Multiple Classifiers System for Reducing Influences of Atypical Observations

Šarūnas Raudys and Masakazu Iwamura

Vilnius Gediminas Technical University, Saulėtekio 11, Vilnius, Lithuania
Tohoku University, Aoba 05, Aramaki, Aoba, Sendai, 980-8579 Japan
E-mail: raudys@ktl.mii.lt, masa@aso.ecei.tohoku.ac.jp

Abstract. Atypical observations, which are called outliers, are one of difficulties to apply standard Gaussian density based pattern classification methods. Large number of outliers makes distribution densities of input features multimodal. The problem becomes especially challenging in high-dimensional feature space. To tackle atypical observations, we propose multiple classifiers systems (MCSs) whose base classifiers have different representations of the original feature by transformations. This enables to deal with outliers in different ways. As the base classifier, we employ the integrated approach of statistical and neural networks. This consists of data whitening and training of single layer perceptron (SLP). Data whitening makes marginal distributions close to unimodal, and SLP is robust to outliers. Various kinds of combination strategies of the base classifiers achieved reduction of generalization error in comparison with the benchmark method, the regularized discriminant analysis (RDA).

1 Introduction

In many real-world practical pattern recognition tasks including printed and handwritten character recognition, we often meet atypical observations, and also meet the classification problem of such observations with Gaussian classifiers. Outliers are the observations which follow another distribution. If the number of outliers is large, the distributions could be multimodal ones. Applying Gaussian model to multimodal distribution produces many outliers.

To deal with multimodal distributions, nonparametric (local) pattern recognition methods such as *k*-NN rule and Parzen window classifier could be used because they approach the Bayes classifier with large training samples. However, in high-dimensional and small sample cases, *sample size/complexity ratio* becomes low. In such situations, utilization of nonparametric methods is problematic [1-3].

To reduce influences of atypical observations, we suggest multiple classifier systems (MCSs) whose several base classifiers have different representations of the original feature. We perform different transformations of the original feature (including no transformation) in order to deal with outliers in different ways.

As the base classifier, we employ the integrated approach of statistical and neural networks. This approach is the combination of data whitening and training

of single layer perceptron (SLP) to recognize patterns. In data whitening, we also perform data rotation to achieve good start, speed up the SLP training, and obtain new features whose marginal distribution densities are close to unimodal ones and often resemble Gaussian distribution. In one base classifier, for data rotation we utilized robust estimates of mean vectors and pooled covariance matrix. The SLP based classifier is inherently robust to outliers.

We considered various kinds of combination strategies of the base classifiers including linear and non-linear fusion rules. We compare their performances with the regularized discriminant analysis (RDA) [2-4] as a benchmark method. RDA is one of the most powerful statistical pattern classification methods.

To test our theoretical suggestions, we considered important task of recognition of handwritten Japanese characters. In handwritten Japanese character recognition, some of the classes can be easily discriminated. However, there are many very similar classes, and recognition of such similar classes is important but difficult problem. To improve this situation, our concern is to study most ambiguous pairs of pattern classes. For illustration, eight pairs of similar Japanese characters are shown in Fig. 1.

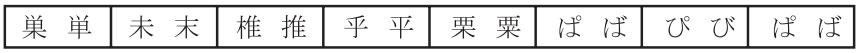


Fig. 1. Eight pairs of similar Japanese characters.

2. Sample Size/Complexity Properties

The standard Fisher discriminant function (we call "discriminant function" DF in short) is one of the most popular decision rules. Let $\bar{\boldsymbol{x}}^{(1)}$ and $\bar{\boldsymbol{x}}^{(2)}$ be sample mean vectors, and **S** be pooled sample covariance matrix. Allocation of a p-variate vector $\boldsymbol{x}=\left(x_1,\cdots,x_p\right)^T$ is performed according to sign of DF [2]

$$g(\boldsymbol{x})=\left(\boldsymbol{x}-\frac{1}{2}\left(\bar{\boldsymbol{x}}^{(1)}+\bar{\boldsymbol{x}}^{(2)}\right)\right)^T \mathbf{S}^{-1}\left(\bar{\boldsymbol{x}}^{(1)}-\bar{\boldsymbol{x}}^{(2)}\right). \qquad (1)$$

