Optimal Classifier Combination Rules for Verification and Identification Systems

Sergey Tulyakov, Venu Govindaraju, and Chaohong Wu

Center for Unified Biometrics and Sensors (CUBS), SUNY at Buffalo, USA

Abstract. Matching systems can be used in different operation tasks such as verification task and identification task. Different optimization criteria exist for these tasks - reducing cost of acceptance decisions for verification systems and minimizing misclassification rate for identification systems. In this paper we show that the optimal combination rules satisfying these criteria are also different. The difference is caused by the dependence of matching scores produced by a single matcher and assigned to different classes. We illustrate the theory by experiments with biometric matchers and handwritten word recognizers.

1 Introduction

Traditionally, the goal of pattern classification algorithms is to minimize the misclassification rate or cost[7]. With the development of biometric field another type of optimization criteria became important - minimizing the cost of verifying the hypothesis of whether the input belongs to the prespecified class. In particular, for biometric verification system we need to determine whether the presented biometric input belongs to the claimed enrolled person. The verification problem is a two-class problem - the input does belong to the hypothesis class (genuine verification attempt) or does not (impostor). On the other hand, the traditional classification problem still takes place in biometrics as an identification problem: given biometric input determine the person among N enrolled persons. Note, that similar task division existed before in other pattern recognition tasks. As an example of verification system in a handwriting application, a bank check recognition system might hypothesize about the value of the check based on the legal field, and numeric string recognition module must confirm that courtesy value coincides with the legal amount[4]. Or, more frequently, a handwriting recognition module is used to identify each word between N words in the lexicon.

It turns out that different tasks might require different optimizations of recognition algorithms. Example 1 of this paper presents two hypothetical recognition algorithms with one more suited for verification task and another for identification task. Similarly, if we have two or more matching algorithms, and we want to combine their results, the best combination algorithms might be different for different tasks. The goal of this paper is to show that this is indeed the case. Whereas the optimal combination algorithm for verification systems corresponds to likelihood ratio combination rule, the optimal combination algorithm for identification systems might be different, and it is rather difficult to find.

1.1 Performance Measures

Different modes of operation demand different performance measures. For verification systems the performance is traditionally measured by means of Receiver Operating Characteristic (ROC) curves or by Detection Error Trade-off (DET) curve. These curves are well suited for describing the performance of two-class pattern classification problems. In such problems there are two types of errors: the samples of first class are classified to belong to second class, and samples of second class are classified to be in first class. The decision to classify a sample to be in one of two classes is usually based on some threshold. Both performance curves show the relationship between two error rates with regards to a threshold (see [2] for precise definition of above performance measures). In our case we will use ROC curves for comparing algorithm performance.

For measuring performance of identification systems we will use ranking approach. In particular, we are interested in maximizing the rate of correctly identifying the input, first-rank-correct rate. If we look at identification task as a pattern classification problem, this performance measure will directly correspond to the traditional minimization of the classification error. Note that there are also other approaches to measure performance in identification systems[2], e.g. Rank Probability Mass, Cumulative Match Curve, Recall-Precision Curve. Though they might be useful for some applications, in our case we will be more interested in correct identification rate.

2 Verification Systems

The problem of combining matchers in verification systems can be easily solved with pattern classification approach. As we already noted, there are two classes: genuine verification attempts and impostor verification attempts. The hypothesis identity of the input is provided before matching. Each matcher j outputs a score s^j corresponding to a match confidence between input sample and hypothesis identity. Assuming that we combine M matchers, our task is to perform two-class classification (genuine and impostor) in M-dimensional score space $\{s^1, \dots, s^M\}$. If the number of combined matchers M is small, we will have no trouble in training pattern classification algorithm.

