2009 10th International Conference on Document Analysis and Recognition

A Novel Rejection Measurement in Handwritten Numeral Recognition Based on Linear Discriminant Analysis

Chun Lei He Louisa Lam Ching Y. Suen
CENPARMI (Centre for Pattern Recognition and Machine Intelligence)
Computer Science and Software Engineering Department, Concordia University
Montreal, Quebec, H3G 1M8, Canada
Emails: {cl_he, llam, suen} @ cenparmi.concordia.ca

Abstract

This paper presents a Linear Discriminant Analysis based Measurement (LDAM) on the output from classifiers as a criterion to reject the patterns which cannot be classified with high reliability. This is important in applications (such as in processing of financial documents) where errors can be very costly and therefore less tolerable than rejections. To implement the rejection, which can be considered to be a two-class problem of accepting the classification result or otherwise, Linear Discriminant Analysis (LDA) is used to determine the rejection threshold at a new approach. LDAM is designed to take into consideration the confidence values of the classifier outputs & the relations between them, and it is an improvement over traditional rejection measurements such as First Rank Measurement (FRM) and First Two Ranks Measurement (FTRM). Experiments are conducted on the CENPARMI Arabic Isolated Numerals Database. The results show that LDAM is more effective, and it can achieve a higher reliability while achieving a high recognition rate.

1. Introduction

In Optical Character Recognition (OCR) application, current research methods have achieved recognition rates higher than 99% on some Latin numeral databases such as MNIST [1, 2] and CENPARMI Database [3]. However, even these low error rates can be costly in some applications. It is the general expectation that OCR machines should achieve a high recognition rate as well as high reliability, which is defined by:

$$Rejection\ rate = \frac{Number\ of\ rejected\ samples}{Total\ number\ of\ test\ samples} \times 100\%$$

$$Reliability = \frac{Recognition\ rate}{100\% - Rejection\ rate} \times 100\%$$

Achieving high recognition and reliability requires methods capable of assigning generally higher confidences to correct recognition results than to incorrect ones. This confidence scoring method may consist of implementing a simple function of appropriate parameters drawn directly from the recognition process, or it may be a learning task in which a classifier is trained to use an array of parameters to distinguish correct recognitions from misclassifications [4].

It seems that the Bayes decision rule embodies a rejection rule, namely, the decision can be based on the maximum confidence value provided this maximum exceeds a certain threshold value. However, this approach did not perform satisfactorily when experiments were performed on the CENPARMI Arabic Isolated Numerals Database. The distribution of incorrectly classified samples is not Gaussian in shape, but remains flat throughout a range of confidence values. This is the case while the correctly classified samples do follow a Gaussian distribution. The results on the training set are shown in Fig. 1.

In this paper, we modify the LDA method [5] and apply it to the measurement level outputs so that samples with low confidence can also have a Gaussian distribution separate from that of the correctly classified data. LDA is a supervised classification method widely used to find the linear combination of features for separating two or more classes. The main idea of LDA is to project high-dimensional data onto a line and perform classification in this one-dimensional space. It provides a linear projection of the data with the outcome of maximum between-class variance and minimum within-class variance. By finding the feature

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.89

space that can best discriminate an object from others, these discriminative methods have been successfully used in pattern classification applications including Chinese character recognition [6], face recognition [7], etc.

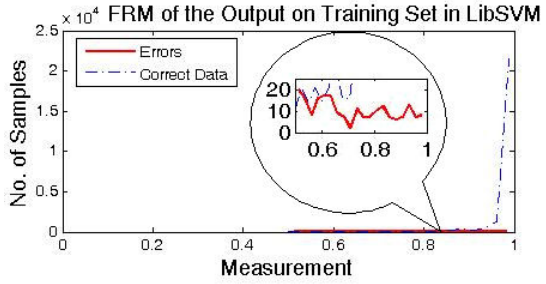

Figure 1. Distribution of the output on the Training set

Offline handwritten character recognition of languages such as English, Chinese, and Japanese has been researched extensively for over twenty years. However, Arabic handwriting recognition research has started only in recent years even though Arabic is one of the most widely spoken languages in the world [8]. Currently, Alamri et al. [9] designed the CENPARMI Arabic database which includes isolated Indian/Arabic numerals, numeral strings, Arabic isolated letters, and Arabic words. Experiments reported in this paper are conducted on the isolated numerals in this database.