Let N be the number of training sets which are used to obtain estimates $\bar{\boldsymbol{x}}^{(1)}$, $\bar{\boldsymbol{x}}^{(2)}$ and **S**. Asymptotic classification error of sample based DF is given as $P_{\mathrm{B}}=\Phi\left(-\frac{1}{2}\delta\right)$, where δ stands for Mahalanobis distance and $\Phi(\cdot)$ is cumulative distribution function of $\mathrm{N}(0,1)$ (see, e.g., [2]). As both sample size N and dimensionality p increase, distribution of sample based DF approaches Gaussian law. After calculation of conditional means and common variance of discriminant function (1), one can find expected probability of misclassification,

$$\mathrm{EP}_N=\Phi\left(-\frac{1}{2}\delta\left(\left(1+\frac{2p}{N\delta^2}\right)\frac{2N}{2N-p}\right)^{-1/2}\right) \qquad (2)$$

[3, 5]. In Eq. (2), term $2N/(2N-p)$ arises due to inexact estimation of covariance matrix, and term $1+2p/N\delta^2$ arises due to inexact estimation of mean vectors. Equation (1) will be Euclidean distance classifier (EDC) if covariance matrix **S** is ignored. EDC has relatively good small sample properties. Similarly, if covariance matrix **S** is described by small number of parameters, DF with better small sample properties could be obtained. An example is the first order decision tree model described in Sect. 4 (see, e.g., [3, 6]). An alternative way to improve small sample properties is RDA. Covariance matrix of RDA is given as $\mathbf{S}_{\mathrm{RDA}}=(1-\lambda)\mathbf{S}+\lambda\mathbf{I}$, where λ is positive constant defined in an interval [0 1]. Optimal value of λ, denoted by λ_{opt}, have to be chosen by taking into account the balance between complexity of pattern recognition task (structure of the true covariance matrix **Σ**) and sample size *N*.

In our investigations, sample size N=100 and the original dimensionality p=196. Suppose Bayes error $P_{\mathrm{B}}=0.1$ ($\delta=2.56$). Then $1+2p/N\delta^2\approx 1.6$ and $2N/(2N-p)\approx 800$. High values of these coefficients indicate that we work in serious deficit of training data. One way to improve the data deficit problem is to reduce dimensionality, i.e. perform feature selection. Another way is to use simpler estimate of covariance matrix. We will use both of them. They are described in Sect. 3 and 4.

3 Representations of Feature Vectors

3.1 The Original Feature Vector

In this paper, 196-dimensional directional element feature [7] was used to represent handwritten Japanese characters in database ETL9B [8]. Preliminary to extracting the feature vector, a character image was normalized nonlinearly [9] to fit in a 64×64 box. Then, skeleton were extracted, and line segments of vertical, horizontal and slanted at ±45 degrees were extracted. An image is divided into 49 sub-areas of 16×16 dots. Sum of each segment in a region is an element of feature vector.

3.2 Three Representations of Feature Vector

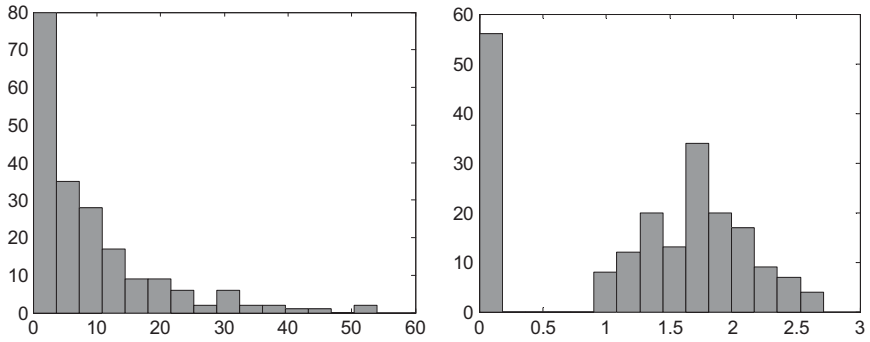
In constructing three base classifiers for multiple classifier system, we performed three kinds of transformations of the feature vector:

A) original (without transformation),

B) transformed by $y_r=x_r^{1/s}$ for r-th element of the feature (r = 1, …, 196, s is arbitrary) and

C) binarized (0 or 1) (non-zero valued components of the feature vector were equalized to 1).

