A Methodology for Handwritten Character Recognition Using SVM

Mohammad Shayganfar [1,2], Mehdi Ghayoumi [1,2], Delbar Jafarpour [3], Parisa Shahamat
[1] Islamic Azad University Shahr-e-Ray Branch, Tehran, Iran
[2] Young Researchers Club, Islamic Azad University(Shahr-e-Ray Branch), Tehran, Iran
[3] Luleå Tekniska Universitet, Luleå, Sweden

Abstract - *This paper discusses a methodology for handwritten character recognition applying feature subset selection to reduce number of features. Its novelty lies in the use of a genetic algorithm for the preparation of input data for a support vector machine which is employed to recognize the handwritten Persian digits in particular. Comprehensive experiments on handwritten Persian digits demonstrate the capability of such a classifier running on reduced extracted features.*

Keywords: Character Recognition, SVM, PCA, GA, OWA.

1 Introduction

An important issue in constructing classifiers is the selection of the best discriminative features. In many applications, it is not unusual to find problems involving hundreds of features. However, it has been observed that beyond a certain point, the inclusion of additional features leads to a worse rather than better performance. Moreover, the choice of features to represent the patterns affects several aspects of the pattern recognition problem such as accuracy, required learning time, and the necessary number of samples [1].

In the context of practical applications such as handwriting recognition, feature selection presents a multicriterion optimization function, e.g. number of features and accuracy of classification. Genetic algorithms offer a particularly attractive approach to solve this kind of problems since they are generally quite effective in rapid global search of large, nonlinear and poorly understood spaces.

In this work we discuss the use of multiobjective genetic algorithms as a mean to search for subsets of features, which contain discriminatory information to classify handwritten digits. This is based on the fact that several studies in the literature have been demonstrated that genetic algorithms would be more effective than other methods when dealing with large-scale feature selection (i.e. more than 50 features). For those readers interested in comparative studies, please see References [2, 3 and 4].

Nowadays Support Vector Machines (SVM) becomes one of the most popular classification methods. They have been used in a wide variety of applications such as text classification [5], facial expression recognition [6], Character recognition [7, 8] and many others. The theoretical foundation of this method is given by statistical learning theory [9].

Principle Component Analysis (PCA) is a useful statistical technique that has found application in fields such as character recognition and image compression, and is a common technique for finding patterns in data of high dimension.

This paper is structured as follows. Section 2 demonstrates the overview of the handwritten character recognition system; Section 3 summarizes our approach to select features subset. Section 4 presents the genetic algorithm based feature selection. Section 5 describes overview of PCA and its outcome on classifier systems, Section 6 describes the application of a support vector machine as a classifier, Section 7 reports the results of experiments, and finally Section 8 presents some discussion and concludes this paper (Microsoft Word version 6.0 or later).

2 Character recognition structure

Figure 1 shows the proposed computational model of our Persian handwritten digit recognition system. In this diagram, an image of a Persian digit will be given as an input image. The input image will be processed at the next step extracting all the required features (Total features = 132). There are three different techniques applied to get the features, concavity of the character, character's contour and character surface which is simply the number of black pixels in the image. Subsequently, a genetic algorithm, as a filter, will reduce the number of features from feature vector of the input image using a selection mechanism. Afterwards, PCA will select effective features to obtain the best number of features which are necessary for classification process. At last, the selected features will be classified by SVM applying Gaussian, polynomial, and Radial Basis Function kernels. The output provides the recognized number according to the input image.