We employ the Bayesian risk minimization method as our classification approach[7]. This method states that the optimal decision boundaries between two classes can be found by comparing the likelihood ratio

$$f_{lr}(s^1, \dots, s^M) = \frac{p_{gen}(s^1, \dots, s^M)}{p_{imp}(s^1, \dots, s^M)} \tag{1}$$

to some threshold θ where p_{gen} and p_{imp} are M-dimensional densities of score tuples $\{s^1, \dots, s^M\}$ corresponding to two classes - genuine and impostor verification attempts. In order to use this method we have to estimate the densities p_{gen} and p_{imp} from the training data.

The likelihood ratio combination method is theoretically optimal for verification systems and its performance only limited by our ability to correctly estimate score densities. But, since our problem is the separation of genuine and impostor classes, we could apply many existing pattern classification techniques as well. For example, support vector machines have shown good performance in many tasks, and can be definitely used to improve the likelihood ratio method. In [8] we performed some comparisons of likelihood ratio method with SVMs on an artificial task and found that on average (over many random training sets) SVMs do have slightly better performance, but for a particular training set it might not be true. The difference in performance is quite small and decreases with the increasing number of training samples.

3 Identification Systems

In identification systems a hypothesis of the input sample is not available and we have to choose the input's class among all possible classes. Denote N as the number of classes. The total number of matching scores available for combination now is MN: N matching scores for N classes from each of M combined classifiers. If numbers M and N are not big, then we can use generic pattern classifiers in MN-dimensional score space to find the input's class among N classes. For some problems, e.g. digit or character recognition, this is an acceptable approach; the number of classes is small and usually there is a sufficient number of training samples to properly train pattern classification algorithms operating in MN score space.

But for our applications in handwritten word recognition and biometric person identification the number of classes is too big and the number of training samples is too small (there might be even no training samples at all for a particular lexicon word), so the pattern classification in the MN-dimensional score space seems to be out of the question. The traditional approach in this situation is to use some combination rules. The combination rule implies the use of some combination function f operating only on M scores corresponding to one class, $f(s^1, \dots, s^M)$, and it states that the decision class C is the one which maximizes the value of a combination function:

$$C = \arg \max_{i=1,\dots,N} f(s_i^1, \dots, s_i^M) \tag{2}$$

Note that in our notation the upper index of the score corresponds to the classifier, which produced this score, and lower index corresponds to the class for which it was produced. The names of combination rules are usually directly derived from the names of used combination functions: the sum function $f(s^1, \dots, s^M) = s^1 + \dots + s^M$ corresponds to the sum rule, the product function $f(s^1, \dots, s^M) = s^1 \dots s^M$ corresponds to the product rule and so on.

Many combination rules have been proposed so far, but there is no agreement on the best one. It seems that different applications require different combination rules for best performance. Anyone wishing to combine matchers in real life has to test few of them and choose the one with best performance.

3.1 Likelihood Ratio Combination Rule

As we already know, likelihood ratio function is the optimal combination function for verification systems. We want to investigate whether it will be optimal for identification systems. Suppose we performed a match of the input sample by all M matchers against all N classes and obtained MN matching scores $\{s_i^j\}_{i=1,\dots,N;j=1,\dots,M}$. Assuming equal prior class probabilities, the Bayes decision theory states that in order to minimize the misclassification rate the sample should be classified as one with highest value of likelihood function $p(\{s_i^j\}_{i=1,\dots,N;j=1,\dots,M}|\omega_k)$. Thus, for any two classes ω_1 and ω_2 we have to classify input as ω_1 rather than ω_2 if

$$p(\{s_i^j\}_{i=1,\dots,N;j=1,\dots,M}|\omega_1) > p(\{s_i^j\}_{i=1,\dots,N;j=1,\dots,M}|\omega_2) \tag{3}$$

Let us make an assumption that the scores assigned to each class are sampled independently from scores assigned to other classes; scores assigned to genuine class are sampled from M-dimensional genuine score density, and scores assigned to impostor classes are sampled from M-dimensional impostor score density:

$$\begin{aligned} &p(\{s_i^j\}_{i=1,\dots,N;j=1,\dots,M}|\omega_k) \\ &= p(\{s_1^1,\dots,s_1^M\},\dots,\{s_{\omega_k}^1,\dots,s_{\omega_i}^M\},\dots,\{s_N^1,\dots,s_N^M\}|\omega_k) \\ &= p_{imp}(s_1^1,\dots,s_1^M)\dots p_{gen}(s_{\omega_k}^1,\dots,s_{\omega_k}^M)\dots p_{imp}(s_N^1,\dots,s_N^M) \end{aligned} \tag{4}$$

