Classifier Ensemble Generation for the Majority Vote Rule

Carlos Orrite[1], Mario Rodríguez[1], Francisco Martínez[1], and Michael Fairhurst[2]

[1] Aragon Institute for Engineering Research, University of Zaragoza, Spain
[2] Electronics Department, University of Kent, UK

Abstract. This paper addresses the problem of classifier ensemble generation. The goal is to obtain an ensemble to achieve maximum recognition gains with the lowest number of classifiers. The final decision is taken following a majority vote rule. If the classifiers make independent errors, the majority vote outperforms the best classifier. Therefore, the ensemble should be formed by classifiers exhibiting individual accuracy and diversity. To account for the quality of the ensemble, this work uses a sigmoid function to measure the behavior of the ensemble in relation to the majority vote rule, over a test labelled data set.

Keywords: Combining classifiers, Ensemble generation, Majority vote.

1 Introduction

Majority vote is considered one of the simplest and most intuitive methods for combining classifier outputs [1], [2]. Majority vote counts the votes for each class over the input classifiers and selects the majority class. The idea of selecting a number of classifiers to make up an ensemble instead of using all classifiers has been dealt with in different ways. In [3] a method was proposed to select some neural networks based classifiers by a genetic algorithm. Other approaches are based on the measures of diversity to make different ensembles. Theoretically, if the classifiers make independent errors, the majority vote outperforms the best classifier. So, the use of diversity to generate ensembles of classifiers has been considered as an important concept in classifier fusion and has been addressed from different perspectives through research [1], [4]. In a recent work [5], a theoretic framework for combination of classifiers, taking into account accuracy and diversity, is presented. The author discusses the diversity/accuracy dilemma, giving as a conclusion that the two measures are somehow contradictory.

This paper introduces a new proposal to combine accuracy and diversity of classifiers from a final recognition point of view. To account for the quality of the ensemble, a sigmoid function is used to measure the behavior of the ensemble in relation to the majority vote rule over a test labelled data set. On the contrary to other feature selection methods, it doesn't need to know the number of classifiers to be in the final ensemble in advance. Instead, it decides which is the optimal one, working in an incremental way so, once a priori ensemble is established, it

J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 340–347, 2008.
© Springer-Verlag Berlin Heidelberg 2008

decides whether or not the inclusion of a new classifier improves the quality of the ensemble. If so, the new classifier becomes a member of the new ensemble.

The classifiers in the initial ensemble can be built on different subsets of features, or the same features but different classifier algorithms. The first approach is more related with traditional feature selection procedures and the second with combining classifiers methods or fusion of classifiers. So, the present proposal shares some aspects from one and the other, but it is focused in the majority vote rule. Thus, the main contribution of this paper is a suitable criterion function to evaluate the effectiveness of the final ensemble.

The present proposal is applied to the selection of classifiers for the development of biometrics recognition systems. For verification, the subject gives his/her identity and some biometrics features, the goal is to obtain the best set of Gabor filters for face recognition. In this way, it could be considered as a feature selection. The experiments are accomplished on the XM2VTS data base [8]. Results show a great improvement in the performance of the ensemble in relation to isolated classifiers. On the other hand, the algorithm is used for hand biometric identification, where no identity is supplied to the system. In this case, the set of features are geometrical relations and several neural networks based classifiers are trained over the same feature data.

2 Measure of Diversity in Classifier Ensembles

Consider a labelled data set $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ sampled from the classification problem in question. Given a set of L classifiers $\boldsymbol{P} = [D_1, \ldots, D_L]$ and considering an oracle output for classifier D_i, it is organized as an N-dimensional binary vector $\boldsymbol{y}_i = [y_{1,i}, \ldots, y_{N,i}]^T$, such that $y_{j,i}$= 1, if D_i recognizes correctly $\boldsymbol{x}_j$, and 0, otherwise, $i = 1, \ldots, L$, see Table 1.

Table 1. Structure of ensemble outputs

	D_1	$\cdots$	D_i	$\cdots$	D_L
$\boldsymbol{x}_1$	$y_{1,1}$	$\cdots$	$y_{1,i}$	$\cdots$	$y_{1,L}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$\boldsymbol{x}_j$	$y_{j,1}$	$\cdots$	$y_{j,i}$	$\cdots$	$y_{j,L}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$\boldsymbol{x}_N$	$y_{N,1}$	$\cdots$	$y_{N,i}$	$\cdots$	$y_{N,L}$

Taking into account Table 1, it shows that $N_1(D_i) = \sum_{j=1}^{N} y_{j,i}$ is the number of correct assignments and $N_0(D_i) = N - \sum_{j=1}^{N} y_{j,i}$ is the number of mistakes.

