2009 10th International Conference on Document Analysis and Recognition

Rejection Strategies with Multiple Classifiers for Handwritten Character Recognition

Xu-Cheng Yin [1] Hong-Wei Hao [1] Yun-Feng Tang [1] Jun Sun [2] Satoshi Naoi [2]
[1] Department of Computer Science, School of Information Engineering, University of Science and Technology Beijing, China
[2] Fujitsu R&D Center Co. Ltd.
xuchengyin@ies.ustb.edu.cn; hhw@ies.ustb.edu.cn; tang.yunfeng@hotmail.com; sunjun@cn.fujitsu.com; naoi@cn.fujitsu.com

Abstract

With rejection strategies in a handwriting recognition system, we are able to improve the reliability and accuracy of the recognized characters. In this paper, we propose several rejection strategies with multiple classifiers for handwritten character recognition. First, the rejection strategy for the single classifier is introduced, which is composed of three stages: initial scaling, confidence measure calculation, and rejection performing. Then, we analyze rejection strategies for multiple classifiers. We divided our rejection strategies into two categories: (1) for voting combination; and (2) for linear combination with multiple classifiers. In the voting combination style, three rejection strategies, OR, AND, and VOTING, are proposed. And for the linear combination one, rejection strategies for average and weighted combination are analyzed respectively. We also experiment and compare our rejection strategies with handwritten digit recognition.

1. Introduction

After about forty years of research, recognition of unconstrained offline handwritten characters is still a difficult problem. While high-accuracy character recognition has been achieved, in some applications even very few recognition errors are extremely costly. By implementing rejection strategies in a handwriting recognition system, we are able to improve the reliability of the rejected characters, and increase the accuracy of the remaining characters [1].

A common way to reject recognized characters is to compute a confidence measure for each processed character. For such an approach rejection strategies can be formulated as simple thresholding operations. A large number of confidence measures have been proposed in the literature [1, 3, 9, 13, 14]. There are two main categories for producing confidence measures. In the first category, confidence scoring may consist of a simple function of appropriate parameters drawn directly form the recognition process, or it may be considered a learning task in which a classifier is trained to use an array of such parameter to distinguish correct recognition from incorrect. In another category, without integrating confidence scoring with recognition, a post-processing approach to confidence is adopted, in which confidence is measured following recognition [14]. In the post-processing mode, most confidence measures for offline handwritten character recognition systems are recognition score, likelihood ratio, estimated posterior probability, exponential probability, and negative entropy [14]. Some more sophisticated methods are to use learning techniques and complicated architectures [16, 17].

Multiple classifier combination has been intensively studied with the aim of overcoming the limitations of individual classifiers [5, 6, 8, 12]. Classifiers differing in feature representation, architecture, learning algorithm, or training data exhibit complementary classification behavior and the fusion of their decisions can yield higher performance than the best individual classifier. Based on a given classifier set, the combination methods can be categorized according to the level of classifier outputs: abstract level (class label), rank level (rank order), and measurement level (class scores) [12, 18]. Yin et al. analyze multiple classifier systems from the perspective of feature combination [15]. And most popular systems are with multiple classifiers for character recognition [4, 12, 15].

We focus on rejection strategies with multiple classifier systems for abstract level and measurement level. For abstract level, rejection strategies with OR, AND and VOTING of voting combination are analyzed; and for measurement level, rejection strategies of linear combination: sum-rule (average) and weight combination are investigated. At both levels, the rejection strategy for a single classifier is the fundamental step. The constituent classifiers in multiple classifiers have different discriminant functions, which given measurements with diverse scales and physical meanings. The classifier outputs should be transformed to uniform measures that have similar scales. Preferably, the transformed measures represent the degree of confidence of decision, like the class posterior probability or likelihood [5, 11, 12].

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.45

Our proposed rejection strategies include strategies in a single classifier and strategies in multiple classifiers, which are both detailedly described in Section 2. Section 3 describes some experiments for handwritten digit recognition. Finally, some conclusions are drawn in Section 4.

2. Proposed rejection strategies

For handwritten character recognition with multiple classifiers, different classifiers have different characteristics. Confidence measures from different classifiers are not uniform. Consequently, before confidence measure calculation, classifier output scaling is necessary. Our technology, improved from the method of confidence transformation for multiple classifiers systems [5, 11], are suitable here. Moreover, we propose several rejection frameworks which are able to adaptively combine multiple classifiers.

2.1. Rejection strategies for the single classifier

Our rejection strategy for single classifier is composed of three stages (parts): (1) **initial scaling**, (2) **confidence measure calculation**, and (3) **rejection performing**. The stage of confidence measure calculation is similar to the one in [14]. In a post-processing mode, confidence measures (scores) include likelihood ratio, estimated posterior probability, negative entropy, etc., are calculated.