Detecting samples recognized with low confidence for rejection and thus achieving a high reliability in handwritten numeral recognition is our objective in this research. We define a novel rejection measurement – LDA Measurement (LDAM), and we compare it to other rejection measurements, such as First Rank Measurement (FRM) and First Two Ranks Measurement (FTRM) in Section 2. Then, we describe the experiments on the CENPARMI Arabic Isolated Numerals Database and compare the results obtained from using the three measurements (Section 3). Finally, in Section 4, we conclude that the LDAM is more effective than the other two measurements.

2. Rejection Measurements

In considering the outputs of classifiers for the rejection option as a two-class problem, (acceptable or rejected classification), the outputs at the measurement level can be considered as features for the rejection option. In an output vector, whose components may represent distances or probabilities, we expect the confidence value (measure) of the first rank (most likely class) to be far distant from the confidence values or measures of the other classes. In other words, good outputs should be easily separated into two classes: the confidence value of the first rank and others. In the following discussion, we assume that the classifier which outputs the probabilities of the patterns for each class would be analogous to the case when the classifier outputs the distances.

2.1. First Rank Measurement (FRM) & First Two Ranks Measurement (FTRM)

Generally, rejection strategies can be directly applied to the outputs with a probability estimation. In an M-class problem, suppose $P(x) = \{p_1(x), p_2(x),..., p_M(x)\}$ is the classification output vector of the given pattern x, with probabilities $p_i(x)$ in descending order. The decision function would be based on $\mathrm{sgn}(\Phi_1(x) - T_1)$, where T_1 is a threshold derived from the training data, and $\Phi_1(x) = p_1(x)$.

If $\Phi_1(x) \le T_1$, the classifier rejects the pattern and does not assign it to a class (it might instead be passed to a human operator). This has the consequence that on the remaining patterns, a lower error rate can be achieved. This method uses the First Rank Measurement (FRM) [10].

Under this method, the frequency distribution according to confidence values of samples in the training set is considered and the threshold T_1 is determined.

However, FRM cannot distinguish between reliable and unreliable patterns with the probability distribution of erroneous samples shown in Fig. 1.

To overcome this deficiency of FRM, we have designed First Two Rank Measurement (FTRM) [11], which uses the difference between the probabilities $p_1(x)$ and $p_2(x)$ of the first two ranks as a condition of rejection. In FTRM, the measurement function is $\Phi_2(x) = \| p_1(x) - p_2(x) \|$, where $\|.\|$ can be any distance measurement, and the decision function is based on $\mathrm{sgn}(\Phi_2(x) - T_2)$, where T_2 is a threshold derived from the training set.

However, FTRM cannot solve the problem in some cases. For example, if $\| p_1(x) - p_2(x) \|$ is relatively large compared to T_2, but the difference between $\| p_2(x) - p_3(x) \|$ is much larger, this pattern may still be accepted, when this pattern should have been rejected since the top two classes are close together in terms of relative distance.

2.2. LDA Measurement (LDAM)

To consider the relative difference between the measurements in the first two ranks and all other measurements, LDAM is defined and applied. Since rejection in classification can be considered as a two-class problem (acceptance or rejection), we apply LDA [5] for two classes, to implement rejection. LDA approaches the problem by assuming that the conditional probability density functions of the two classes are both normally distributed. There are $n = n_1 + n_2$ observations with d features in the training set, where $\{x_{1i}\}_{i=1}^{n_1}$ arise from class ω_1 and $\{x_{2i}\}_{i=1}^{n_2}$ arise from class ω_2. Gaussian-based discrimination assumes two normal distributions: $(x \mid \omega_1) \sim N(\mu_1, \Sigma_1)$ and $(x \mid \omega_2) \sim N(\mu_2, \Sigma_2)$. In LDA, the projection axis (discriminant vector) w for discriminating between two classes is estimated to maximize the Fisher criterion:

$$J(w) = tr((w^T S_w w)^{-1}(w^T S_B w))$$

where tr(·) denotes the trace of a matrix, $\mathbf{S}_B$ and $\mathbf{S}_w$ denote the between-class scatter matrix and within-class scatter matrix respectively, and w is the optimal discriminant vector. For the two classes ω_1 and ω_2, with a priori probabilities p_1 and p_2 (it is often assumed that $p_1 = p_2 = 0.5$), the within-class and between-class scatter matrices can be written as

$$S_w = p_1\Sigma_1 + p_2\Sigma_2 = \Sigma_{12},$$
$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T,$$

where Σ_{12} is the average variance of the two classes. It can be shown that the maximum separation occurs when:

$$w = S_w^{-1}(\mu_1 - \mu_2) = (p_1\Sigma_1 + p_2\Sigma_2)^{-1}(\mu_1 - \mu_2) = \Sigma_{12}^{-1}(\mu_1 - \mu_2).$$