We comment the reasons of using feature B and C. In Fig. 2a, we have a histogram of an element of the feature vector, which corresponds to the sub-area at a boundary of an image. The distribution density is highly asymmetric. It is well known that estimation of covariance matrix requires Gaussian density [10], and nonlinear transformations such as transformation (B) often helps reveal correlation structure of the data better. Thus, the histogram of nonlinearly transformed by $y_r = x_r^{1/4}$ is performed (Fig. 2b). We notice that the distribution of single feature is obviously bimodal and one peak is at zero. One possible way to tackle bimodality problem is to ignore "outliers" (in our case "zero valued features"). As mentioned in Sect. 3.3, our dimensionality reduction strategy has similar effect to this for feature B. However, the deletion may cause loss of information. Therefore, the third expert classifier utilized binary vectors.



(a) Original measurement. (b) After transformation by $y_r = x_r^{1/4}$.

Fig. 2. Histograms of distribution of 5th feature, x_5.

3.3 Dimensionality Reduction

When the number of training vectors is unlimitedly large, local pattern recognition algorithms (such as Parzen window classifier and *k*-NN rule) could lead to minimal (Bayes) classification error if properly used. Unfortunately, in practice, the number of training vectors is limited, and the dimensionality of feature vector is usually high. Therefore, one needs utilize prior information available to build the classification rule. An optimal balance between complexity and training sample size has to be retained. If sample size is not very large, one has to restrict complexity of base classifiers. In small sample case, optimistically biased resubstitution error estimates of the base classifiers supplied to fusion rule designer could ruin performance of MCS [11].

When we have notably smaller coefficient in Eq. (2), i.e. for dimensionality $p^* = 20,\ \ 1 + 2p^* \big/ N\delta^2 \approx 1.06.$ In order to improve small sample properties of base classifiers, for each kind of features transformation, we selected only twenty "better" features. Selection was performed on bases of sample Mahalanobis distances of each original feature. Since sample size was relatively

small, we could not use complex feature selection strategy. Here, the written character occupies only a part of 7×7 area. Like in many of similar character recognition problem, some of the sub-areas are almost "empty". Thus, our feature selection strategy is: at first, the *r*-th elements of 196-dimensional feature vectors were divided by their standard deviation of each class s_r (r = 1, 2, …, 196); then, twenty features (dimensions) whose 1-dimensional sample means of two classes are more distant were selected. The experiments showed that this feature selection also had important secondary effect: many non-informative features which contain a large number of outliers (zero valued measurements) were discarded.

4 Three Base Classifiers

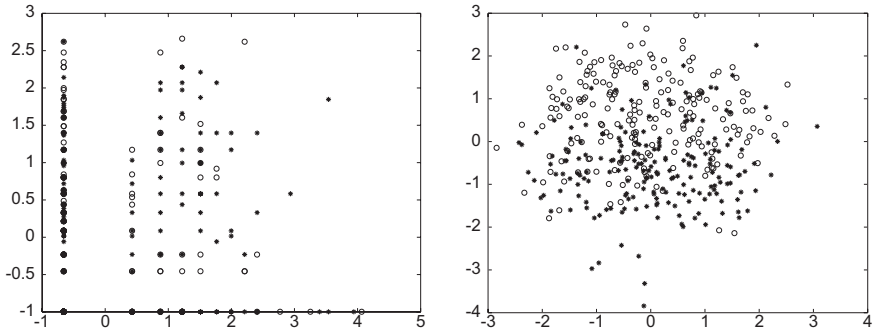
Three base classifiers were built in the three feature spaces (A, B, C) respectively. To construct robust base classifiers, we utilized the integrated approach of statistical and neural network [3]. This approach consists of data whitening and training of SLP. This can offer better linear DF by taking advantage of both statistical methods and neural networks.