The scaling function shift and re-scales the classifier output to a moderate range such that the outputs of different classifiers are comparable. The re-scaled output is transformed to confidence measure to confidence measure using an activation function corresponding to one of three confidence types: log-likelihood, likelihood, and sigmoid. The scaling functions include global normalization, one-dimensional Gaussian density modeling, multivariate Gaussian density. The confidence types and scaling functions are briefly reviewed in the following, and more details can be found in [5, 11].

2.1.1. Initial scaling

An essential requirement to the scaling function is that the re-scaled classifier outputs distribute in a moderate range around 0. It is desired the transformed confidence measures represent the probability as of the input pattern to belong to a specific class.

To manage the range of classifier outputs, one simple strategy is to re-scale the output values to zero mean and standard deviation 1:

$$f_i(d) = \frac{d_i - \mu_0}{\sigma_0} \tag{1}$$

where μ_0 and σ_0^2 are the mean and variance of the pooled classifier outputs respectively. We refer to this scaling functions are Global Normalization.

The other two scaling functions are derived from Gaussian densities of classifier outputs. Assuming multivariate or one-dimensional Gaussian densities to the classifier outputs, the class probabilities are shown to be calculated from soft-max or sigmoid, from which we extract the scaling functions.

Assume for each class, the density of classifier outputs is a multivariate Gaussian with identity variance σ^2. Considering that the outputs of a strong classifier are well ordered such that the target class generally has high measure while other classes have low outputs, we assume that all classes share two distinct mean values, μ^+ for target class and μ^- for other classes such that for class ω_i , $m_{ii} = \mu^+$ and $m_{ij} = \mu^-$, $j \neq i$. With the mean μ^r of competing negative samples, we get one scaling function:

$$f_i(d) = \frac{\mu^+ - \mu^-}{\sigma^2}(d_i - \frac{\mu^+ + \mu^r}{2}) \cdot \tag{2}$$

The next scaling function is obtained by assuming one-dimensional Gaussian density to the output of each class,

$$f_i(d) = \alpha[d_i - (\beta + \gamma / \alpha)], \tag{3}$$

with $\alpha = \frac{\mu^+ - \mu^-}{\sigma^2}$, $\beta = \frac{\mu^+ + \mu^-}{2}$, $\gamma = \ln(P(\overline{\omega_i}) / P(\omega_i))$.

We set $P(\overline{\omega_i}) / P(\omega_i) = M$ considering that the negative samples may include those out of the *M* hypothesized classes and the *M*+1 classes are assumed to have equal prior probabilities.

2.1.2. Confidence measure calculation

As prevalently used in neural networks, the sigmoid function behaves well in squashing neuronal outputs to approximate probability measures. We take it as an activation function for confidence transformation:

$$g_i(d) = \frac{1}{1 + e^{-f_i(d)}} \tag{4}$$

In many parametric classifiers such as LDF and QDF, the class measurement is the logarithm or negative logarithm of Bayesian likelihood:

$$d_i(x) = \log[p(\omega_i) p(x \mid \omega_i)]$$

In this case, the class posterior probability can be calculated by soft-max as

$$p(\omega_i \mid x) = \frac{\exp[d_i(x)]}{\sum_{j=1}^{M} \exp[d_i(x)]}$$

We take the exponential before normalization to unity as a type of confidence

$$g_i(d) = e^{f_i(d)} \tag{5}$$

in which using re-scaled output instead of the raw output may give better confidence estimates.

The third type of confidence is the log-likelihood. When approximating the Bayesian likelihood using exponential, the log-likelihood is simply the linear form of the scaling function:

$$g_i(d) = f_i(d) \tag{6}$$

To give the class posterior probabilities that satisfy the axiom of probabilities, the exponential likelihood and the sigmoid measure are to be normalized:

$$p(\omega_i \mid d) = \frac{g_i(d)}{\sum_{j=1}^{M} g_j(d)}.$$

2.1.3. Rejection performing

After initial scaling (Equation (1), (3) and (4)) and confidence measure calculation (Equation (4), (5) and (6)), the last step is to perform rejection. The simple way can be formulated as thresholding operations.