To use this principle to the outputs for the rejection option as a one-dimensional application, we define the two sets $G^{(1)}(x) = \{p_1(x)\}$, and $G^{(2)}(x) = \{p_2(x), p_3(x), ..., p_M(x)\}$. Then,

$$\mu_1 = p_1(x),$$
$$\mu_2 = \frac{1}{M-1}\sum_{i=2}^{M} p_i(x),$$
$$\Sigma_1 = (p_1(x) - \mu_1)^2 = 0,$$
$$\Sigma_2 = \frac{1}{M-1}\sum_{i=2}^{M}(p_i(x) - \mu_2)^2, \text{ and } \Sigma_{12} = \tfrac{1}{2}\Sigma_2.$$

Thus, in LDA,

$$w = \Phi_3(x) = \frac{\sum_{i=2}^{M} \| p_1(x) - p_i(x) \|}{(M-1)\cdot\Sigma_{12}}.$$

Then the decision function would be based on $\mathrm{sgn}(\Phi_3(x) - T_3)$, where T_3 is a threshold derived from the training set, and all values are scaled to [0, 1].

In summary, when compared to FRM and FTRM, LDAM should be more reliable and informative since it compares the relative difference of the measures in the first two ranks with all other measures.

3. Experiments and Results

The CENPARMI Arabic Isolated Numerals Database is used for the experiments. It contains 18,585, 6,199, and 6,199 samples in the Training, Validation, and Test sets, respectively. Since validation was not implemented in this experiment, the Training and Validation sets were combined to be the Training set. There are 10 classes (0-9), and five samples of each numeral are shown in Figure 2.

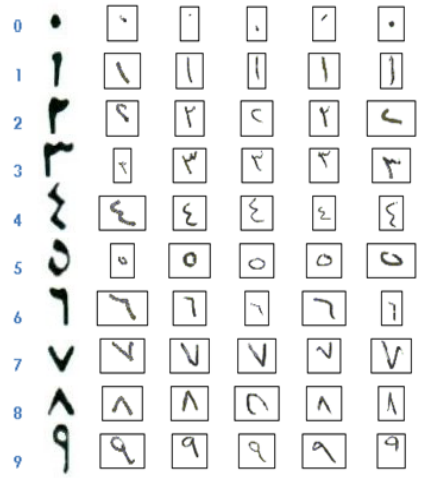

Figure 2. Samples from CENPARMI Arabic Isolated Numerals Database

In the recognition process, the standard procedures of image pre-processing, feature extraction, and classification were implemented. In image pre-processing, we perform noise removal, grayscale normalization, size normalization, and binarization of the grayscale images. Gradient features were extracted from pseudo gray-scale images [12]. The Robert filter, which uses the masks $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$, was applied to calculate the gradient strengths and directions of pixels. The direction of the gradient was quantized to 32 levels with an interval of $\pi/16$. The normalized character image was divided into 81 (9×9) blocks. After

extracting the strengths and directions in each image, the spatial resolution was reduced from 9×9 to 5×5 by down sampling every two horizontal and every two vertical blocks with a 5×5 Gaussian filter. Similarly, the directional resolution was reduced from 32 to 16 levels by down sampling with a weight vector $[1\,4\,6\,4\,1]^T$, to produce a feature vector of size 400 (5×5×16). Moreover, the transformation $y = x^{0.4}$ was applied to make the distribution of the features Gaussian-like. The feature set size was reduced to 400 by principal component analysis (KL transform). Finally, we scaled the feature vectors by a constant factor so that the values of feature components range from 0 to 1.

Support Vector Machines [13] was chosen as a classifier. SVMs with different kernel functions can transform a non-linear separable problem into a linear separable one by projecting data into the feature space, and then SVMs can find the optimal separating hyperplane. Radial Basis Function (RBF) was chosen as the kernel in this research. The recognition rate on the test set is 98.48% (Table 1), which is significantly higher than the performance (93.60%) in [9] on the same database, and the number of errors was 94 (1.52%), most of which are also unrecognizable by human beings.

