Classifier-Independent Visualization of Supervised Data Structures Using a Graph

Hiroshi Tenmoto[1], Yasukuni Mori[2], and Mineichi Kudo[3]

[1] Kushiro National College of Technology,
Kushiro, Hokkaido 084-0916, Japan
tenmo@kushiro-ct.ac.jp
http://www.kushiro-ct.ac.jp/tenmo/
[2] Chiba University, Chiba 263-8522, Japan
yasukuni@faculty.chiba-u.jp
[3] Hokkaido University, Sapporo 060-8628, Japan
mine@main.eng.hokudai.ac.jp

Abstract. Supervised data structures in high dimensional feature spaces are displayed as graphs. The structure is analyzed by normal mixture distributions. The nodes of the graph correspond the mean vectors of the mixture distributions, and the location is carried out by Sammon's nonlinear mapping. The thickness of the edges expresses the separability between the component distributions, which is determined by Kullback-Leibler divergence. From experimental results, it was confirmed that the proposed method can illustrate in which regions and to what extent it is difficult to classify samples correctly. Such visual information can be utilized for the improvement of the feature sets.

1 Introduction

In usual pattern recognition systems, both construction of classifiers and classification of unknown samples are carried out in a high dimensional vector space, called "feature space." The space is spanned by the "features" extracted from raw patterns in the observation space. Therefore, the system's accuracy strongly depends on the feature extraction part. However, the extraction process cannot be generalized, so the system has to be improved heuristically by evaluating the extracted features.

For this purpose, many feature selection method have been proposed. They can find and remove redundant features, while they cannot illuminate what features should be added. Therefore, on the other hand, many trials of visualizations that illustrate the distributions of samples in the feature space have been carried out [1-7]. By using such methods, the following properties can be observed: (1) in what "shape" the samples distribute in the high dimensional space, (2) whether the class regions are separated from the others sufficiently or not, and (3) if it is not, in which regions the classes are not separated well. These matters may help us in improving the system's accuracy, and are also useful for superclass finding problem [8].

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 1043–1051, 2004.

2 Visualization Methods

To achieve the purpose mentioned above, many "mapping" methods [1-7] can be applied. These mapping techniques are characterized from the following viewpoints:

1. Using supervise information or not.
2. Displaying the individual points or representative points.
3. Linear mapping or nonlinear mapping.

The relationship among these methods can be illustrated as Fig.1.

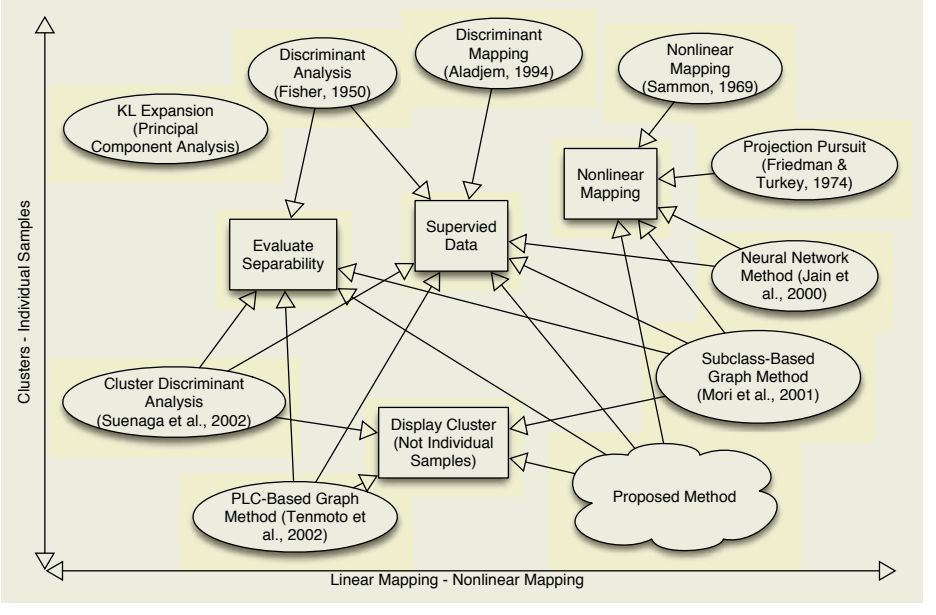


Fig. 1. Visualization methods of high-dimensional data.

