Improving Writer Identification Through Writer Selection

Diego Bertolini [1], Luiz S. Oliveira[2], and Robert Sabourin[3]

[1]Universidade Tecnologica Federal do Paraná, Campo Mourão (PR), Brazil
[2]Universidade Federal do Paraná, Curitiba (PR), Brazil
[3]École de Tecnologie Supérieure, Montreal, Canada

Abstract. In this work we present a method for selecting instances for a writer identification system underpinned on the dissimilarity representation and a holistic representation based on texture. The proposed method is based on a genetic algorithm that surpasses the limitations imposed by large training sets by selecting writers instead of instances. To show the efficiency of the proposed method, we have performed experiments on three different databases (BFL, IAM, and Firemaker) where we can observe not only a reduction of about 50% in the number of writers necessary to build the dissimilarity model but also a gain in terms of identification rate. Comparing the writer selection with the traditional instance selection, we could observe that both strategies produce similar results but the former converges about three times faster.

1 Introduction

The concept of dissimilarity [9] has been used successfully to deal with several pattern recognition problems. In the case of writer identification [2] and signature verification [1], the dissimilarity-based classifiers using a dichotomy transformation have been proved a good alternative since i) they can deal with a large number of classes by reducing any pattern recognition problem to a 2-class pattern, ii) the ability of using disjoint sets for training and testing, and iii) the model is scalable in the sense that we do not need to train it each time a new class (writer) is enrolled into the system. In this approach the feature vectors are extracted from both questioned and reference samples and then the dissimilarity feature vectors are computed. In ideal conditions, if both samples come from the same writer (genuine), then all the components of such a vector should be close to 0, otherwise (forgery), the components should be far from 0. Figure 1 illustrates this transformation.

The difference vectors plotted in Figure 1b compose then the training set T that will be used to train the 2-class classifier. As pointed out in the literature [6], in practice, T contains useless information for the classification task (that is, superfluous instances which can be noisy or redundant) therefore a process to discard them from T is needed. This process is known as instance selection [6] or prototype selection [5].

Similarly to feature selection, instance selection algorithms can also be classified into two categories based on whether or not instance selection is performed independently of the learning algorithm used to construct the classifier. If instance selection is done independently of the learning algorithm, the technique is said to follow a filter approach. Otherwise, it is said to follow a wrapper approach. The literature shows that the wrapper approach produce better results, however, it involves the computational overhead of evaluating candidate instance subsets by executing a given learning algorithm on the database using each instance subset under consideration.

The literature shows different strategies for wrapping a classifier into the instance selection process. Several of them formulate the problem as a search problem using different algorithms, such as Tabu Search [11], Sequential Floating Search [7], and Genetic Algorithms [3]. In the case of Genetic Algorithms, the straightforward approach consists in using a binary coded chromosome where the size of the chromosome is the number of instances available for training. This strategy produce good results but it shows its limits as the number of instances available for training gets larger.

In this work we deal with instance selection for writer identification using the dissimilarity representation. To overcome the limitations imposed by large training sets, we proposed selecting writers instead of instances. In other words, if a given writer is not selected, all his instances are removed from the training set. In the proposed method, the selection takes place before the dichotomy transformation, therefore only the instances of the selected writers are used to build the dissimilarity space.

To assess the proposed method we have used the writer identification system described in [2]. In this system the handwriting is first transformed into a texture and then different descriptors are used to generate the vectors in the feature space. Then, it applies the dichotomy transformation to create the dissimilarity representation where a SVM classifier is trained to discriminate between positive (writer) and negative (not writer) classes. Through a set of comprehensive set of experiments on three different databases, using a classifier trained with two textural descriptors (LBP and LPQ), we show that the proposed writer selection method is able to reduce considerably the number of writers necessary to build the dissimilarity model, in about 50%, while improving the identification rates. We also show that the performance of the selection mechanism is related to the number of references available for training and testing. Results show that when few references are available, which is true in most of real problems, the writer selection process appears to be more relevant. Finally, we compare the proposed approach with instance selection and show that it converges much faster producing similar results.