Table 1. Performance with different thresholds in LDAM compared with [9]

Threshold	0	0.05	0.35	[9]
Recog. Rate (%)	98.48	92.40	90.06	93.60
Error Rate (%)	1.53	0.27	0.11	6.40
Reliability (%)	98.48	99.70	99.87	93.60

The experimental results on the test set are shown in Figure 3. The solid lines represent the distributions of errors, and the dotted lines represent the distribution of correctly recognized samples. The distributions based on FRM are shown in Figure 3(a). It is similar to the distributions of the Training data. Although the correct samples display a Gaussian distribution, the errors are distributed almost evenly for confidence values (measurements) ranging from 0.4 to 1, so the graph is too flat to separate correctly and incorrectly classified samples based on FRM. When compared to FRM, FTRM is more discriminating, as the range of measurements in FTRM is wider than FRM. However, the distribution of errors in FTRM is flat as well. LDAM is more discriminating than FRM and FTRM. This is because the errors plus correctly classified samples with low confidence values are assigned small measurements. In LDAM, most incorrectly classified samples (78/94) retain very low measurements (less than 0.05). Thus, LDAM enables the rejection of samples with low reliability with the thresholds obtained from the training set.

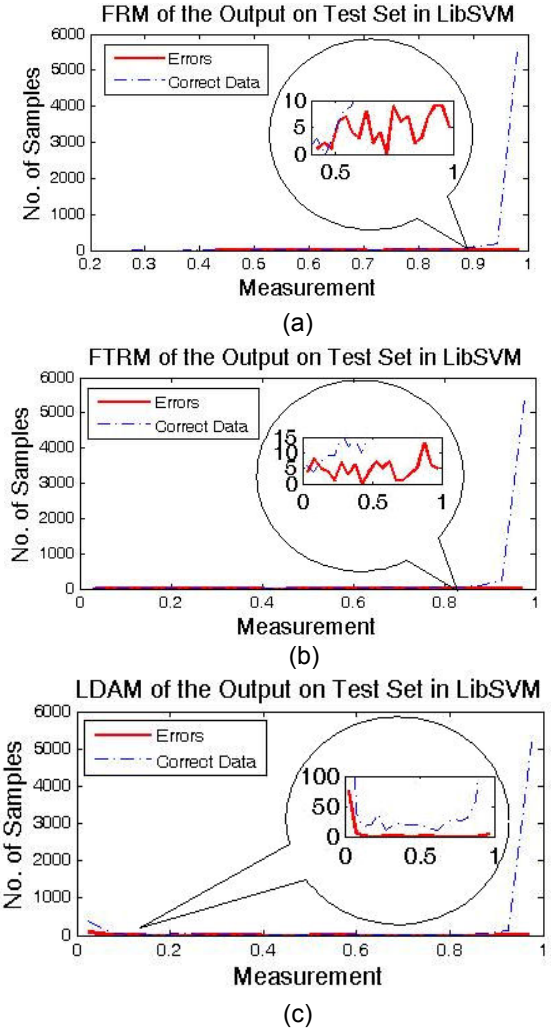

Figure 3. Distributions of the three measurements on the Test Set:

(a) FRM (b) FTRM (c) LDAM

The performances using different thresholds on the three measurements are shown in Figure 4. As illustrated, when the threshold T_3 is set at 0.05, the reliability increases from 98.48% to 99.70% with LDAM, while the reliabilities with FTRM and FRM are 98.52% and 98.48% respectively. It shows that LDAM is more effective in increasing reliability.

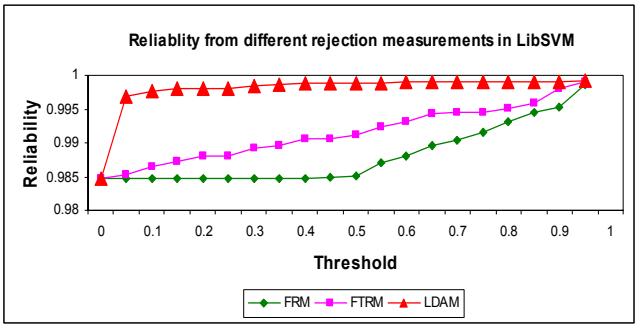

Figure 4. Reliability with different thresholds used on the three measurements

When the reliability of LDAM is 99.70%, there are 16 errors, all of which are shown in Figure 5. As can be seen, these errors are reasonable since even human beings would find it difficult to recognize these samples, or to distinguish between samples of "2" and "3" written in the same styles in Arabic.