After substituting 4 into 3 and canceling out common factors we obtain the following inequality for accepting class ω_1 rather than ω_2:

$$p_{gen}(s_{\omega_1}^1,\dots,s_{\omega_1}^M)p_{imp}(s_{\omega_2}^1,\dots,s_{\omega_2}^M) > p_{imp}(s_{\omega_1}^1,\dots,s_{\omega_1}^M)p_{gen}(s_{\omega_2}^1,\dots,s_{\omega_2}^M)$$

or

$$\frac{p_{gen}(s_{\omega_1}^1,\dots,s_{\omega_1}^M)}{p_{imp}(s_{\omega_1}^1,\dots,s_{\omega_1}^M)} > \frac{p_{gen}(s_{\omega_2}^1,\dots,s_{\omega_2}^M)}{p_{imp}(s_{\omega_2}^1,\dots,s_{\omega_2}^M)} \tag{5}$$

The terms in each part of the above inequality are exactly the values of the likelihood ratio function f_{lr} taken at the sets of scores assigned to classes ω_1 and ω_2. Thus, the class maximizing the MN-dimensional likelihood function of inequality 3 is the same as a class maximizing the M-dimensional likelihood ratio function of inequality 5. The likelihood ratio combination rule is the optimal combination rule under used assumptions of score independence.

The main assumption that we made while deriving likelihood ratio combination rule is that the score samples in each identification trial are independent. That is, genuine score is sampled from genuine score distribution and is independent from impostor scores which are independent and identically distributed

Matchers	$first_{imp}$	$second_{imp}$	$third_{imp}$	$mean_{imp}$
CMR	0.4359	0.4755	0.4771	0.1145
WMR	0.7885	0.7825	0.7663	0.5685
li	0.3164	0.3400	0.3389	0.2961
C	0.1419	0.1513	0.1562	0.1440
G	0.1339	0.1800	0.1827	0.1593

Table 1. Correlations between s_{gen} and different statistics of the impostor score sets produced during identification trials for considered matchers.

according to impostor score distribution. We can verify if this assumption is true for our matchers.

Table 1 shows correlations between genuine score and some functions of the impostor score sets obtained in the same identification trial. $first_{imp}$ column has correlations between genuine and the best impostor score, and so on. Non-zero correlations indicate that the scores are dependent, and likelihood ratio combination rule will not necessarily be optimal for our applications.

The main reason for the dependence among matching scores produced during identification trial is that they are derived using same input signal. The next two examples will illustrate the effect of score dependences on the performance of identification systems. In particular, second example confirms that if identification system uses likelihood ratio combination, then its performance can be worse than the performance of a single matcher.

Example 1 Suppose we have an identification system with one matcher and, for simplicity, $N = 2$ classes. During each identification attempt a matcher produces two scores corresponding to two classes, and, since by our assumption the input is one of these two classes (closed set identification), one of these scores will be genuine match score, and another will be impostor match score. Suppose we collected a data on the distributions of genuine and impostor scores and reconstructed score densities (let them be gaussian) as shown in Figure 1.

Consider two possible scenarios on how these densities might have originated from the sample of the identification attempts:

1. Both scores s_{gen} and s_{imp} are sampled independently from genuine and impostor distributions.
2. In every observed identification attempt : $s_{imp} = s_{gen} - 1$. Thus in this scenario the identification system always correctly places genuine sample on top. There is a strong dependency between scores given to two classes, and score distributions of Figure 1 do not reflect this fact.