In the data whitening, *one moves data mean vectors to the origin of coordinates*, and then performs *data whitening transformation* by use of $\boldsymbol{y} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Phi}^T\left(\boldsymbol{x} - 1/2\left(\overline{\boldsymbol{x}}^{(1)} + \overline{\boldsymbol{x}}^{(2)}\right)\right)$, where $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ are the matrices of eigenvalues and eigenvectors of simplified covariance matrix $\mathbf{S}_{\text{RDA\&Tree1}}$. We used regularization and the first order tree dependence model [6] for the simplified covariance matrix $\mathbf{S}_{\text{RDA\&Tree1}} = \mathbf{S}_{\text{Tree1}}\left(1-\lambda_{\text{opt}}\right) + \lambda_{\text{opt}}\mathbf{I}$, where $\mathbf{S}_{\text{Tree1}}$ is covariance matrix of the first order tree dependence model described only by 2*p*-1 independent parameters. This simplification makes the estimate of the covariance matrix more reliable in small sample case. Thus, in data whitening, we perform data rotation by means of orthogonal matrix $\boldsymbol{\Phi}$ and variance normalization by multiplying rotated data by matrix $\boldsymbol{\Lambda}^{-1/2}$. This transformation has a secondary effect which have not been discussed in the robust statistics literature. Linear transformation of multidimensional data produces weighted sums of the original features (see Fig. 3). For this reason, the distribution densities of the new features are closer to univariate and unimodal, and often resemble Gaussian distribution. In our experiments we noticed that time and again in whitened feature space, the first components give good separation of the data.

After data whitening, SLP was trained in space of $\boldsymbol{y}$. The training of SLP started with zero valued weight vector. After the first batch iteration, we obtain DF (1) whose $\mathbf{S}$ is replaced by $\mathbf{S}_{\text{RDA\&Tree1}}$. If assumptions about structure of covariance matrix are truthful, estimate $\mathbf{S}_{\text{RDA\&Tree1}}$ helps to have quite good DF with relatively small error rate and good small sample properties just at the very beginning.

If starting regularization parameter λ_{opt} is suitable, proper stopping could help to obtain the classifier of optimal complexity. To determine optimal number of batch iterations (epochs) to train SLP, we utilized independent pseudo-validation data sets with colored noise injection [12]. The pseudo-validation sets were formed

by adding many (say n) randomly generated zero mean vectors to each training pattern vector. The detail is as follows. For each vector $\boldsymbol{x}_i$, its k nearest neighbors $\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{ik}$ are found in the same pattern class; then, k lines which connect $\boldsymbol{x}_i$ and $\boldsymbol{x}_{iq}$, (q=1, 2,…, k) are prepared; along the q-th line, one adds random variables which follow Gaussian distribution $\mathrm{N}\left(0,\left(\sigma \mid \boldsymbol{x}_i - \boldsymbol{x}_{iq} \mid\right)^2\right)$; after adding k components, a new artificial vector is obtained. This procedure is repeated n times. Three parameters have to be defined to realize a noise injection procedure. In our experiments, we used: k = 2, n = 10, and σ = 1. In fact, noise injection introduces *additional non-formal information*: it declares in an inexplicit way that the space between nearest vectors of one pattern class could be filled with vectors of the same category (for more details see [11, 12]).



(a) Original feature space. (b) Whitened feature space.

Fig. 3. Effect of whitening transformation; "*" and "o" stand for feature vectors of two classes respectively, which were transformed by $y_r = x_r^{1/2}$.

Second expert, B, was working in transformed by $y_r = x_r^{1/s}$ feature space where bimodality of the data was clearly visible. For robust estimation, an influence of outliers was reduced purposefully. We ignored measurements with zero feature values for *robust estimation of mean vectors and covariance matrix*. While estimating the mean values and variance of j-th feature, we rejected zero valued training observations. To estimate ρ_{ij}, a correlation coefficient between i-th and j-th elements of feature vector, we utilized training vectors with only nonzero i-th and j-th components.

