Confusion Matrix Disagreement for Multiple Classifiers

Cinthia O.A. Freitas[1], João M. de Carvalho[2], José Josemar Oliveira Jr[2], Simone B.K. Aires[3], and Robert Sabourin[4]

[1] Pontificia Universidade Católica do Paraná – PUCPR - Brazil
[2] Universidade Federal de Campina Grande – UFCG - Brazil
[3] Universidade Tecnológica Federal do Paraná – Campus Ponta Grossa – UTFP-PG - Brazil
[4] École de Technologie Supérieure - ETS - Canada
cinthia@ppgia.pucpr.br, carvalho@dee.ufcg.edu.br, josemar@dee.ufcg.edu.br, sbkaminski@pg.cefetpr.br, robert.sabourin@etsmtl.ca

Abstract. We present a methodology to analyze Multiple Classifiers Systems (MCS) performance, using the disagreement concept. The goal is to define an alternative approach to the conventional recognition rate criterion, which usually requires an exhaustive combination search. This approach defines a Distance-based Disagreement (DbD) measure using an Euclidean distance computed between confusion matrices and a soft-correlation rule to indicate the most likely candidates to the best classifiers ensemble. As case study, we apply this strategy to two different handwritten recognition systems. Experimental results indicate that the method proposed can be used as a low-cost alternative to conventional approaches.

Keywords: multiple classifiers systems, pattern recognition, classifiers diversity, handwriting recognition.

1 Introduction

The traditional pattern recognition approach divides the recognition task in two steps: first, a feature set is extracted from the images; second, the classifier computes the class-conditional probabilities based on the extracted features. Different feature sets can be proposed as well as many distinct classifiers can be designed. Therefore, the problem is to find the best combination of feature set and classifier.

In order to solve this problem many researchers have recently adopted the strategy of utilizing Multiple Classifiers Systems (MCS). The aim is to design a composite system that outperforms any of its individual component classifiers. The underlying principle is that it is more difficult to design one single complex classifier than to optimize a combination of relatively less complex ones. Several combination methods are found in the literature, however, one open question remains: what are the requirements to be fulfilled by the classifier components?

The concept of diversity has been used to answer that question, given that an ensemble of identical classifiers will not outperform its components individually. Evidence indicates that diversity within an ensemble is vital for its success [1] [2]. However, there is no general agreement on how to quantify neither diversity nor its

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 387–396, 2007.

relation with the ensemble average recognition rate. Aiming to group classification problems in a consistent way, Duin et al. [3] applied disagreement to measure the difference between two distinct classifiers trained on the same classification problem, which may be helpful in selecting appropriate tools for solving those problems. They admit, however, that there is still much to be investigated about disagreement between classifiers, although preliminary results are encouraging.

We are interested in designing a method that does not use first-order information (classifier's score output) to evaluate the ensemble. The idea is to use information from the confusion matrix for each individual classifier and compute distances between those matrices that represent classifier disagreements. Thus, distances will provide a mechanism for *a priori* evaluation of the possible classifier combinations.

2 Distance-Based Disagreement Classifiers Combination

A method for designing pattern recognition systems, known as MCS or committee/ ensemble approach, has emerged over recent years to tackle the practical problem of designing classification systems with improved accuracy and efficiency [1].

Attempts to understand the effectiveness of the MCS framework have prompted the development of several measures, like margin, bias and variance. Recently, many diversity measures have been studied to determine how they correlate with ensemble accuracy [4].

In trying to achieve this goal, the main question is: How can we measure the efficiency of a MCS? Our answer is to measure disagreements between classifiers, using their confusion matrices. This information can provide a mechanism to understand which classifiers can effectively contribute to boost the efficiency of the ensemble.

2.1 Diversity and Disagreement

Diversity measures can be categorized into two types [4]:

- **pair-wise:** calculates the average of a particular distance metric between all possible pairings of classifiers in the ensemble. The distance metric used determines the characteristics of the diversity measure;
- **non-pair-wise:** uses entropy or another similar measure to calculate a correlation of each ensemble member with the averaged ensemble output.

The main difficulty with the use of diversity measures is the so-called accuracy-diversity dilemma. As explained by Hadjitodorov et al., it is not clear how to choose the degree of diversity which produces the best performance, leading to a desired tradeoff between diversity and accuracy [5]. These authors also point that no convincing theory or experimental study has emerged to indicate a reliable measure to predict the generalization error of an ensemble. Other authors have stressed the need to find a balance point between diversity and accuracy [1], [4], reaching no agreement, however, regarding the choice of disagreement measure.