2 The Writer Identification System

As stated before, the method introduced in [2] was used to assess the proposed method. For the sake of clarity we reproduce it in this section. Given a queried

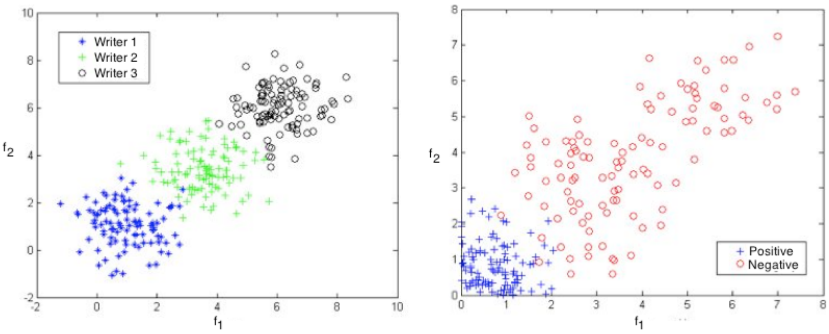

Fig. 1. The dichotomy transformation: (a) three classes in the feature space and (b) two classes in the dissimilarity space

handwritten document and a reference handwritten document, the aim is to determine whether or not the two documents were produced by the same writer. Let V and Q be two vectors in the feature space, labeled l_V and l_Q respectively. Let Z be the dissimilarity feature vector resulting from the dichotomy transformation $Z = |V - Q|$, where $|\cdot|$ is the absolute value. This dissimilarity feature vector has the same dimensionality as V and Q.

In the dissimilarity space, there are two classes that are independent of the number of writers: the within class (+) and the between class (−). The dissimilarity vector Z is assigned the label l_Z,

$$l_Z = \begin{cases} + \text{ if } l_V = l_Q, \\ - \text{ otherwise} \end{cases} \tag{1}$$

Figure 1 illustrates this transformation. Suppose there are three writers, $\{\omega_1, \omega_2, \omega_3\}$, and each one of them provides some samples. The feature extraction process extracts a vector from each sample, and these are shown in Figure 1a. Then, a dichotomy transformation takes place and computes the dissimilarity between the features of each pair of samples to form vectors. The distribution of such vectors, which we call dissimilarity feature vectors, are shown in Figure 1b.

We can see in Figure 1 that the dichotomy transformation affects the geometry of the distribution. In the feature space, multiple boundaries are needed to separate all the writers. In the dissimilarity space, by contrast, only one boundary is necessary, since the problem is reduced to a 2-class classification problem. The number of samples in the dissimilarity space is larger, because these samples are made up of every pair of feature vectors. We can also see in Figure 1 that, if both samples come from the same writer (genuine), then all the components of such a vector should be close to 0, otherwise they come from different writers (a forgery), in which case the components should be far from 0. This is true under favorable conditions. However, as in any other feature representation, the dissimilarity feature vector can be affected by intra-writer variability. This variability could generate values that are far from zero, even when the dissimilarity between the samples produced by the same writer is measured.

As mentioned earlier, one advantage of this approach is that even writers whose specimens were not used for training can be identified by the system. This characteristic is quite attractive, since it obviates the need to train a new model every time a new writer is introduced. In our experiments, we emphasize this feature by using disjoint sets of writers for training and testing.

The framework underpinning the identification system works as follows. Initially, a handwritten document is converted to a texture image. Then, the texture is split into n equal parts, $R_i(i = 1, 2, \dots, n)$, which are sent to the feature extraction module. The resulting feature vectors, V_i, are stored in a database. When a queried handwritten document is presented to the system, it is also converted to a texture and split into m equal parts, $S_i(i = 1, 2, \dots, m)$. These m textures undergo the same feature extraction process, and so creating the feature vectors Q_i. Then, the dissimilarity feature vectors $Z_i = |Vi - Qi|$ are computed and sent to the SVM classifier, which yields a decision on each dissimilarity feature vector. The final decision, is based on combining these partial decisions, and is obtained by means of a fusion rule.

The dissimilarity framework requires the classifiers to discriminate between genuine (positive) and forgeries (negative). To generate the positive samples to train the SVM classifier, we computed the dissimilarity vectors among the R genuine samples (references) of each writer which resulted in $\binom{R}{2}$ different combinations. The same number of negative samples is generated by computing the dissimilarity between one reference of one writer against one reference of other writers picked at random. In this work we assess the impact of the number of references per writer in the writer selection process.

2.1 Feature Extraction

In order to generate the texture, the document is binarized and scanned top-down and left-right to detect all the connected components of the image. The bounding box of the remaining components is then used to extract the original components of the gray level image. The components in gray levels are then aligned with the new image using the center of mass of the bounding box. This algorithm, described in details in [2], compacts the handwriting generating texture images. Then, the texture is segmented into nine 256 × 256 blocks. Figure 2 shows two examples of the handwriting texture produced by two different writers.

After creating the textural fragments, the next step consists in dealing with representation. The literature shows us a long story of research in texture representation but recent works have shown that Local Binary Pattern (LBP) and Local Phase Quantization (LPQ) appear to be a very interesting alternatives to represent texture. They have been successfully applied to different problems achieving promising results. Besides, they are quite easy to implement. In this work, we have used the traditional LBP configuration, i.e., $\mathrm{LBP}^{u2}_{8,2}$ with 59 components and the 256 dimensional feature vector produced by the LPQ.