Given an input sample *x*, the outputs of the recognition systems with *M* classes (after scaling and confidence calculation) are $\{g_{i_1}(x), g_{i_2}(x), ..., g_{i_M}(x)\}$, which are listed in a descending sort. Many researchers use the first result to do thresholding [13, 14], i.e.,

$$r_1(x) = g_{i_1}(x) < TH_1 \tag{7}$$

And we get $0 \le r_1(x) \le 1$. Some other researchers calculate the relative ratio of the first two results [17] for rejecting unreliability recognized characters, i.e.,

$$\frac{g_{i_1}(x)}{g_{i_2}(x)} < TH_2^0$$

Moreover, some transforms (e.g., normalization) of the above equation should be used [17]. In our technology, we use the following equation for transformation,

$$r_2(x) = \frac{g_{i_1}(x) - g_{i_2}(x)}{g_{i_1}(x)} < TH_2 \tag{8}$$

Similarly, we get $0 \le r_2(x) \le 1$.

The above thresholding ways are reasonable for different situations. For example, if the confidence score is very high, Equation (7) is suitable. However, if classes are obviously different, Equation (8) is more reasonable. Our new technique is a hybrid way which combines the above two styles with a linear weighting,

$$\begin{aligned} r_3(x) &= \alpha r_1(x) + \beta r_2(x) \\ &= \alpha g_{i_1}(x) + \beta \frac{g_{i_1}(x) - g_{i_2}(x)}{g_{i_1}(x)} < TH_3 \end{aligned} \tag{9}$$

where $\alpha + \beta = 1$, and α and β can be experimental determined or learned by some learning technologies (e.g., GA algorithms). In our experiments, $\alpha = \beta = 0.5$. Similarly, we also get $0 \le r_3(x) \le 1$.

2.2. Rejection strategies for multiple classifiers

After performing rejection in each single classifier, some strategies should be adapted in multiple classifiers. Given *K* classifiers, $\{H_1, H_2, ..., H_K\}$, each classifier adopts a rejection strategy with Equation (9). We set

$$H_k(x \mid \omega_{i_1}) = \begin{cases} 1 & r_3^k(x) = \alpha g_{i_1}^k(x) + \beta \dfrac{g_{i_1}(x) - g_{i_2}(x)}{g_{i_1}(x)} < TH_3 \\ 0 & otherwise \end{cases}. \tag{10}$$

That is to say, if $H_k(x \mid \omega_{i_1}) = 1$, the recognition result is rejected; otherwise, accepted.

2.2.1. Voting combination with multiple classifiers

In our proposed rejection strategies with multiple features and multiple classifiers for handwritten character identification, three rejection frameworks are: (1) **OR**, (2) **AND**, and (3) **VOTING**. The precondition of these rejection strategies is that the class label outputs of multiple classifiers should be same. That is, in Equation (10), for one sample *x*, there is

$$\omega_{i_1}(x) = \omega_{i_2}(x) = ... = \omega_{i_M}(x).$$

If the output class labels of *x* are different, we do reject directly.

(1) OR $\quad F_I(x \mid \omega_i) = \coprod_{k=1}^{K} H_k(x \mid \omega_i)$

If $F_I(x \mid \omega_i) = 1$, the recognition result is rejected; otherwise, accepted.

(2) AND $\quad F_{II}(x \mid \omega_i) = \prod_{k=1}^{K} H_k(x \mid \omega_i)$

If $F_{II}(x \mid \omega_i) = 1$, the recognition result is rejected; otherwise, accepted.

(3) VOTING $\quad F_{III}(x \mid \omega_i) = \sum_{k=1}^{K} H_k(x \mid \omega_i)$

If $F_{III}(x \mid \omega_i) > N_{thres}$, the recognition result is rejected; otherwise, accepted. N_{thres} can be pre-defined. Always, $N_{thres} > N/2$, which means the majority voting.

3.2.2. Linear combination with multiple classifiers

We perform rejection strategies in two simple ways for classifier combination: sum-rule (average) and weighted combination.

(1) Average combination

In combination of fixed rules, we only investigate the most common usage: *sum-rule (averaging)*. Given *K* classifiers for *M*-class classification, the classifier outputs are re-scaled and transformed to confidence measures $g_m^k(x)$, $k = 1,...,K$, $m = 1,...,M$. In combination, the sum-rule computes the combined class scores by

$$g_m(x) = \frac{1}{K}\sum_{k=1}^{K} g_m^k(x), \; m = 1,...,M$$

This is equivalent to the averaging of confidence measures over the classifiers.

And the rejection strategy is the same to the one (Equation (10)) for the single classifier described in Section 2.1.3. That is,

$$r_3(x) = \alpha g_{m_1}(x) + \beta \frac{g_{i_1}(x) - g_{i_2}(x)}{g_{i_1}(x)} < TH_3$$

where $\{g_m(x), m=1,...,M\}$ are sorted in a descending order as $\{g_{m_1}(x), g_{m_2}(x),..., g_{m_M}(x)\}$.