The goal of this paper is to obtain an ensemble $\boldsymbol{E}$ of M classifiers $\boldsymbol{E} = \{D_1^e, \ldots, D_M^e\}, M < L$, from the previous pool in order to achieve maximum recognition gains following the majority vote rule.

As mentioned before, diversity is an important consideration in classifier ensembles; it can be potentially exploited in order to obtain higher classification

accuracy. Several papers can be found in research in this area dealing with combination of classifiers [1], [4]. Most of them are based on the concept of diversity for ensemble generation. To date, there is no widely accepted formal definition of diversity in classifier ensembles. Practically, there are pairwise measures which are calculated for each pair of classifiers in the ensemble and then averaged. There are also non-pairwise measures that use the idea on entropy or correlation of individual outputs with the averaged output of the ensemble, see [1], for more details.

However, these metrics do not take into account the accuracy of the classifier in order to select a new one for the final ensemble. As mentioned before, the goal is to obtain an ensemble, from a pool of classifiers, in some way that they exhibit good individual classifier accuracy as well as ensemble diversity in order to achieve maximum recognition gains with the lowest number M of classifiers.

To obtain a new ensemble we combine individual accuracy and diversity. To measure the degree of accuracy for the D_i classifier we use:

$$p_{\mathrm{a}}(D_i) = \frac{N_1(D_i)}{N} \tag{1}$$

where N and N_1, have been defined before.

So, classifiers exhibiting high values will be good candidates for the ensemble. Note that p_a varies between 0 and 1. Besides, we have to measure the diversity in classifier ensembles, or in other words, we seek classifiers committing errors in different samples. Given a classifier D_i from the pool $\boldsymbol{P}$, we need to measure the correlation degree of this classifier with the ensemble $\boldsymbol{E}$. In the previous section we have introduced some metrics Δ, for diversity measurement, based on pairwise or non-pairwise approaches, and now we use them to obtain the correlation degree p_c. It is important to notice that some normalization has to be carried out in the diversity measures to obtain a p_c score between 0 (no correlation at all) and 1 (completely correlated).

For non-pairwise diversity for $D_i \cup \boldsymbol{E}$ we have:

$$p_{\mathrm{c}}(D_i, \boldsymbol{E}) = \Delta(D_i \cup E) \tag{2}$$

For pairwise diversity for $D_i \cup \boldsymbol{E}$:

$$p_{\mathrm{c}}(D_i, \boldsymbol{E}) = \max_j \Delta(D_j, D_i) \tag{3}$$

So, the acceptability degree $\mathcal{A}$ for D_i to be included in the ensemble $\boldsymbol{E}$ is given by:

$$\mathcal{A}(D_i, \boldsymbol{E}) = \alpha \cdot p_{\mathrm{a}}(D_i) - (1-\alpha) \cdot p_{\mathrm{c}}(D_i, \boldsymbol{E}) \tag{4}$$

where α weights the contribution of the classifier accuracy to the ensemble independency, note again that p_a and p_c varies between 0 and 1. The chosen classifier D_i will be that with the highest acceptability degree given by (4).

The proposed algorithm starts choosing from $\boldsymbol{P}$ the classifier with the highest accuracy. We then proceed to choose the next classifier giving the highest score

given by (4) and so on until the predefined number M of classifiers is reached. The final decision from the ensemble is taken following a majority vote rule.

3 A New Proposal for Classifier Ensemble Generation

The selection of classifiers for fusion at matching stage, has to be based on some "goodness" statistic to avoid performance degradation when using classifier combination techniques like majority vote rule. To count for the "goodness" of the ensemble (i.e., individual accuracy and diversity) this paper uses a function to assign a value between 0 and 1 for the number of classifiers that correctly classify $\boldsymbol{x}_j$ (0 if all are wrong and 1 if all are right). Bearing in mind the majority vote rule, the critical point is close to half the number of classifiers, so the function should be non-linear, giving more relevance to this critical point. To account for this non-linear relation this paper uses the sigmoid function given by the expression:

$$sigmoid = \frac{1}{1 + \mathrm{e}^{-\rho}} \tag{5}$$

To establish the correspondence between the ρ parameter and the number of classifiers, a normalization function is introduced. *Normalize* is a linear function to assign $\rho = -5$ when all the classifiers are wrong ($N_1(\boldsymbol{x}_j) = 0$) and $\rho = +5$ when all are right($N_1(\boldsymbol{x}_j) = L$). Therefore, for the half number of classifiers correct a value $\rho = 0$ is obtained.