Duin et al. [3], use the disagreement concept to measure the difference between two classifiers C_1 and C_2 trained on a classification problem P_j(j = 1,…,N; N is the size of the set of problems). The disagreement is formulated as in Equation 1:

$$d_j(C_1,C_2) = Prob(C_1(x) \sim= C_2(x) \mid x \in P_j) \tag{1}$$

where $C_i(x)$ returns the label for object x according to classifier C_i. M classifiers constitute an MxM disagreement matrix D_j^C for problem P_j, with elements $D_j^C(m,n) = d_j(C_m,C_n)$.

In this work we take a different approach from that of Duin et al. [3], although also based on disagreement. The idea is to use the confusion matrix for each individual classifier to compute distances that represent classifier disagreements. We call our approach Distance-based Disagreement (DbD) criterion.

2.2 Confusion Matrix

A consistent analysis of classifier behavior can be provided by the semi-global performance matrix, known as the Confusion Matrix. This matrix provides a quantitative performance representation for each classifier in terms of class recognition. The Confusion Matrix can be denoted as in Equation 2 [2]:

$$A = \begin{bmatrix} RR_{1,1} & RR_{1,2} & \cdots & RR_{1,N} \\ \vdots & \vdots & & \vdots \\ RR_{2,1} & RR_{2,2} & \cdots & RR_{2,N} \\ \vdots & \vdots & & \vdots \\ RR_{N,1} & RR_{N,2} & \cdots & RR_{N,N} \end{bmatrix} \tag{2}$$

where $RR_{i,j}$ corresponds to the total number of entities in class C_i which have been classified in class C_j. Hence, the main diagonal elements indicate the total number of samples in class C_i correctly recognized by the system. From matrix A, it is possible to compute a global performance index for classifier A, defined by Equation 3:

$$RR^A = \frac{1}{N}\sum_{i,j=1}^{N} RR_{i,j} \tag{3}$$

For the ensemble of classifiers $A, B,…,M$ (considering that all confusion matrices are of the same size), a distance measure D^A between classifier A and all other classifiers is provided by Equation 4:

$$D^A = \sum_{i=1}^{N}\sum_{j=1}^{N}\left|RR_{i,j}^A - RR_{i,j}^B ... - RR_{i,j}^M\right| \tag{4}$$

where $RR_{i,j}^k$, k=$A,B,…M$, are the elements of the confusion matrix for classifier k. This distance can be similarly calculated for all members of the ensemble. Therefore, for an ensemble of M classifiers, we can define a Distance-based Disagreement (DbD) measure, expressed by Equation 5:

$$DbD = \sum_{k=A,B,...M} D^k \tag{5}$$

which uses information from the confusion matrix of each individual classifier and expresses classifiers disagreement.

2.3 Hypothesis: Soft-Correlation Rule

Our hypothesis is based on the following idea proposed by Hadjitodorov et al. [5]: "The ensembles selected through median diversity will fare better than randomly selected ensembles or ensembles selected through maximum diversity". We are calling this hypothesis the Soft-Correlation Rule. These authors observed that excessively increasing diversity does not lead to more accurate ensembles. They intuitively explain this phenomenon with the notion that in pattern clustering more diversity is associated with many clusters not getting the clustering structure right, leading to lower individual accuracy.

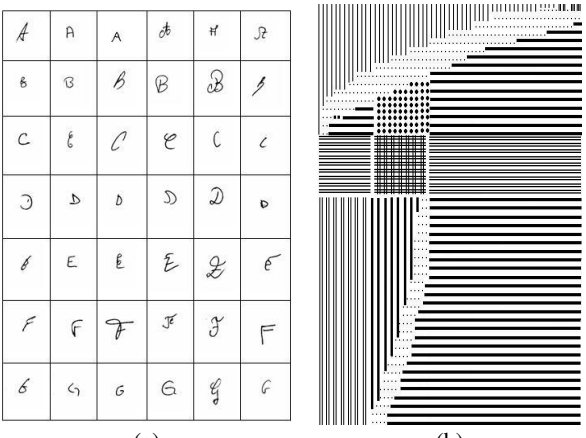
Considering this hypothesis, our proposal is to compute ensemble diversity from the distances (Eq. 4) between confusion matrices of the component classifiers, and to verify whether or not the best ensemble performance corresponds to the median diversity value. This methodology has been applied to two handwriting recognition problem, as described next.