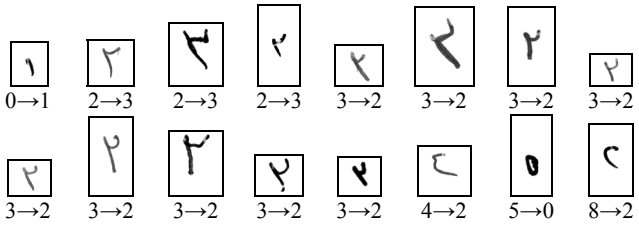

Figure 5. Incorrectly classified images

This method has also been trained and tested on the MNIST database, and the resulting reliabilities display a pattern very similar to that shown in Figure 4. In particular, when the threshold is set at 0.1, reliabilities of 98.97%, 99.09% and 99.88% are achieved on the test set for FRM, FTRM and LDAM, with recognition rates of 98.97%, 98.84% and 96.16% respectively.

4. Conclusion

The rejection option is very useful in preventing misclassifications, especially in applications which require a high reliability. We designed a novel rejection criterion using the LDA Measurement (LDAM), which relies on the principle of LDA and considers relationships among the probabilities in each output vector. We trained this rejection measurement on the training set and tested it on the test set of different databases. At the same time, we compared LDAM with other measurements such as FRM and FTRM. The results indicate that LDAM achieved a higher reliability than the other measurements when a small threshold was set.

In the future, we can apply this rejection method to train multi-classifier systems. Moreover, we can also apply this methodology to semi-supervised learning, so that we could reject the data with unreliable classification results produced by supervised learning.

5. References

[1] F. Lauer, C. Y. Suen, and G. Bloch, "A trainable feature extractor for handwritten digit recognition," *Pattern Recognition*, 40(6), 2007, pp. 1816–1824.

[2] P. Zhang, T. D. Bui, and C. Y. Suen, "A novel cascade ensemble classifier system with a high recognition performance on handwritten digits," *Pattern Recognition*, 40(12), 2006, pp. 3415–3429.

[3] C. L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: Investigation of normalization and feature extraction techniques," *Pattern Recognition*, 37(2), 2004, pp. 265–279.

[4] J. F. Pitrelli and M. P. Perrone, "Confidence-scoring post-processing for off-line handwritten-character recognition verification," *Proc. of 7th Int. Conference on Document Analysis and Recognition (ICDAR'03)*, I, 2003, pp. 278–282.

[5] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7, 1936, pp. 179–188.

[6] T.-F. Gao and C.-L. Liu, "High accuracy handwritten Chinese character recognition using LDA-based compound distances," *Pattern Recognition*, 41, 2008, pp. 3442–3451.

[7] F. Tang and H. Tao, "Fast linear discriminant analysis using binary bases," *Pattern Recognition Letters*, 28 (16), 2007, pp. 2209–2218.

[8] A. El. Sagheer, N. Tsuruta, and R.-I. Taniguchi, "Arabic lip-reading system: A combination of Hypercolumn Neural Network Model with Hidden Markov Model," *Artificial Intelligence and Soft Computing ASC*, Marbella, Spain, 2004, 9.1–9.3.

[9] H. Alamri, J. Sadri, C. Y. Suen, and N. Nobile, "A Novel Comprehensive Database for Arabic Off-Line Handwriting Recognition," *Proc. of the 11th Int. Conference on Frontiers in Handwriting Recognition (ICFHR 2008)*, Montreal, Canada, 2008, pp. 664–669.

[10] J. X. Dong, A. Krzyzak, and C.Y. Suen, "Fast SVM training algorithm with decomposition on very large datasets," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(4), 2005, pp. 603–618.

[11] C. Y. Suen, C. Nadal, R. Legault, T. A. Mai, and L. Lam, "Computer recognition of unconstrained handwritten numerals," *Proc. IEEE*, 80(7), 1992, pp. 1162–1180.

[12] M. Shi, Y. Fujisawa, T. Wakabayashi, and F. Kimura, "Handwritten numeral recognition using gradient and curvature of gray scale image," *Pattern Recognition*, 35(10), 2002, pp. 2051–2059.

[13] V. Vapnik and A. Lerner, "Pattern recognition using generalized portrait method," *Automation and Remote Control*, 24, 1963, pp. 774–780.