From a practical point of view, the sigmoid function gives a non linear value in relation to the number of classifiers that correctly classify x_j. In this sense, the addition of a new classifier exhibiting an error is not relevant when the rest of the classifiers in the ensemble are all wrong or correct (sigmoid value equal to -5 or +5). The critical point is just in the middle, when the majority vote rule may change from 0 to 1 (or 1 to 0). So, the inclusion of a wrong classifier has to be penalized when the ensemble outputs are near to this critical point.

Finally, for the whole labelled data set $\boldsymbol{X}$ the following expression for the quality of the ensemble is proposed:

$$\mathcal{S}(\boldsymbol{E}; \boldsymbol{X}) = \frac{1}{N} \sum_{\mathrm{j}=1}^{N} sigmoid\big(Normalize(N_1(\mathrm{x_j}))\big) \tag{6}$$

Note that $\mathcal{S}(\boldsymbol{E}; \boldsymbol{X})$ is between 0 and 1.

The proposed algorithm, from this point onwards will be referred to as the "Sigmoid Algorithm" (**SA**), starts choosing the classifier with the highest accuracy. A second classifier is chosen exhibiting the lower correlation with the previous one, following a pair-wise diversity measure. Afterwards the rest are considered, selecting in first place that classifier given the highest score in equation (6), and finishing when the inclusion of a new classifier does not improve the quality of the ensemble.

However, this strategy assumes that the classifier with the highest accuracy has to be in the final ensemble and this might not be always true. Therefore, after

Algorithm 1. New Proposal for Ensemble Generation: Sigmoid Algorithm (**SA**)

Input: training set $\boldsymbol{X}$, pool of classifiers $\boldsymbol{P}$

1. **Ensemble initialization** $\boldsymbol{E} = \{\boldsymbol{D_i}, \boldsymbol{D_j}\}$
   $D_i \in \boldsymbol{P}$; $D_i = \arg\max_i p_a(D_i)$.
   $D_j \in \boldsymbol{P}$; $D_j = \arg\max_j \Delta(D_i, D_j)$
2. **Find the classifier** $\boldsymbol{D_i}$ **from** $\boldsymbol{P}$ **:**
   $D_i = \arg\max_i \mathcal{S}(\boldsymbol{E} \cup D_i; \boldsymbol{X})$
3. **While** $\mathcal{S}(\boldsymbol{E} \cup \boldsymbol{D_i}; \boldsymbol{X}) > \mathcal{S}(\boldsymbol{E}; \boldsymbol{X})$
   $\boldsymbol{E} = \boldsymbol{E} \cup D_i$
   For all classifiers $\boldsymbol{D_j^e}$ **in** $\boldsymbol{E}$
   If $\mathcal{S}(\boldsymbol{E} - D_j^e; \boldsymbol{X}) > \mathcal{S}(\boldsymbol{E} \cup D_i; \boldsymbol{X})$ then
   Remove D_j^e from $\boldsymbol{E}$
   end For
   Find the classifier D_i from $\boldsymbol{P}$:
   $D_i = \arg\max_i \mathcal{S}(\boldsymbol{E} \cup D_i; \boldsymbol{X})$
   End while

Output: ensemble of classifiers $\boldsymbol{E}$

a new inclusion of a new classifier, the algorithm should check if the removing of any of the old one improves the ensemble performance measured by (6).

The final algorithm for ensemble generation is given in Algorithm 1. This algorithm resembles the Floating Search algorithms used for feature selection, [6]. It is worth noting that the **SA** allows the inclusion of a classifier several times in the final ensemble. In this sense, the combination scheme becomes a weight majority vote rule.