When the supervise information, i.e., the class labels, are used, the separation status between the class regions can be observed. Therefore, almost latest methods are included in that group. In addition, the latest studies display representative points instead of the individual points. This is effective especially in the case of numerous samples. Once the points in a high dimensional space were mapped onto two dimensional space, the visual can trick us that the near points on the resultant map are also near in the feature space. This phenomenon can be happened both in linear mapping and nonlinear mapping methods.

To avoid these misunderstandable situations, Mori *et al.* proposed a novel visualization method [6]. In that technique, the individual samples in the feature space are organized to special type of overlapping clusters by a nonparametric classifier called subclass method [9], and such clusters are displayed as the

nodes of a graph. The clusters called "subclasses" are formed so as to include the samples belonging to a certain class (positive samples) as many as possible and exclude the samples belonging to the other classes (negative samples) completely. In addition, the method connects the nodes according to the overlapping hypervolumes between the subclasses in order to display the separability.

However, the method tends to produce much subclasses in order to exclude the negative samples, then the resultant graph becomes complex. Although such a parameter that is used to eliminate too small subclasses in the graph is prepared, it is difficult to judge whether such nodes can be removed or important for showing the separability of the class regions.

From these points of view, in this study, normal mixture distributions are employed for the analysis of the structures in order to absorb the effects of the noise or outlier samples. Hence, Kullback-Leibler divergence is used to illustrate the separability between the component distributions.

3 Proposed Method

3.1 Overview

The supervised data structure is displayed as a "graph" by the following procedure:

1. Estimate normal mixture distributions on the samples in the feature space.
2. Project the mean vectors of the component distributions onto two dimensional space by Sammon's nonlinear mapping [1], and let them the nodes of a graph.
3. Calculate Kullback-Leibler divergences between the component distributions, and connect the nodes according to the values of the divergences.

The details of the steps are described in the following sections.

3.2 Normal Mixture Distributions

In this study, normal mixture distributions on the samples are estimated by EM algorithm [10] that maximizes the following likelihood:

$$L = \sum_{i=1}^{N} \log p(\boldsymbol{x}_i) = \sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{K} c_k \mathrm{N}(\boldsymbol{m}_k, \Sigma_k)(\boldsymbol{x}_i) \right\},$$

where $\boldsymbol{x}_i$ is the ith feature vector (sample), N is the number of the samples, K is the number of the components, $\mathrm{N}(\boldsymbol{m}_k, \Sigma_k)(\cdot)$ is a normal distribution with a mean vector $\boldsymbol{m}_k$ and a covariance matrix Σ_k, and c_k is the weight of the kth component $\left(\sum_{k=1}^{K} c_k = 1\right)$. Such mixture models are estimated for the individual classes.

Here, the appropriate value of K is determined by MDL (minimum description length) criterion [11]. Each hypothesis is evaluated by the following formula:

$$\mathrm{MDL}(K) = -L + \frac{1}{2}\left[(K-1) + K\left\{D + \frac{D(D+1)}{2}\right\}\right]\log N,$$

where D is the number of the features. The appropriate number of components $\hat{K}$ is determined by varying K from 1 through a fixed number $K_{\max}$, e.g., 10, as:

$$\hat{K} = \arg\min_{K} \mathrm{MDL}(K).$$

3.3 Nonlinear Mapping

In this study, the nodes of a graph are located by Sammon's nonlinear mapping [1]. This method projects the points in a high dimensional space so as to maintain the distances of every pair as far as possible, that is achieved by minimizing the following evaluation formula:

$$E = \frac{1}{\gamma}\sum_{i<j}^{N} \frac{\{\delta(\boldsymbol{x}_i, \boldsymbol{x}_j) - \delta(\boldsymbol{y}_i, \boldsymbol{y}_j)\}^2}{\delta(\boldsymbol{x}_i, \boldsymbol{x}_j)},$$

where $\boldsymbol{x}_i$ is the original vector and $\boldsymbol{y}_i$ is the corresponding projected vector,

$$\gamma = \sum_{i<j} \delta(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

and $\delta(\cdot,\cdot)$ is the Euclidean distance.

