also show how a classifier changes with increases in the number of learning samples or with changes in control parameters. We now plan to develop an interactive visualization system with the proposed method to facilitate classifier design.



(a) GM by FLDA (b) SGM by FLDA (c) QRL by FLDA

(d) GM by KDA (e) SGM by KDA (f) QRL by KDA

Fig. 2: Two dimensional visualization of GM, SGM, QRL classifiers by FLDA(a)∼(c) and KDA(d)∼(f). The symbols are the same as in Figure 1.

References

[1] G. Baudat and F. Anouar, Generalized discriminant analysis using a kernel approach, *Neural Computation*, 12:2385–2404, 2000.

[2] I. Dhillon, D. Modha and W. Spangler, Class visualization of high-dimensional data with applications, *Computational Statistics and Data Analysis*, 41:59–90, 2002.

[3] R. Fisher, The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7:179–188l, 1950.

[4] T. Hastie, R. Tibshirani, J. Friedman and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York, 2001.

[5] T. Iwata, K. Saito, N. Ueda, S. Stromsten, T. Griffiths and J. Tenenbaum, Parametric embedding for class visualization, *Advances in Neural Information Processing Systems*, 17:617–624, 2005.

[6] D. C. Liu, and J. Nocedal, On the limited memory BFGS method for large scale optimization, *Math. Probramming*, 45(3):503–528, 1989.

[7] S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K. Muller, Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX*, IEEE, pages 41–48, 1999.