5 Experiments with handwritten Japanese Characters

5.1 Fusion Rules

We utilize a number of different fusion rules and compare classification performances of MCSs with RDA used as a benchmark method. From an

abundance of known fusion rules (see, e.g., [13]), eight linear and non-linear rules below were considered to make final decision.

BestT) The best (single) base classifier is selected according to classification results using the test set. Actually, this is the ideal classifier which achieves the minimum error rates in use of the three base classifiers. This classifier and BestV (the next item) are weighted voting MCSs that only one weight is unequal to zero.

BestV) The best (single) base classifier is selected according to classification results using the pseudo-validation set. This classifier was used as a benchmark MCS.

MajV) Majority voting. This is a fixed (non-trainable) fusion rule.

WStv) Weighed sum of the outputs of the base classifiers. SLP was used as fusion rule. The original training data set was used to train SLP classifier and produce coefficients of weighted sum. Optimal stopping was performed according to classification error estimated from pseudo-validation set.

BKS) The original behavior knowledge space method (see, e.g., [3, 14]). Allocation is performed according to probabilities $P_{11},\ldots,P_{18}$, $P_{21},\ldots,P_{28}$ of eight combinations of binary outputs of three base classifiers; training set were used to estimate the probabilities mentioned.

BKSn) Modified BKS method aimed to reduce expert adaptation to training data [11]. An independent pseudo-validation set was used to estimate $P_{11},\ldots,P_{28}$.

G&R) Nonlinear classification in 3D space of the three outputs of the base classifiers. To make final allocation, the Parzen window classifier was utilized (this MCS utilizes of expert outputs. Its decision making procedure resembles that of Giacinto and Roli [15], therefore, it is marked by G&R).

R&E) Nonlinear fusion of the outputs of the base classifiers where a sample-based oracle uses the input vector $\boldsymbol{x}$ in order to decide which expert is the best competent to classify this particular vector $\boldsymbol{x}$. The fusion rule allocates vector $\boldsymbol{x}$ to one of three virtual pattern classes (experts). The competence of the j-th expert is estimated as a "potential"

$$\hat{p}_j(\boldsymbol{x}) = \sum_{s=1}^{K}\sum_{l=1}^{N_s} q_{jl}^{s} \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{x}_l^{(s)})^T(\boldsymbol{x}-\boldsymbol{x}_l^{(s)})}{h^2}\right\}, \quad (j=1,2,3), \qquad (\mathbf{3})$$

where $q_{jl}^{s}=1$ if l-th training vector of s-th class, $\boldsymbol{x}_l^{(s)}$, was classified by the j-th expert correctly, and $q_{jl}^{s}=-1$ if vector $\boldsymbol{x}_l^{(s)}$ was classified incorrectly, $\exp\{\cdot\}$ is a *kernel* and h is a smoothing constant. This approach corresponds to Rastrigin and Erenstein [16] fusion rule introduced three decades ago. We marked it by R&E.

RDA) RDA is one of the most powerful pattern classification tools, and was described in Sect. 4.

5.2 Experiment

Each pair of handwritten character contains two similar classes. Each class consists of 200 vectors. 100 vectors were randomly selected as the training set (N=100),

and remaining 100 vectors were the test set. To reduce an influence of randomness, the experiments were performed 100 times for each pair of characters. Every time, random permutation of vectors was performed in each category.

For each data representation, individual feature selection and subsequent data rotation and normalization procedures were performed. After few preliminary experiments following parameters were determined: $p^* = 20$, $\lambda_{\text{opt}} = 0.2$, and $s = 2$.

In all experiments, only training set and its "product", artificial pseudo-validation set described in Sect. 4, were used to design decision making rules. The SLPs which were used as the base classifiers were trained on training set. Optimal number of iterations was determined by the recognition results of pseudo-validation set. While building some of trainable fusion rules, we interchanged training and pseudo-validations sets: the fusion rules were trained on validation set, and optimal number of iterations was found according to error rates of the training set. The test set was used only once, for final evaluation of generalization errors.