Unfortunately, the original Sammon's method cannot project additional points after the calculation of the mapping. Therefore, in this study, the mean vectors of the component distributions are tentatively added to the samples before performing the projection, and are picked up from the result.

3.4 Kullback-Leibler Divergence

In order to take into account the scattering information of the samples, the separability between the classes are evaluated by Kullback-Leibler divergence.

Because, in this study, normal mixture distributions are used for the analysis of the structure, the divergence between two components p, q is calculated directly from the distribution parameters as follows:

$$D(p||q) = \frac{1}{2}\left\{\log\det(\Sigma_q\Sigma_p^{-1}) - D + \mathrm{tr}(\Sigma_q^{-1}\Sigma_p) + (\boldsymbol{m}_p - \boldsymbol{m}_q)^t\Sigma_q^{-1}(\boldsymbol{m}_p - \boldsymbol{m}_q)\right\}.$$

The thickness of the edge between two nodes is determined proportional to $1/D(p||q)$. As a result, well-separated nodes have no or thin edges, while nonseparated nodes have thick edges.

Here, a threshold parameter θ for $D(p||q)$ is introduced to eliminate the edges with too large value of $D(p||q)$, because the resultant too thin edges are redundant and obstacle for the observation.

4 Experiments

The proposed method was tested on SHIP dataset [12]. This dataset aims to distinguish eight types of navy ships with eleven features, that were extracted from the ships' silhouette images. The total number of samples is 2545.

The result of Sammon's nonlinear mapping is shown in Fig.2, and the graphs obtained by the proposed method are shown in Fig.3 (a) and (b). In addition, the graphs obtained by Mori *et al.*'s method are also shown in Fig.4(a) and (b), corresponding to two different types of location methods. The parameters in their method were adjusted according to their guideline.

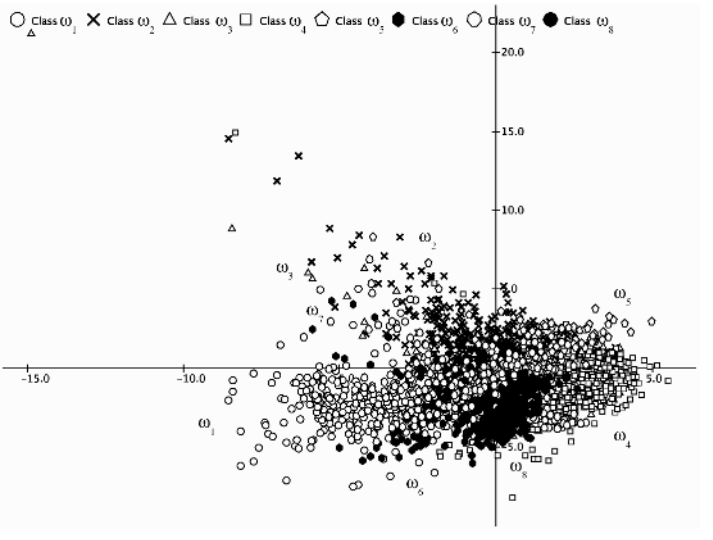
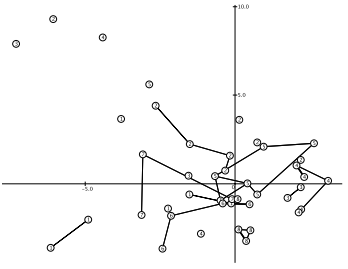


Fig. 2. Result for SHIP data by Sammon's nonlinear mapping.

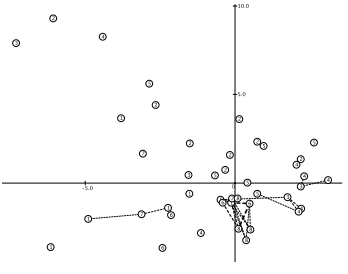
The result of Sammon's nonlinear mapping is very condensed, therefore it is difficult to find where the class $\omega_6, \omega_7, \omega_8$ samples distribute.

On the other hand, from the result of the proposed method (a), the essential distribution structure of each class is easily observed. For example, the class ω_5 forms "ring" structure, the class ω_8 forms unimodal cluster, and the other classes forms nonlinear belt structure and some isolated clusters.