4 Experiments and Discussion

4.1 Sigmoid Algorithm (SA) for Face Verification

One of the mostly commonly deployed and successful appearance-based methods for face recognition is the Gabor decomposition. Gabor filters have been used with great success in face recognition, see as a review in [7] because they can capture salient visual properties such as spatial localization, orientation selectivity, and special frequency characteristics. The Gabor representation of a facial image $\boldsymbol{G}_{u,v}(\boldsymbol{z})$ can be obtained by convolving the image $\boldsymbol{I}(\boldsymbol{z})$ with the family of Gabor filters $\Phi_{u,v}(\boldsymbol{z})$, where u denotes orientation and v scale. In this work seven scales and eight orientations are used.

The classification method used is based on eigenfaces, but instead of dealing with the gray level image, the input data are the Gabor representations of the facial image.

The proposed method was tested on the XM2VTS face database, [8] which contains 2360 images of 295 subjects, with 8 faces for each subject, divided into 4

sessions of 2 images in each one. The testing is performed following the Laussane protocol, which splits the database into 3 different sets for training, evaluation and test. The database comprises of 200 people used as "clients" and 95 persons used as "impostors". The client images for training were acquired from the first two sessions and the client images for evaluation from the third session. The fourth session is exclusively used for testing the algorithm. The impostors are divided in two groups: 25 persons for training and 70 for test.

All images were cropped and rectified according to the manually located eye positions supplied with the XM2VTS data. The normalized images were 128 x 128 pixels. Every image was filtered with up to 56 Gabor filters. Afterwards, Principal Component Analysis (PCA) was first applied on every Gabor filter output to accomplish dimensionality reduction. The decision of acceptance or rejection is based on a measurement of similarity between the gallery and the average of client's training images with a global threshold. The similarity of two features is defined as the cosine distance, of its projection vector in PCA subspace. The set of a projected Gabor facial representation images in the new space and the cosine metric produce a classifier. The threshold for every classifier is selected at the equal error point, EER, at which the False Acceptance Ratio (FAR) is equal to the False Rejection Ratio (FRR) on the evaluation set. Finally, the (**SA**) is applied to select the best set of classifiers for the majority vote rule.

To measure the performance of the verification system the Total Error Rate (TER) was used. TER is defined by the sum of the false acceptance and false rejection rates: TER=FAR+FRR. Figure 1 represents the scores given by the sigmoid function (6) for different number of classifiers for the test and evaluation sets. Notice that the ensemble is improved with the inclusion of a new classifier to a certain point, though for about 13 classifiers the improvement in the ensemble is practically insignificant. Furthermore, there is a point, exactly 21 classifiers, where the ensemble output degenerates due to the lack of ensemble diversity.

The **SA** is compared with different measurements of diversity used in other works [1], i.e, the Q-statistic (Q), the correlation coefficient (ρ), the disagreement measure (D), the double-fault measure (DF), the entropy measure (E), the measure of difficulty (θ), the Kohavi-Wolpert variance (KW), and the Interrater agreement (k). The ensemble is generated following (4), where parameter α weights the classifiers' accuracy and the ensemble diversity. Bearing in mind that the main goal of the ensemble is to obtain the best output with the lowest number of classifiers, different values of α have been tested, Figure 1 shows the TER for the evaluation data set. The values for the optimum α are: Q (0.00), ρ (0.11), D (0.43), DF (0.10), E (0.00), θ (0.08), KW (0.12) and k (0.03). Every α value is chosen so the classifier exhibits the fastest descendent curve, given a low TER with the lowest number of classifiers.

4.2 Sigmoid Algorithm (SA) for Hand Biometric Identification

To test the algorithm in identification a public database consisting of 5400 images obtained from 280 users with 10 samples of each hand per person was used, [9].

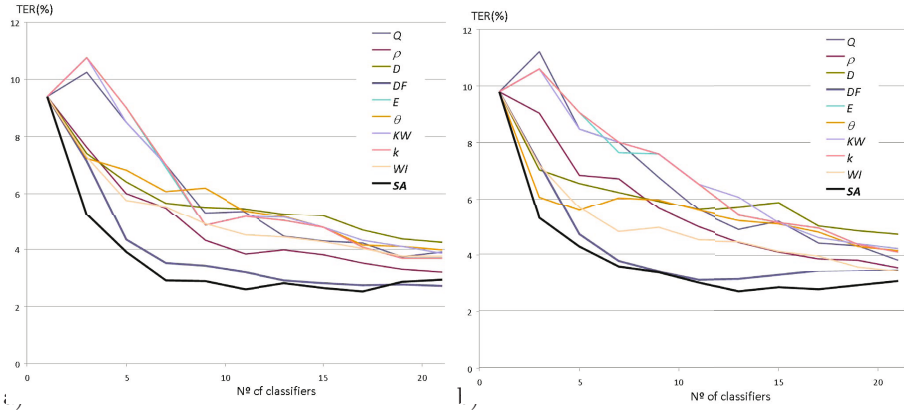

Fig. 1. (a) TER(test). (b)TER(evaluation).