Results obtained in 800 training sessions (100 independent experiments with eight pairs of similar handwritten Japanese characters) are summarized in Table 1. For every pair, averaged test error rates of three experts (1st E, 2nd E and 3rd E), BestT and BestV are presented in the left five columns of Table 1 (next to the index of character pair). Let $\bar{\hat{P}}_{\text{BestV}}$ be the averaged test error rate of BestV (printed in bold in the table) and $P_{\text{Method\_A}}$ be that of "Method A". Further, the ratio of the averaged test error rate of "Method A" (corresponding to remaining 6 fusion rules and RDA) to that of BestV are shown in the right seven columns of Table 1. Namely, the relative error rate is given as $P_{\text{Method\_A}} \big/ \bar{\hat{P}}_{\text{BestV}}$. The very last row in the table contains averaged values of eight cells of the column.

Table 1. Average error rates of single experts, BestT and BestV are in the left five columns next to index, and relative efficacy of six fusion procedures and RDA are in the seven columns in the right. Relative error rates of the most effective fusion rules are underlined.

PAIR	1st E	2nd E	3rd E	BestT	BestV	MVot	WSum	BKS	BKSn	G&R	R&E	RDA
A	0.037	0.037	0.042	0.035	**0.042**	0.882	0.899	1.022	0.930	1.239	0.911	1.349
B	0.129	0.116	0.198	0.112	**0.122**	0.946	1.012	1.039	0.946	1.090	1.008	1.385
C	0.056	0.070	0.159	0.054	**0.066**	1.002	0.964	0.923	1.002	0.955	0.841	1.115
D	0.138	0.139	0.154	0.132	**0.144**	0.950	0.981	1.051	1.018	1.042	0.972	1.453
E	0.087	0.083	0.133	0.079	**0.103**	0.805	0.810	0.865	0.823	0.819	0.842	1.261
F	0.135	0.130	0.190	0.125	**0.138**	0.920	0.962	0.996	0.921	0.986	0.972	1.575
G	0.086	0.088	0.100	0.081	**0.091**	0.940	0.940	0.948	0.941	0.955	0.968	1.675
H	0.120	0.119	0.149	0.112	**0.125**	0.899	0.923	0.960	0.899	0.943	0.967	1.410
ALL	0.098	0.098	0.141	0.091	**0.104**	0.918	0.936	0.976	0.935	1.004	0.935	1.403

In spite of apparent similarity of eight kinds of Japanese character pairs, we have notable variations in experimental results obtained for diverse pairs: both separability of pattern classes (classification error rate) and relative efficacy of the experts differ by pairs.

Nevertheless, for all the pairs, RDA whose parameter λ is adjusted to complexity of the recognition problem and size of training set was outperformed by MCSs designed to deal with outliers and multimodality problems. By comparing RDA and MCS rules, the highest gain among MCS rules (1.675 times in comparison with BestV, and 1.78 times in comparison with Majority Voting) was obtained for pair **G** where all three experts were approximately equally qualified. The lowest gain (1.115 times in comparison with BestV, and 1.325 times in comparison with Rastrigin-Erenstein procedure) was obtained for pair **C** where the third expert was notably worse than two others.

The training set size of the current problem, 100+100 vectors in 196-variate, is rather small. Therefore, sophisticated trainable fusion rules were ineffective: for almost all eight Japanese character pairs considered, the fixed fusion rule, Majority Voting, was the best. Exception is pair **C** because of inefficiency of the third expert, that is, only two experts participated in final decision making. Detailed analysis shows that in general, all three experts are useful: rejection of one of them assists an increase in generalization error of MCS.

6 Concluding Remarks

In this paper, we considered problem of atypical observations in training set in high-dimensional situations where sample size is relatively small. Since Gaussian classifiers are not suitable for atypical observations, as a practical solution, we proposed multiple classifiers systems (MCSs) whose base classifiers have different data representations respectively. The base classifiers are constructed with the integrated approach of statistical and neural networks.