In addition, from the result of the proposed method (b), it can be observed that in which regions correct classification is difficult. Here, two distant nodes in the graph does not necessarily mean that the separation of the corresponding samples is easy, and *vice versa*. For example, the distance between the class ω_1–ω_7 nodes at the left-bottom region in Fig.3(b) is far, but they should be nonseparable because there is the edge between the nodes. On the other hand, the class ω_2–ω_5 nodes at the right-upper region are close to each other, so it

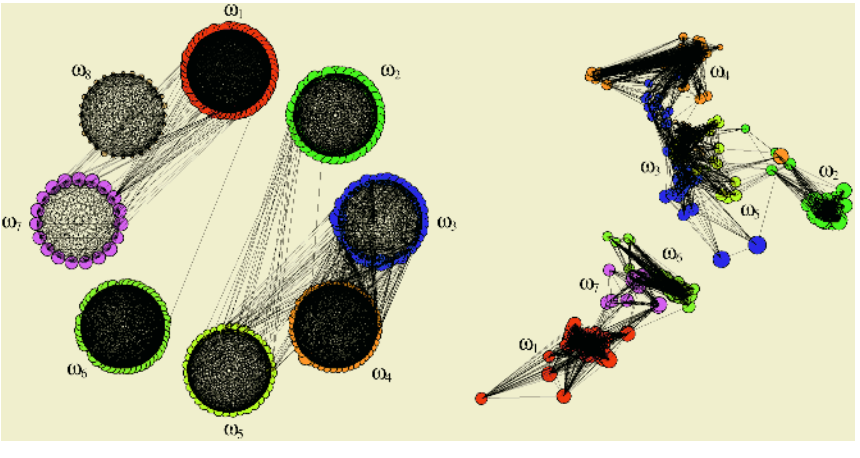


(a) Within-class edges with the threshold $\theta = 70$.



(b) Between-class edges with the threshold $\theta = 35$.

Fig. 3. Result for SHIP data by the proposed method. The numbers in the circles correspond the class labels.



(a) Polygon display type. (b) Principal component display type.

Fig. 4. Result for SHIP data by Mori *et al.*'s method.

seems very difficult to separate them. However, there is no edge between the nodes. This means the separation may be easy in the feature space.

In order to confirm that the nonseparable regions illustrated by the proposed method is right, the resultant graphs were compared to the confusion matrices

obtained by nearest neighbor method and SVM with soft-margin. The hold-out method was used for the calculation. The results are shown in Table 1 (a) and (b).

Table 1. Confusion matrices for SHIP data. Too much erroneous results are underlined.

(a) by nearest neighbor method.

I\O	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	Error
ω_1	143	1	0	0	1	12	11	2	0.159
ω_2	0	214	5	6	1	0	1	0	0.057
ω_3	0	7	72	9	5	0	0	0	0.226
ω_4	0	5	10	225	5	0	0	0	0.082
ω_5	2	6	0	5	154	5	1	1	0.115
ω_6	8	1	0	0	3	114	2	12	0.186
ω_7	9	5	0	0	1	0	102	2	0.143
ω_8	1	0	0	0	1	4	0	98	0.058

(b) by SVM with soft-margin.

I\O	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	Error
ω_1	140	3	0	0	0	19	6	2	0.176
ω_2	0	222	1	3	1	0	0	0	0.022
ω_3	1	6	68	9	9	0	0	0	0.269
ω_4	0	4	0	235	6	0	0	0	0.041
ω_5	0	3	1	5	161	3	0	1	0.075
ω_6	5	2	0	0	2	120	0	11	0.143
ω_7	17	5	0	0	2	1	92	2	0.227
ω_8	0	0	0	0	1	5	1	97	0.067

By comparing Fig.3 to the tables, it can be confirmed that the understanding of the resultant graph is almost consistent with the results of those classifiers. For example, in the confusion matrices, the most erroneous relationships between classes are ω_1–ω_6, ω_3–ω_4, ω_6–ω_8 and ω_1–ω_7. Such classes are also connected strongly in the result of the proposed method. This means that such class regions are very closed and/or the local variance of each class is large, then the local distributions are overlapped. Therefore, new features should be added in order to improve the system's accuracy at these classes. On the other hand, the other classes, for example, ω_1–ω_3 has no error in the tables, and there is also no edge between the classes in the graph. This means the current feature set is sufficient for such classes, and the features may be reduced by feature selection methods.