If a system works in verification mode and we have only one match score to make a decision on accepting or rejecting input, we can only compare this score to some threshold. By doing so both scenarios would have same performance: the rate of false accepts (impostor samples having match score higher than threshold)

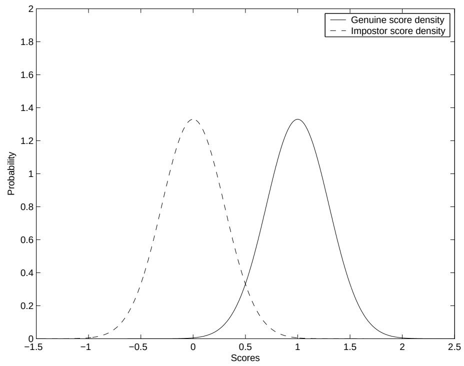


Fig. 1. Hypothetical densities of matching(genuine) and non-matching(impostors) scores.

and the rate of false rejects (genuine samples having match score lower than threshold) will be determined by integrating impostor and genuine densities of Figure 1 no matter what scenario we have. If system works in identification mode, the recognizer of the second scenario will be a clear winner: it is always correct while the recognizer of first scenario can make mistakes and place impostor samples on top.

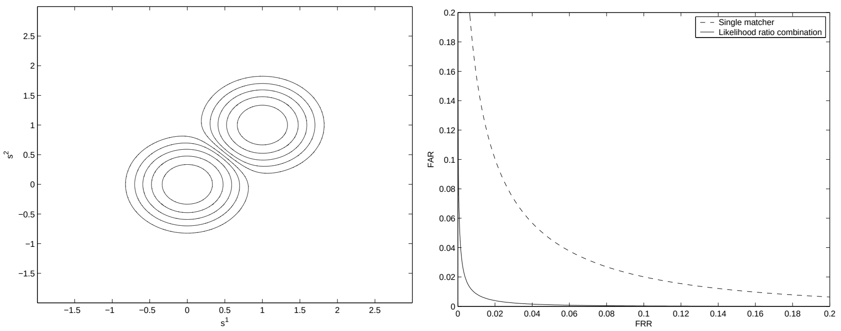
This example shows that the performance of the matcher in the verification system might not predict its performance in the identification system. Given two matchers, one might be better for verification systems, and another for identification systems.

Example 2 Consider a combination of two matchers in two class identification system: one matcher is from the first scenario, and the other is from the second scenario. Assume that these matchers are independent. Let the upper score index refer to the matcher producing this score; s_i^j is the score for class i assigned by the classifier j. From our construction we know that the second matcher always outputs genuine score on the top. So the optimal combination rule for identification system will simply discard scores of first matcher and retain scores of the second matcher:

$$f(s^1, s^2) = s^2 \tag{6}$$

The input will always be correctly classified as $\arg\max_i s_i^2$.

Let us now use the likelihood ratio combination rule for this system. Since we assumed that matchers are independent, the densities of genuine $p_{gen}(s^1, s^2)$ and impostor $p_{imp}(s^1, s^2)$ scores are obtained by multiplying corresponding one-dimensional score densities of two matchers. In our example, impostor scores are distributed as a Gaussian centered at $(0, 0)$, and genuine scores are distributed as a Gaussian centered at $(1, 1)$. Figure 2(a) contains the contours of function $|p_{gen} - p_{imp}|$ which allows us to see the relative position of these gaussians. The gaussians have same covariance matrix, and thus the optimal decision contours are hyperplanes[7] - lines $s^1 + s^2 = c$. Correspondingly, the likelihood ratio combination function is equivalent to the combination function $f = s^1 + s^2$ (note, that true likelihood ratio function will be different, but if two functions have same contours, then their combination rules will be the same). Such combination improves the performance of the verification system relative to any single matcher; Figure 2(b) shows corresponding ROC curves for any single matchers and their combination.



(a) (b)

Fig. 2. (a) Two-dimensional distributions of genuine and impostor scores for examples 2 and 3 (b) ROC curves for single matchers and their likelihood ratio combination.