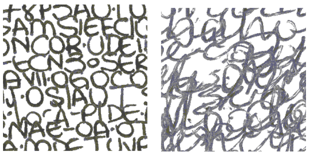

Fig. 2. Examples of handwriting textures from two different writers.

3 The Writer Selection Method

The selection mechanism is based on a Genetic Algorithm (GA) with bit representation, one-point crossover, bit-flip mutation, stochastic uniform selection, and elitism which is implemented using a generational procedure. The following parameter setting were employed: population size: 20, number of generations: 100, probability of crossover: 0.8, and probability of mutation: 0.01. In order to define the probabilities of crossover and mutation, we have used the one-max problem, which is probably the most frequently-used test function in research on genetic algorithms because of its simplicity. This function measures the fitness of an individual as the number of bits set to one on the chromosome. The population size was determined through experimentation.

Let $A = W_1, W_2, \ldots, W_n$ be the pool of n writers and B a chromosome of size n of the population. The relationship between A and B is straightforward, i.e., the gene i of the chromosome B is represented by the writer W_i from A. Thus, if a chromosome has all bits selected, all writers of A will be used to train the classifier.

In this work we have adopted a wrapper approach where each solution created by the genetic algorithm is a SVM classifier trained on the training set (TR) using 5-fold cross validation. It uses a Gaussian kernel and the parameters C and γ are determined through a grid search. After training, the solution is assessed on the validation set (VAL#1) where we compute the Equal Error Rate (EER), which is given by Equation 2

$$\text{Equal Error Rate} = \frac{FP + FN}{TP + TN + FP + FN} \tag{2}$$

where FP, FN, TP, and TN stand for False Positive, False Negative, True Positive, and True Negative, respectively. The minimisation of the ERR is the measure of fitness of the genetic algorithm.

During the search the performance of the classifier on VAL#1 may continue to improve, but its performance on a independent validation set will only improve to a point, where the classifier start to overfit VAL#1. To avoid this overfitting, a second validation set (VAL#2) is used to monitor and stop the evolutionary algorithm. Finally the solution the minimises the ERR on VAL#2 is evaluated in the the independent testing set (TS).

4 Experimental Results

In order to build a reliable experimental protocol to assess the proposed selection method we have performed the same experiments on the three aforementioned databases using two different textural descriptors, LBP and LPQ. As described in Section 2, the dissimilarity-based system uses a certain number of references for training (R) and testing (S). To produce a final decision, the system combines all partial decisions produced on S using a fusion rule. Based on previous experiments [2], the fusion rule that provides the best results is the Sum rule.

To show the reproducibility of the proposed strategy, three databases were considered in this work, the Brazilian Forensic Letter (BFL) database (Freitas et al. [4]), the IAM database (Marti and Bunke, [8]), and the Firemaker database [10]. To meet the requirements of the proposed method, all databases were divided into four independent partitions, i.e., training (TR), validation 1 (VAL#1), validation 2 (VAL#2), and testing (TS). The BFL database was divided into four subsets: 25, 60, 60, and 115 writers for TR, VAL#1, VAL#2, and TS, respectively. Each writer is represented by 9 blocks of texture (256 × 256 pixels). The IAM database was divided into four subsets: 50, 125, 125, and 240 writers for TR, VAL#1, VAL#2, and TS, respectively. Each writer is represented by 9 blocks of 256 × 128 pixels. Finally, the Firemaker dataset was divided into four subsets: 20, 45, 45, and 90 writers for TR, VAL#1, VAL#2, and TS, respectively. Each writer is represented by 9 blocks of texture of 256 × 256 pixels.

Our previous experiments also show that the best results were achieved when the number of references available for training and testing are maximised, i.e., $R, S = 9$. One aspect we analyse in these experiments is the impact of the number of references in the writer selection process. To that end, we have performed the experiments using $R, S = [3, 5, 9]$. All experiments were performed three times so the identification rates are the average of three runs. The four partitions of the databases were randomly generated for each experiment.

Table 1 compares the identification rates on the BFL database. It shows that the proposed method is able to reduce in about 50% the number of writers in all scenarios. In the case of the classifier trained with LBP using few references (R,S=3), besides reducing the number of writers the writer selection method also brought an important gain in terms of performance, about 5 percentage points.

Table 1. Results on the BFL database

	Ref.	With Writer Selection				Without Writer Selection	
	R,S	%	σ	# of writers	σ	%	# of writers
LBP	3	95.1	0.01	9.6	2.3	89.5	25
	5	95.7	0.02	12	1.0	94.7	25
	9	98.0	0.01	14	4.3	99.8	25
LPQ	3	95.5	0.01	11.3	0.6	96.5	25
	5	98.3	0.01	13	3.5	99.1	25
	9	99.4	0.01	14	1.0	99.0	25

In the second experiment we have applied the same protocol in a bigger database, the IAM. Table 2 shows the results for IAM database where we can observe a similar behavior, i.e., reduction of the number of writers in about 50% and improvement in terms of identification rate.