(2) Weighted combination

In weighted combination, usually each constituent classifier has exactly one weight for sharing for all classes. Accordingly, the combined class score (confidence measure score) is computed by

$$g_m(x) = \sum_{k=1}^{K} w_k g_m^k(x),\ m = 1,...,M$$

where the classifier weights, $w = \{w_1, w_2,..., w_K\}$, can be estimated by regression on a validation data set to optimize the CE, MSE, or MCE criterion.

3. Experiments

The experiment data is the MNIST handwritten digits, which include 60,000 training examples and 10,000 testing examples. We experiment with three digit classifiers: three-layer BP neural network classifier, modified quadratic discriminant function (MQDF), and SVMs. The features used by the classifiers are weight direction code histogram features [7]. And we use the LIBSVM with the RBF kernel [2] for our SVM classifiers. Moreover, the measures of SVMs are the probabilities directly from SVM outputs. And all experiment results in the following are performed on the testing data. With simplicity and efficiency [5, 11, 12], we only use Equation (1) as the initial scaling function, and Equation (4) as the confidence measure calculation step.

We analyze the classifier accuracy with two measures: rejection rate on all characters which are recognized, and recognition error rate on characters which are not rejected. Assume *#ALL*, *#RE*, *#ER* are the number of all input characters, rejected characters, and incorrectly recognized characters respectively, then the rejection rate is *rate(RE) = #RE/#ALL*, and the recognition error rate is *rate(ER) = #ER/(#ALL-#RE)*. These two types of verification error naturally trade off; for example, raising the rejection threshold reduces the recognition error. Therefore, for each rejection strategy, we sweep a rejection threshold across its entire range values, plotting the two rate types, recognition error rate on recognized characters against the rejection rate on all the processed characters, as a receiver-operating-characteristic (ROC) curve. Generally speaking, a curve reaching closer to the origin indicates a superior confidence measure.

3.1. Experiments with the single classifier

The ROC curves with three different classifiers and three different rejection strategies (Equation (7), (8) and (9)) are shown in Figure 1 and Figure 2 is a scaling-up part of Figure 1. The red, green and blue lines are for the BP, SVM, and MQDF classifiers respectively; and the "——", "– –", and "····" lines are for the rejection strategies of Equation (7), (8), and (9) respectively.

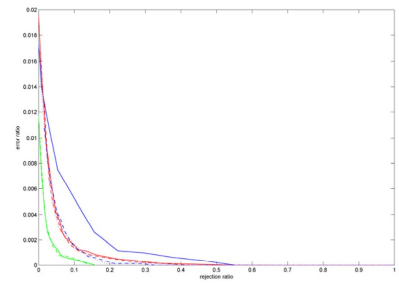

Figure 1. ROC curves with rejection and recognition error rates for the single classifier.

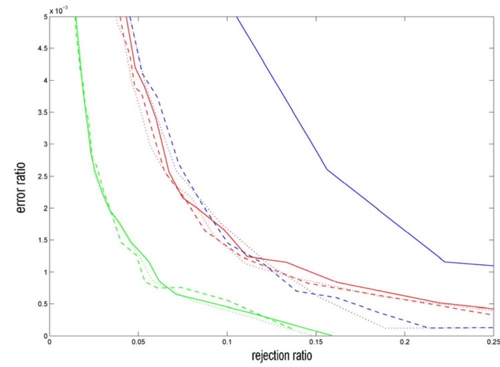

Figure 2. One scaling-up part near to the origin of Figure 1.

For different classifiers, the performance of SVM is the best. For different rejection strategies in the single classifier, strategies of Equation (8) and (9) have better performances, and the improvement is much obvious for the MQDF classifier. And our proposed new one (9) has the best ROC curve, which is more smoothing.

In our experiments, we only use Equation (1) as the initial scaling function, and Equation (4) as the confidence measure calculation step. To improve our systems, we should investigate other scaling functions and confidence measure calculation functions, such as the ones described in [11].

3.2. Experiments with multiple classifiers

In experiments for rejection strategies with multiple classifiers, we only use our new rejection strategy (Equation (9)) in each single classifier.

For rejection strategies with multiple classifiers, the ROC curves of the rejection rate and the recognition error rate are shown in Figure 3, where the red, green and blue "——" lines are for the OR, AND, and VOTING strategies in voting combination with multiple classifiers respectively; and the red and blue "····" lines are for the Average and Weighted Combination strategies in linear combination with multiple classifiers respectively.