3 Case 1: Character Recognition

For feature extraction, the baseline system for handwritten character recognition used in this work combines global and local (based on a zoning mechanism) approaches, and uses feedforward MLP (Multiple Layer Perceptron) Class-Modular architecture in the classification stage, where the modular MLP classifier consists of K sub-networks, M_i for $0 \leq i \leq K$-1, each responsible for one of the K classes [6].

The system gets as input a 256 grey-level image, as depicted in Fig. 1a. The preprocessing step is composed of binarization and bounding box definition. The feature set is obtained by labeling the background pixels of the input image as belonging to either a concavity or a convexity region [7], as presented in Fig. 1b. The alphabet of symbols was adapted to handwritten characters, resulting in 24 different symbols.

Several authors have presented zoning mechanisms or regional decomposition methods to investigate the recognition of patterns from their parts, similarly to what the human brain does during the reading process. Suen et al. [8] and Li et al. [9] applied a zoning mechanism in their experiments with hand printed characters. They analyzed different zone configurations, framing the character by a rectangle partitioned into Z parts. Based on these studies, we tested several zoning mechanisms, for Z equal to 4, 5 horizontal (5H), 5 vertical (5V) and 7, as shown in Fig. 2. Each image zone defines one classifier and these classifiers combined constitute a MCS. The zoning approach is beyond the scope of this paper [10].



(a) (b)

Fig. 1. Feature extraction: a) sample images from the character database and b) feature extraction (character "T")

4 Case 2: Word Recognition

The word recognition problem analyzed in this work is the recognition of handwritten month words on Brazilian bank checks. This is an important task, since it constitutes a sub-problem of bank check date recognition. This study deals only with recognition of the portuguese language month names represented by a limited lexicon of 12 classes: *Janeiro, Fevereiro, Março, Abril, Maio, Junho, Julho, Agosto, Setembro, Outubro, Novembro*, and *Dezembro. S*ome of these classes share sub-strings of characters, therefore adding to the problem complexity.

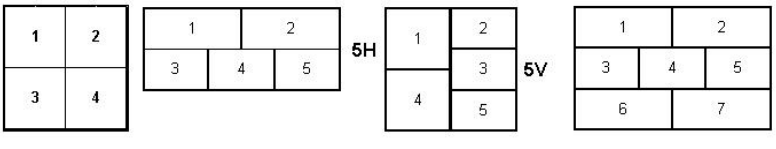


Fig. 2. Zoning mechanism: Z = 4, 5V, 5H and 7 parts

The baseline system utilizes multiple classifiers to avoid the intrinsic difficulties of the lexicon, by combining complementary information obtained from distinct sources (classifiers). Therefore, two different classifiers (Class-Modular Neural Networks [6] and Hidden Markov Models [11]) based on five different feature sets were evaluated. Similarly as used in character recognition, zoning mechanisms were utilized for feature extraction in order to add robustness to the system [8], [9]. The implemented zoning schemes are the following (Fig. 3):

- 2 fixed sub-regions (2-FS): Zoning splits the image in two areas defined at the right and at the left of the word center of gravity (Fig. 3a). These system extracts 14 features from each word in order to generate a feature vector of dimension 24 [12];

- 8 fixed sub-regions (8-FS): Each sample image is divided in 8 sub-regions of equal size (Fig. 3b). This number corresponds to the average number of letters in the lexicon words. In this zoning mechanism, three different feature extraction were evaluated [12]: perceptual, directional, and topological feature sets;
- N-variable sub-regions (N-VS): The features are extracted from the words images and a pseudo-segmentation process is applied to obtain a sequence of corresponding observations (Fig. 3c). Between two black-white transitions over the maximum peak of the horizontal transition histogram, called the Median Line, a segment is delimited and a corresponding symbol is designated to represent the extracted set of features, making up a grapheme [12], [13].

5 Experimental Results

This Section presents the database used in the experiments performed to evaluate the DbD criterion and the experimental results obtained with the MCS applied to the two cases described in Sections 3 and 4.

5.1 Character and Word Databases

The experiments applying characters were carried out using the handwritten character database called IRONOFF (IRESTE/University of Nantes-France), consisting of 26 classes of uppercase characters from Form B [14]. The IRONOFF database was selected because it is fully cursive. Samples were collected from about 700 writers, mainly of French nationality. The off-line data were scanned at 300 dpi with 8 bits per pixel.

To develop the word recognition system it was initially necessary to construct a database that can represent the different handwriting styles present in Brazilian Portuguese language. This was done by collecting samples of each month name, from 500 writers of different levels of education. Each writer was asked to fill a specific form where the word corresponding to each month name would be written once, as presented in Fig. 4 [12].