Suppose that (s_1^1, s_1^2) and (s_2^1, s_2^2) are two score pairs obtained during one identification trial. The likelihood ratio combination rule classifies the input as a class maximizing likelihood ratio function:

$$\arg\max_{i=1,2} \frac{p_{gen}(s_i^1, s_i^2)}{p_{imp}(s_i^1, s_i^2)} = \arg\max_{i=1,2}\ s_i^1 + s_i^2 \tag{7}$$

Let the test sample be $(s_1^1, s_1^2) = (-0.1, 1.0)$, $(s_2^1, s_2^2) = (1.1, 0)$. We know from our construction that class 1 is the genuine class, since the second matcher assigned score 1.0 to it and 0 to the second class. But its score pair $(1.1, 0)$ is located just above the diagonal $s^1 + s^2 = 1$, and the score pair $(-0.1, 1.0)$ corresponding to class 1 is located just below this diagonal. Hence class 2 has bigger ratio of genuine to impostor densities than class 1, and the likelihood ratio combination method would incorrectly classify class 2 as the genuine class.

Thus the optimal for verification system likelihood ratio combination rule (7) has worse performance than a single second matcher. On the other hand, the optimal for identification system rule (6) does not improve the performance of the verification system. Recall, that in section 3.1 we showed that if scores assigned by matchers to different classes are independent, then likelihood ratio combination rule is optimal for identification systems, as well as for verification systems. Current example shows that if there is a dependency between scores, this is no longer a case, and the optimal combination for identification systems can be different from the optimal combination for verification systems.

4 Experiments

We have performed three sets of experiments for this paper - one for combining two word recognizers and two for combining fingerprint and face biometric matchers. Two handwritten word recognizers are Character Model Recognizer (CMR)[3] and Word Model Recognizer (WMR)[5]. Both recognizers employ similar approaches to word recognition: they oversegment the word images, match the combinations of segments to characters and derive a final matching score for each lexicon word as a function of character matching scores. Still, the correct identification rates of these recognizers (see Table 2) reveal that these matchers produce somewhat complementary results and their combination might be beneficial.

Our data consists of three sets of 2654, 1723 and 1770 word images representing UK postal town and county names of approximately same quality (the data was provided as these three subsets and we did not regroup them). The word recognizers were run on these images and their match scores for the total of 1681 lexicon words were saved. Since our data was already separated into three subsets, we used this structure for producing training and testing sets. Each experiment was repeated three times, each time one subset is used as a training set, and two other sets are used as test sets. Final results are derived as averages of these three training/testing phases.

We used biometric matching score set BSSR1 distributed by NIST[1]. This set contains matching scores for a fingerprint matcher and two face matchers 'C' and 'G'. Fingerprint matching scores are given for left index 'li' finger matches and right index 'ri' finger matches. In this work we used both face matching scores and fingerprint 'li' scores and we do two types of combinations: 'li'&'C' and 'li'&'G'. We used bigger subsets of this data set with 6000 identification attempts to create a set of virtual persons and their matching scores. After discarding enrollees and identification trials with failed biometric enrollment we obtained two equal sets - 2991 identification trials with 2997 enrolled persons with each part used as training and testing sets in two phases.

For our applications the number of matchers M is 2 and the number of training samples is large (bigger than 1000), so we can successfully estimate the score densities for the likelihood ratio combination method. We approximate both densities as the sums of 2-dimensional gaussian Parzen kernels. The window parameter is estimated by the maximum likelihood method on the training set[6] using leave-one-out technique. Note that window parameter is different for genuine and impostor density approximations.

4.1 Identification System Experiments

Table 2 shows the performance of likelihood ratio rule on our data sets. Whereas the combinations of biometric matchers have significantly higher correct identification rates than single matchers, the combination of word recognizers has lower correct identification rate than a single WMR matcher. Example 2 provides an explanation to this result; there is a strong dependence in matching scores for WMR and it affects the performance of likelihood ratio combination.