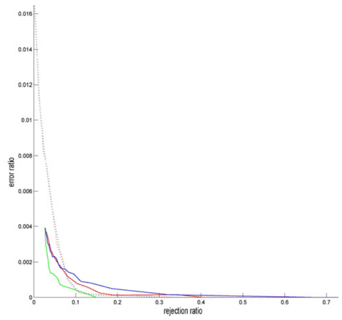

Figure 3. ROC curves with rejection and recognition error rates for multiple classifiers.

In voting combination with multiple classifiers, the "AND" rejection strategy gives the best performance. And the "OR" strategy greatly emphasizes the strategy target. Be surprised, the "VOTING" strategy is inferior to others. This may be because that the precondition of these rejection strategies is that the class label outputs from multiple classifiers must be same. In linear combination with multiple classifiers, the two rejection strategies have similar performances. Though Weighted Combination in classifier combination has shown better performance than the sum-rule method [12], for rejection strategies with multiple classifiers, this phenomenon is not obvious. Figure 3 also shows that the "AND" rejection strategy gives the best performance for both voting combination and linear combination with multiple classifiers.

4. Conclusions

In this paper, rejection strategies with multiple classifiers for handwritten character recognition are investigated. Our technology is composed by two parts: rejection strategies with the single classifier, and rejection strategies with multiple classifiers. There are three stages for reject strategies in the single classifier: initial scaling, confidence measure calculation, and rejection performing. For abstract level with classifier combination, rejection strategies with OR, AND and VOTING are analyzed; and for measurement level, rejection strategies of average and weighted combination are investigated.

One next issue is to investigate rejection strategies with more confidence transformation techniques and classifier combination methods, such as the ones described in [11, 12]. Another future direction is to analyze rejection strategies with not only multiple classifiers but also multiple features, especially for handwritten Chinese character recognition.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China under Grant No. 60675006.

References

[1] R. Bertolami, M. Zimmermann, and H. Bunke. Rejection strategies for offline handwritten text line recognition. *Pattern Recognition Letters*, 27(16): 2005-2012, 2006.

[2] C.-C. Chang, and C.-J. Lin. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[3] N. Gorski. Optimizing error-reject trade off in recognition systems. *Proceedings of ICDAR*, pages 1092-1096, 1997.

[4] H.W. Hao, X.H. Xiao, and R.W. Dai. Handwritten Chinese character recognition by metasynthetic approach. *Pattern Recognition*, 30(8): 1321-1328, 1997.

[5] H.W. Hao, C.-L. Liu, and H. Sako. Confidence evaluation for combining diverse classifier. *Proceedings of ICDAR*, pages 760-764, 2003.

[6] T.K. Ho, J. Hull, and S.N. Srihari. Decision combination in multiple classifier systems. *IEEE Trans. PAMI*, 16(1): 66-75, 1994.

[7] F. Kimura, T. Wakabayashi, S. Tsuruoka, and Y. Miyake. Improvement of handwritten Japanese character recognition using weighted direction code histogram. *Pattern Recognition*, 30(9): 1329-1337, 1997.

[8] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. PAMI*, 20(3): 226-239, 1998.

[9] A.L. Koerich. Rejection strategies for handwritten word recognition. *Proceedings of IWFHR*, pages 479-484, 2004.

[10] A.L. Koerich, R. Sabourin, and C.Y. Suen. Recognition and verification of unconstrained handwritten words. *IEEE Trans. PAMI*, 27(10): 1509-1522, 2005.

[11] C.-L, Liu, H.W. Hao, and H. Sako. Confidence transformation for combining classifiers. *Pattern Analysis and Applications*, 7(1): 2-17, 2004.

[12] C.-L, Liu. Classifier combination based on confidence transformation. *Pattern Recognition*, 38(11): 11-28, 2005.

[13] J. Pitrelli, and M.P. Perrone. Confidence modeling for verification post-processing for handwriting recognition. *Proceedings of IWFHR*, pages 30-35, 2002.

[14] J. Pitrelli, and M.P. Perrone. Confidence-scoring post-processing for off-line handwritten-character recognition verification. *Proceedings of ICDAR*, pages 278-282, 2003.

[15] X.-C. Yin, C.-P. Liu, and Z. Han. Feature combination using boosting. *Pattern Recognition Letters*, 26(14): 2195-2205, 2005.

[16] 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): 3415-3429, 2007.

[17] L. Zhen, Z.-S. Hu, J.-Y. Yang, K. Liu, and C.Y. Suen. Handwritten numerals recognition based on segmented contour feature. *Chinese Journal of Computers*, 22(10): 1065-1073, 1999. (In Chinese)

[18] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. SMC*, 22(3): 418-435, 1992.