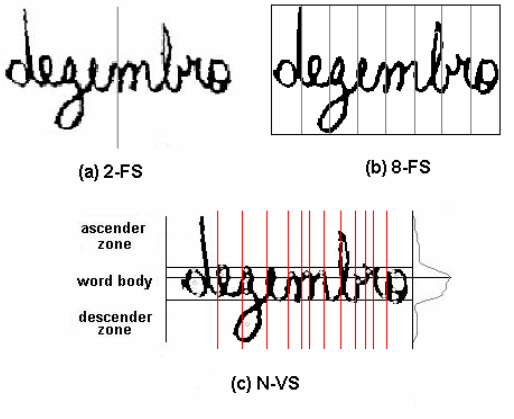


Fig. 3. Example of each zoning mechanism

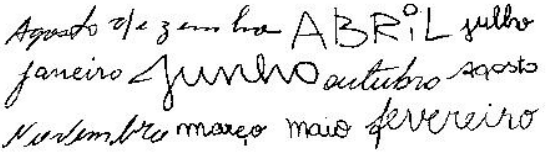


Fig. 4. Sample images from the word database

The databases have a total of 10,510 images of characters and 6,000 images of words, respectively. We split both databases into three sub-sets: Training (60%), Validation (20%), and Test (20%).

5.2 Results

For each zoning Z in the character recognition system presented in Section 3 and for each 2-FS and 8-FS feature set presented in Section 4, one NN was trained and tested, as proposed by Oh et al. [6]. The class decision module considers only the O_0 outputs from each sub-network and uses a simple winner-takes-all scheme to determine the final class.

The N-VS classifier for word recognition was evaluated with the same sets used for the other classifiers and for each class one model was trained and validated. The model that assigns maximum probability to one test image represents the class recognized.

The computational time for the experiments is beyond the scope of this paper. Both systems account for the entire recognition process (preprocessing, feature extraction, recognition, DbD calculation). As described in Section 2.2, the DbD measure is computed from the confusion matrices, therefore after the recognition stage.

Tables 1 and 2 show the results obtained for each zoning scheme both for the character and the word recognition systems, respectively. It can be seen that the best results were obtained using Z=7 for character recognition and 8-FS with directional features for word recognition.

Table 1. Recognition rate for each classifier in character recognition system

Classifier	Recognition Rate (%)
Z=4	83.2
Z=5Vertical	82.4
Z=5Horizontal	84.7
Z=7	**88.9**

Table 2. Recognition rate for each classifier in word recognition system

Classifier	Recognition Rate (%)
2-FS	73.9
8-FS-P (Perceptual)	86.3
8-FS-D (Directional)	**91.4**
8-FS-T (Topological)	85.0
N-VS	81.7

The DbD approach described in Section 2 was applied to the classifiers confusion matrices, for both systems. Distance was calculated considering groups of 2 classifiers. The calculated disagreement measures were compared with the results obtained combining the classifiers outputs with the weighted sum rule (WSC) defined by Kittler et al. [15], as presented in Tables 3 and 4. Other combination rules (average and product) were also tested but the best results were obtained using WSC. The boldfaced entries in the tables correspond to the median diversity ensembles, obtained by the DbD methodology, and to the best recognition rate associated to them.

Tables 3 and 4 show that, in general, the best combination results produced by the WSC rule correspond to one of the median diversity (DbD) values. One exception is the 4-7 combination (last line of Table 3), that achieved the best recognition score (85.8%) with the largest (5.28) DbD value, for the character recognition case. The reason for this deviation is not yet clear (shall be further investigated), however, it does not invalidate the general observed behavior, that recognition rates decay as classifiers diversity moves away from the median values.

Table 3. Classifier combination using DbD – Character recognition system

Character Classifier	DbD	WSC (%)
5H-5V	4.78	83.9
4-5H	4.87	83.4
4-5V	**5.01**	**83.7**
5V-7	**5.13**	**85.2**
5H-7	5.22	85.1
4-7	5.28	85.8

Table 4. Classifier combination using DbD – Word recognition system

Word Classifier	DbD	WSC (%)
8-FS-D – 8-FS-T	1.30	91.7
8-FS-P – 8-FS-T	1.33	89.4
8-FS-P – 8-FS-D	1.34	92.9
8-FS-P – N-VS	1.44	93.2
8-FS-D – N-VS	**1.77**	**95.0**
8-FS-T – N-VS	**1.88**	**93.4**
2-FS – N-VS	2.07	89.9
2-FS – 8-FS-P	2.28	90.3
2-FS – 8-FS-T	2.28	88.9
2-FS – 8-FS-D	2.57	93.9

The DbD measure has also been applied to experiments using three different classifiers for character recognition systems. Preliminary results obtained are shown in Table 5. We can observe that the best results correspond to the diversity (DbD) values close to the median.