Matchers	Total	1st matcher is correct	2nd matcher is correct	Either one is correct	Likelihood Ratio Rule	Weighted Sum Rule
CMR&WMR	6147	3366	4744	5105	4293	5015
li&C	5982	4870	4856	5789	5817	5816
li&G	5982	4870	4635	5731	5737	5711

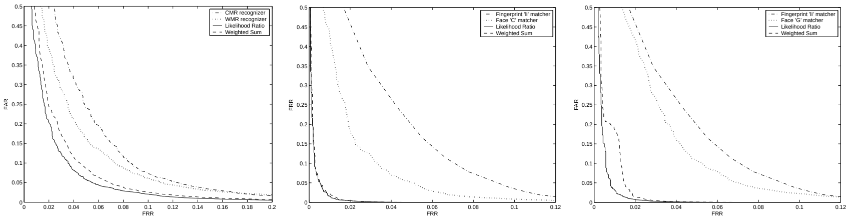
Table 2. Correct identification rate for likelihood ratio and weighted sum combination rules.

We compare the performance of the likelihood ratio combination method with the weighted sum combination rule $f(s^1, \ldots, s^M) = w_1 s^1 + \cdots + w_M s^M$. We train the weights so that the number of successful identification trials on the training set is maximized. Since we have two matchers in all configurations we use brute-force method: we calculate the correct identification rate of combination function $f(s^1, s^2) = ws^1 + (1-w)s^2$ for different values of $w \in [0, 1]$, and find w corresponding to highest rate.

The numbers of successful identification trials on the test sets is presented in Table 2. In all cases we see an improvement over the performances of single matchers. The combination of word recognizers is now successful and is in line with the performance of other combinations of matchers. Weighted sum method seems to perform slightly worse than likelihood ratio for biometric matchers, which can be explained by its simplicity. Another possible reason for this is that likelihood ratio combination rule is actually the optimal rule for classifiers with independent identification trial scores, and scores of biometric matchers show less dependence than scores of word recognizers.

4.2 Verification System Experiments

Figure 3 contains ROC curves likelihood ratio and weighted sum combination rules in verirification tasks. The weights in the weighted sum rule are the same as trained in identification experiments. In all cases we get slightly worse performance from the weighted sum rule than from the likelihood ratio rule. This confirms our assertion that the likelihood ratio is the optimal combination method for verification systems.



(a) (b) (c)

Fig. 3. ROC curves for combinations of (a) CMR and WMR, (b) 'li' and 'C', (c) 'li' and 'G'

5 Conclusion

The combination of matchers for verification problems is relatively easy task with likelihood ratio combination rule being the optimal method, as well as many other two-class pattern classification methods. On the other hand, the combination in identification problems might require different methods, and it is rather difficult task. In practice, presented results argue that we can not effectively use same combination method for both verification and identification. Though the weighted sum rule shows good performance in identification systems, there is a need to develop more finely trainable combination methods.

References

1. Nist biometric scores set. http://www.nist.gov/biometricscores/.
2. Ruud M. Bolle, Jonathan H. Connell, Sharath Pankanti, Nalini K. Ratha, and Andrew W. Senior. *Guide To Biometrics*. Springer, New York, 2004.
3. J.T. Favata. Character model word recognition. In *Fifth International Workshop on Frontiers in Handwriting Recognition*, pages 437–440, Essex, England, 1996.
4. G.Kim and V.Govindaraju. Bank check recognition using cross validation between legal and courtesy amounts. *Int'l J. Pattern Recognition and Artificial Intelligence*, 11(4):657–674, 1997.
5. G. Kim and V. Govindaraju. A lexicon driven approach to handwritten word recognition for real-time applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(4):366–379, 1997.
6. B W Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.
7. S. Theodoridis and Koutroumbas K. *Pattern Recognition*. Academic Press, 1999.
8. S. Tulyakov and Govindaraju V. Using independence assumption to improve multimodal biometric fusion. In *6th International Workshop on Multiple Classifiers Systems (MCS2005)*, Monterey, USA, 2005. Springer.