While, the resultant graphs obtained by Mori *et al.*'s method do not necessarily match the results of the proposed method and the two classifiers. For example, Fig.4 insists that there are much overlaps between the class ω_3–ω_4, ω_1–ω_7 and ω_2–ω_5, but no or not so much overlaps between the class ω_1–ω_6 and ω_6–ω_8. Two options can be considered for the reason: (1) the subclass classifier is so strong that succeeded to separate these class regions completely, (2) Important but small nodes were removed by the threshold parameter. As a result, the corresponding edges did not appear in the graph.

5 Conclusion

A new graph type visualization method for supervised data structures was proposed. The method uses normal mixture distributions for the analysis of the distribution structures, and uses Kullback-Leibler divergence for the evaluation of the separability between the class regions.

From the experimental results, it was confirmed that the proposed method can display the essential distribution structure by within-class edges, and also

can show by between-class edges in which regions and to what extent the class separation is not achieved sufficiently by the current feature set.

Compared to Mori *et al.*'s method, the proposed method shows the within-class structure simply and the between-class relationships appropriately. This property comes from the normal mixture distributions and MDL criterion in the complexity selection.

However, the proposed method also has a drawback in nature. The estimation of normal mixture models is difficult when the number of samples is relatively smaller than the number of features. Therefore, the estimation should be carried out with special care as far as possible. For example, it may be effective to substitute the basic EM algorithm to SMEM (Split and Merge EM) algorithm [13]. In addition, MDL framework for the selection of the appropriate number of component distributions does not work well if the number of samples is insufficient. Also at this point, good substitutions should be researched for the visualization purpose.

Acknowledgements

The authors are grateful to Professor Dr. Sklansky of University of California Irvine for providing the SHIP dataset.

This work was partly supported by the Ministry of Public Management, Home Affairs, Posts and Telecommunications of Japan under grant Strategic Information and Communications R&D Promotion Programme (SCOPE-S).

References

1. Sammon, J. W.: A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers **18** (1969) 401–409
2. Friedman, J. H., Tukey, J. W.: A Projection Pursuit Algorithm for Exploratory Data Analysis. IEEE Transactions on Computers **23** (1974) 1974 881–889
3. Aladjem, M.: Multiclass Discriminant Mappings. Signal Processing **35** (1994) 1–18
4. Mao, J., Jain, A. K.: Artificial Neural Networks for Feature Extraction and Multivariate Data Projection. IEEE Transactions on Neural Networks **6** (1995) 296–317
5. Jain, A. K., Duin, P. W., Mao, J.: Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 4–37
6. Mori, Y., Kudo, M., Toyama, J., Shimbo, M: Comparison of Low-Dimensional Mapping Techniques Based on Discriminary Information. Proceedings of the Second International ICSC Symposium on Advances in Intelligent Data Analysis (2001) CD-ROM Paper #1724-166
7. Tenmoto, H., Mori, Y., Kudo, M.: Visualization of Class Structures using Piecewise Linear Classifiers. Proceedings of Logic Applied to Technology (LAPTEC) (2002) 104–111.
8. Taylor, P. C., Hand, D. J.: Finding 'Superclassifications' with an Acceptable Misclassification Rate. Journal of Applied Statistics **26** 579–590
9. Kudo, M., Torii, Y., Mori, Y., Shimbo, M.: Approximation of Class Regions by Quasi Convex Hulls. Pattern Recognition Letters **19** (1998) 777–786

10. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B **39** (1977) 1–38
11. Rissanen, J.: A Universal Prior for Integers and Estimation by Minimum Description Length. Annals of Statistics **11** (1983) 416–431
12. Park, Y., Sklansky, J.: Automated Design of Multiple-Class Piecewise Linear Classifiers. Journal of Classification **6** (1989) 195–222
13. Ueda, N.: EM Algorithm with Split and Merge Operations for Mixture Models (Invited). Transactions of IEICE, **E83-D** (2000) 2047–2785.