The biometric recognition was only concerned to the right hand. Of these 10 right-hand samples, 5 samples were used for training, 3 for evaluation and 2 for testing. Geometric features were obtained following a previous work [10]. Up to 21 neural classifiers were trained with different numbers of neurons. After the evaluation stage, the SA determined that no improvement was reached for more than 7 classifiers. As a matter of fact, the sigmoid function (6) gave similar scores for 5 and 7 classifiers (i.e, 0.9795 for the former and 0.9797 for the later).

Table 2 shows the testing recognition rates (**RR**) and the total number of errors corresponding to the best classifier, the majority vote rule for all classifiers, and the majority vote rule for the ensemble given by SA, with 5 and 7 classifiers. As can be seen, the majority vote rule over the whole ensemble gives the worst score. The final ensemble given by the **SA** gives a slightly better improvement.

It is worthy to mention that the best classifier, from the accuracy point of view, is repeated 3 times in the final ensemble of 7 classifiers and 2 for the ensemble of 5 classifiers.

Table 2. Hand geometric-based identification

	Best classifier	**MV (21)**	**MV (5)**	**MV (7)**
Total errors	11	18	10	10
RR (%)	98.59	97.69	98.72	98.72

5 Conclusions

This paper introduces a new method for classifier ensemble generation. The fusion takes place at the matching stage following a majority vote rule. To count for the ensemble quality, a sigmoid function was proposed to measure the behavior of the ensemble in relation to the majority vote rule over a test labelled data set from the classification problem in question. The proposed algorithm works in

an incremental way so, once a priori ensemble is established, it decides whether or not the inclusion of a new classifier improves the quality of the ensemble. If so, it becomes a member of the new ensemble.

The algorithm has been applied to the selection of the best Gabor filters for facial representation. The main goal of this work was not to obtain the best face verification procedure, but to present a new approach for ensemble generation based on majority vote rule. More classifiers in the ensemble do not mean a better recognition rate due to redundancy among some classifiers. The comparison with other measurements of diversity has resulted in the fact that the **SA** reaches the highest recognition, with the lowest number of classifiers, much quicker than the other methods.

Acknowledgments. This work is supported by Spanish Grant TIN2006-11044 (MEyC), FEDER and BioSecure Network of Excelence (IST-2002-507634).

References

1. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley InterScience, Chichester (2004)
2. Oh, S.: On the relationship between majority vote accuracy and dependency in multiple classifier systems. Pattern Recog. Letters 24, 359–363 (2003)
3. Zhou, Z., Wu, J., Tang, W.: On the Relation between Dependence and Diversity in Multiple Classifier Systems. In: Proceedings of the International Conference on Information Technology, ITCC 2005, Las Vegas, USA, pp. 134–139 (2005)
4. Narasimhamurthy, A.: Evaluation of diversity measures for binary classifier ensembles. In: Oza, N.C., Polikar, R., Kittler, J., Roli, F. (eds.) MCS 2005. LNCS, vol. 3541, pp. 267–277. Springer, Heidelberg (2005)
5. Meynet, J.: Information theoretic combination of classifiers with application to face detection, Ph.D. Thesis (2007)
6. Pudil, P., Ferri, F.J., Novovicova, J., Kittler, J.: Floating search methods for feature selection with nonmonotoniccriterion functions. In: Conference on Pattern Recognition, 1994, vol. 2, pp. 279–283 (1994)
7. Shen, L., Bai, L., Fairhurst, F.: Gabor wavelets and General Discriminat Analysis for face identification and verification. Image and Vision Computing 25(5), 553–563 (2007)
8. The Xm2VTSDB, `http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/`
9. Hand dataset, `http://visgraph.cs.ust.hk/biometrics/Visgraph_web/index.html`
10. Martinez, F., Orrite, C., Herrero, E.: Biometric Hand Recognition using Neural Networks. In: Cabestany, J., Gonzalez Prieto, A., Sandoval, F. (eds.) IWANN 2005. LNCS, vol. 3512, pp. 1164–1171. Springer, Heidelberg (2005)