Table 5. Classifier combination using DbD– Three different classifiers

Character Classifier	DbD	Recognition Rate (%)
4-5H-5V	27.49	85.8
4-5V-7	**27.52**	**90.9**
Median	27.70	------
4-5H-7	**27,88**	**91,0**
5H-5V-7	29,64	90,1

6 Discussions and Conclusion

This paper proposes an approach to evaluate ensembles of classifiers without actually having to exhaustively combining them to measure classification performance. The main motivation for this is the high computational cost of performing an exhaustive search in classifier combination space when we have a large number of classifiers. The DbD approach uses information from the confusion matrices of each individual classifier to compute distances that represent classifiers diversity. The results obtained with this approach reinforce the idea that median, rather than high, diversity is in general synonymous with high accuracy. The proposed methodology thus constitutes a new method for *a priori* evaluating multiple classifier systems, indicating the strongest candidates to the best classifiers combination. Using this approach, the search space is drastically reduced, in general to two strong candidates, which can then have their performance evaluated to determine the best MCS for the problem at hand. In this work, the method was applied to two handwriting recognition problems, although it can in fact be used for any pattern recognition problem. The validity of the DbD approach is supported by the experimental results. Future work will focus on the analysis of different distance criteria and the application of the DbD measure for ensembles of more than two classifiers, for the word recognition system.

References

1. Windeatt, T.: Diversity Measures for Multiple Classifier System Analysis and Design. Information Fusion 6(1), 21–36 (2005)
2. Zouari, H.K.: Contribution à l'évaluation des méthodes de combinaison parallèle de classifieurs par simulation. Doctor Thesis, Université de Rouen (2004)
3. Duin, R.P.W., Pekalska, E., Tax, D.M.J.: The Characterization of Classification Problems by Classifier Disagreements. In: ICPR 2004, vol. 1, pp. 140–143 (2004)
4. Kuncheva, L.I., Whitaker, C.J.: Measures of Diversity in Classifier Ensembles. Machine Learning 51, 181–207 (2003)
5. Hadjitodorov, S.T., Kuncheva, L.I., Todorova, L.P.: Moderate Diversity for Better Cluster Ensembles (2005), http://www.informatics.bangor.ac.uk/kuncheva/-recent_publications.htm
6. Oh, I.-S., Suen, C.Y.: A Class-Modular Feedforward Neural Network for Handwriting Recognition. Pattern Recognition 35(1), 229–244 (2002)
7. Parker, J.R.: Algorithms for Image Processing and Computer Vision. Jonh Wiley & Sons, Chichester (1997)
8. Suen, C.Y., Guo, J., Li, Z.C.: Analysis and Recognition of Alphanumeric Handprints by Parts. IEEE Trans. on Systems, Man and Cybernetics 24(4), 614–631 (1994)

9. Li, Z.C., Suen, C.Y., Guo, J.: A Regional Decomposition Method for Recognizing Handprinted Characters. IEEE Trans. on Systems, Man and Cybernetics 25(6), 998–1010 (1995)
10. Freitas, C.O.A., Oliveira, L.E.S., Bortolozzi, F., Aires, S.B.K.: Handwritten Character Recognition Using Nonsymmetrical Perceptual Zoning. International Journal of Pattern Recognition and Artificial Intelligence, IJPRAI 21(1), 135–155 (2007)
11. Rabiner, L., Juang, B.H.: Fundamental of Speech Recognition. Prentice-Hall, Englewood Cliffs (1993)
12. Oliveira, Jr.,J.J., Kapp, M.N., Freitas, C.O.A., Carvalho, J.M., Sabourin, R.: Handwritten month word recognition using multiple classifiers. In: XVII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), vol. 1, pp. 82–89 (2004)
13. Freitas, C.O.A., Bortolozzi, F., Sabourin, R.: Study of Perceptual Similarity Between Differnet Lexicons. International Journal of Pattern Recognition and Artificial Intelligence, IJPRAI 18(7), 1321–1338 (2004)
14. Viard-Gaudin, C.: The Ironoff User Manual. IRESTE, University of Nantes, France (1999)
15. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On Combining Classifiers. IEEE Trans. on PAMI 20(3), 226–239 (1998)