To test the proposed MCSs, we considered recognition task of similar pairs of handwritten Japanese characters. For all eight similar Japanese character pairs considered, all the proposed MCSs outperformed the benchmark classification method, the RDA, in the situation of small sample and high-dimensional problem. Utilization of MCSs with base classifiers working in differently transformed feature space contains supplementary information that nonlinear transformations are important in revealing atypical observations. Dealing with the outlier problem, dissimilarity of features allowed the MCSs to reduce generalization error.

With a simple feature selection procedure, all three base classifiers worked in reduced feature spaces. The feature selection procedure utilizes additional information: a part of the features are notably less important for the linear classification rules designed to operate with unimodal distributions. Analysis of histograms of rejected features showed that often rejected features had bimodal distribution density functions, i.e. substantial part of data contained zero-valued measurements. This means that our feature selection lightened outliers and multimodality problem.

Training sample size used to train the experts and the trainable fusion rules of MCSs is too small for given high-dimensional pattern recognition problem. Therefore, fixed fusion rule performed the best. No doubt that in situations with larger number of samples, more sophisticated fusion rules would be preferable and could lead higher gain in dealing with outlier and multimodality problems.

Acknowledgments

The authors thank Assoc. Prof. Shinichiro Omachi, Prof. Hirotomo Aso, Dr. Ausra Saudargiene and Giedrius Misiukas for useful discussions, shared with us data sets and Matlab codes.

References

1. S. Raudys. and A.K. Jain. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-13:252–64, 1991.
2. R.O. Duda, P.E. Hart and D.G. Stork. *Pattern Classification*. 2nd ed. Wiley, NY, 2000.
3. S. Raudys. *Statistical and Neural Classifiers: An integrated approach to design*. Springer-Verlag, NY, p. 312, 2001.
4. J.M. Friedman. Regularized discriminant analysis. *Journal of American Statistical Association* 84:165–75, 1989.
5. S. Raudys. On the amount of a priori information in designing the classification algorithm. Proc. Ac. Sci. USSR, Ser. *Engineering Cybernetics*, N4:168–74, 1972 (in Russian).
6. S. Raudys and A. Saudargiene. Tree type dependency model and sample size - dimensionality properties. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 23(2):233-239, 2001.
7. N. Sun, Y. Uchiyama, H. Ichimura, H. Aso and M. Kimura. Intelligent recognition of characters using associative matching technique. *Pacific Rim Int'l Conf. Artificial Intelligence (PRICAI'90)*, 546-551, 1990.
8. T. Saito, H. Yamada and K. Yamamoto. On the data base ETL9 of handprinted characters in JIS Chinese characters and its analysis. *Trans. IEICE*, J68-D(4):757-764, 1985 (in Japanese).
9. H. Yamada, K. Yamamoto and T. Saito. A nonlinear normalization method for handprinted Kanji character recognition: line density equalization. *Pattern Recognition*, 23(9):1023-1029, 1990.
10. H. Ujiie, S. Omachi, and H. Aso. A Discriminant Function Considering Normality Improvement of the Distribution. *16th International Conference on Pattern Recognition (ICPR 2002)*, 2:224-227, 2002.
11. S. Raudys. Experts' boasting in trainable fusion rules. *IEEE Trans. on Pattern Analysis and Machine Intelligence*. PAMI 25(9):1178-1182, 2003.
12. M. Skurichina, Š. Raudys and R.P.W. Duin. K-nearest neighbors directed noise injection in multilayer perceptron training, *IEEE Trans. on Neural Networks*, 11(2):504–511, 2000.
13. A.F.R. Rahman and M.C. Fairhurst. Multiple classifier decision combination strategies for character recognition: A review. *Int.J.Document Analysis and Recognition*, 5(4):166–194, 2003.
14. Y.S. Huang and C.Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(1):90-94, 1995.
15. G. Giacinto and F. Roli. Dynamic classifier selection based on multiple classifier behaviour. *Pattern Recognition*, 34(9):1879-1881, 2001.
16. L.A. Rastrigin and R.Ch Erenstein. On decision making in a collective of decision rules. *Priborostroenie* , LITMO, St-Petersburg. 16(11):31-35, 1973 (in Russian).