Table 2. Results on the IAM database

	Ref.	With Writer Selection				Without Writer Selection	
	R,S	%	σ	# of writers	σ	%	# of writers
	3	68.2	0.03	26	2.6	60.0	50
LBP	5	76.5	0.02	25.6	1.38	74.0	50
	9	91.3	0.01	28.6	1.1	91.0	50
	3	77.5	0.08	26.3	1.1	75.0	50
LPQ	5	81.8	0.02	22	1.7	77.0	50
	9	93.1	0.01	27.3	1.15	92.0	50

In the third experiments we have considered the Firemaker database, which contains different handwriting styles such as upper-case and copied text and also forgeries. The results for the Firemaker database are reported in Table 3 and follows the same pattern exhibited by the experiments on BFL database. The number of writers was reduced in about 50% and the performance was improved when few references were available.

Table 3. Results on the Firemaker database

	Ref.	With Writer Selection				Without Writer Selection	
	R,S	%	σ	# of writers	σ	%	# of writers
	3	96.7	0.02	10.3	1.5	94.4	20
LBP	5	91.9	0.01	9.3	1.5	91.1	20
	9	96.7	0.01	12.3	0.5	97.7	20
	3	98.1	0.01	8.6	0.5	96.6	20
LPQ	5	98.9	0.01	11.6	4.9	96.6	20
	9	97.8	0.02	10.3	1.5	98.8	20

A final experiment was performed to compare the results of the proposed method with the traditional instance selection approach. In this case we have used the BFL database for R = 9 and the LPQ-based classifier. The size of the chromosome is the number of instances available per writer. Considering the 25 writers times 9 references, the size of the chromosome is 225.

Table 4. Instance versus writer selection on the BFL database for R = 9 and LPQ

Method	Rec. Rate (%)	writers	instances	time (seconds)
Writer Selection	99.4	14	117	84021
Instance Selection	99.4	25	113	215680

The consequence of dealing with a large search space is a higher computational overhead. Table 4 shows that writer and instance selection arrives to very similar solution, in terms of performance and number of instances, but the proposed strategy uses a considerably reduced amount of time.

5 Conclusion

In this paper we have discussed a method for selecting instances for a writer identification system underpinned on the dissimilarity representation and a holistic representation based on texture. The proposed method is based on a genetic algorithm that surpasses the limitations imposed by large training sets by selecting writers instead of instances.

Our experiments on different databases show that the proposed method is able to reduce in about 50% the number of writers necessary to build the dissimilarity model while improving the identification rates. Comparing the writer selection with the traditional instance selection, we could observe that both strategies produce similar results but the former converges about three times faster.

References

1. D. Bertolini, L. S. Oliveira, E. Justino, and R. Sabourin. Reducing forgeries in writer-independent off-line signature verification through ensemble of classifiers. *Pattern Recognition*, 43(1):387–396, 2010.
2. D. Bertolini, L. S. Oliveira, E. Justino, and R. Sabourin. Texture-based descriptors for writer identification and verification. *Expert Systems With Applications*, 40(6):2069–2080, 2013.
3. J. C. Bezdek and L. Kuncheva. Nearest prototype classifier designs: an experimental study. *nt J Hybrid Intell International Journal of Hybrid Intelligent Systems*, 16(12):1445–14473, 2001.
4. C. Freitas, L.S. Oliveira, R. Sabourin, and F. Bortolozzi. Brazilian forensic letter database. In *11th Int. Workshop on Frontiers on Handwriting Recognition*, 2008.
5. S. Garcia, J. Derrac, J. R. Cano, and F. Herrera. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Trans. on Pattern Anaysis and Machine Intelligence*, 34(3):417, 435 2012.
6. J. Lopes, J. Ochoa, J. Trinidad, and J. Kittler. A review of instance selection methods. *Artificial Intelligence Review*, 34:133–143, 2010.
7. J. Lopes, J. Trinidad, J. Ochoa, and J. Kittler. Prototype selection based on sequeintial search. *Intelligent Data Analysis*, 13(4):599–631, 2009.
8. U. V. Marti and H. Bunke. The IAM-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46, 2002.
9. E. Pekalska and R. P. W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition*, 23:943–956, 2002.
10. L. Schomaker and L. Vuurpijl. Forensic writer identification: A benchmark data set and a comparison of two systems. Technical report, Nijmegen, February 2000.
11. H. Zhang and G. Sun. Optimal reference subset selection for nearest neighbor classification by tabu search. *Pattern Recognition*, 35:1481–1490, 2002.