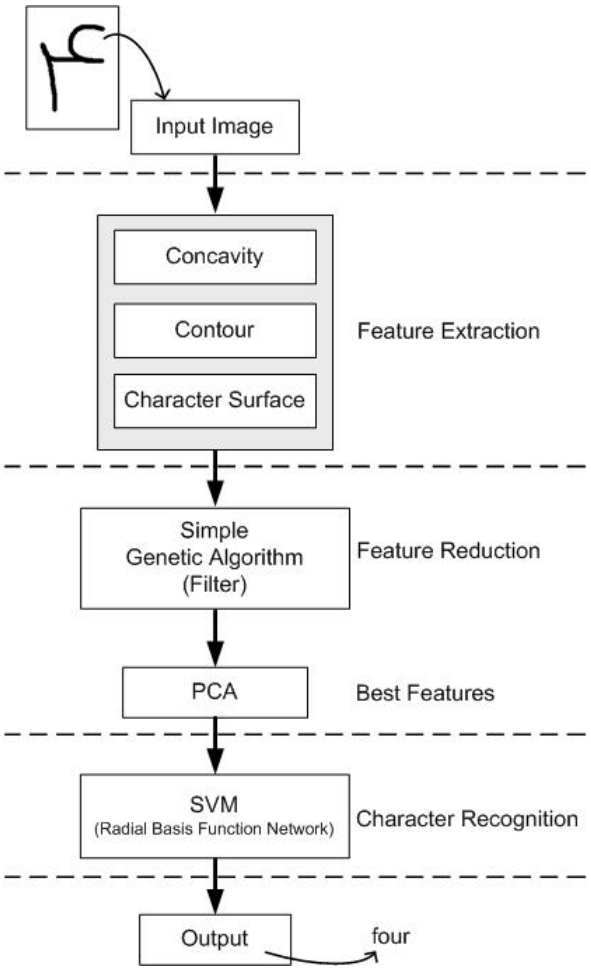

Fig.1. An overview of the Persian handwritten digit recognition system

component has two labels. The superior label means the number of black pixels found during the search while the inferior label means the directions where the black pixels were not reached. For example, the pixel x1 [Figure 2(b)] reaches the black pixel in all directions except in direction 1. Therefore, the position 6 of the feature vector is incremented. For pixel x2, the position 1 is incremented because it reaches the black pixel in four directions. However, using the auxiliary directions s1 - s4, we confirm that it is not inside a closed contour if previous step shows that the pixel is inside of a closed contour. When the pixel is inside a closed contour, the position incremented is the eighth [1].

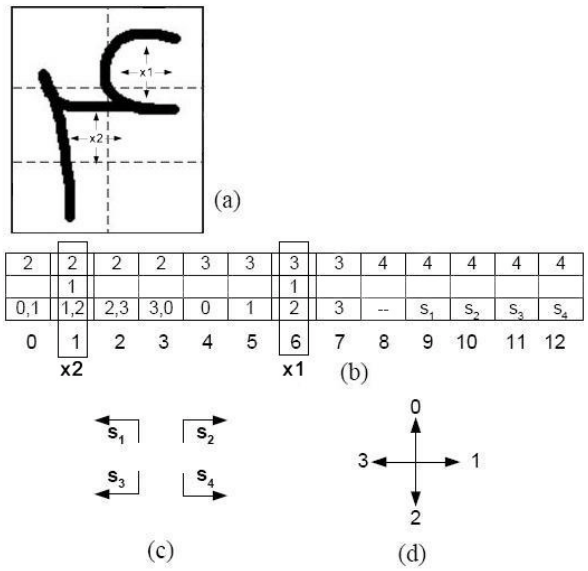

Fig.2. Concavity measurements: (a) feature vector, (b) concavities, (c) auxiliary directions and (d) 4-Freeman directions.

3 Feature extraction

Since the focus of most of our experiments is the recognition of the Persian digits based on feature set extracted from the image, we will describe it in this section. Such a feature vector is composed of a mixture of concavity and contour information [1]. The basic idea of concavity measurements is the following: for each white pixel in the component, we search in 4-Freeman directions [Figure 2], the number of black pixels that it can reach as well as which directions the black pixel is not reached. When black pixels are reached in all directions [e.g. point x1 in Figure 2(b)], we branch out in four auxiliary directions [s1 to s4 in Figure 2(c)] in order to confirm if the current white pixel is really inside a closed contour. Those pixels that reach just one black pixel are discarded.

Thereafter, we increment the position in the feature vector that fits with results returned by the search [Figures 2(a) and 2(b)]. In Figure 2(a), we represent the feature vector where each

Since we are dividing the image into six zones, we consider six feature vectors with 13 components each. Therefore, in the example presented above, the pixel x1 will update the second vector while the pixel x2 will update the third vector. Finally, the overall concavity feature vector is composed of (13 x 6) 78 components which are normalized between 0 and 1 by summing up their values and then dividing each one by this summation.

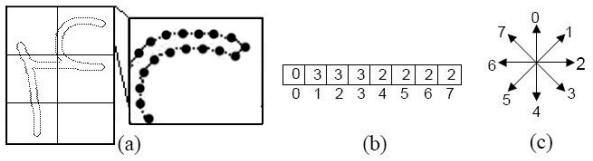

Fig.3. Contour measurements: (a) contour image of the upper right corner zone, (b) feature vector and (c) 8-Freeman directions.

The contour information is extracted from a histogram of contour directions. For each zone, the contour line segments between neighboring pixels are grouped regarding 8-Freeman directions [Figure 3(c)]. The number of line segments of each orientation is counted [Figure 3(b)]. Therefore, the contour feature vector is composed of (8 x 6) 48 components normalized between 0 and 1.

Finally, the last part of the feature vector is related to the surface of the character. We simply count the number of black pixels in each zone and normalize these values between 0 and 1. Thus, the final feature vector has (78 + 48 + 6) 132 components.

4 Feature selection

Here, we present the choice of a representation for encoding candidate solutions to be manipulated by the genetic algorithm.

Each individual in the population represents a candidate solution to the feature subset selection problem. Let m be the total number of features available to choose from to represent the patterns to be classifier (m = 132 in our case). The individual (chromosome) is represented by a binary vector of dimension m. If a bit is a 1, it means that the corresponding feature is selected, otherwise the feature is not selected. This is the simplest and most straightforward representation scheme [10].

Since we are representing a chromosome through a binary string, the operators mutation and crossover operates in the following way: Mutation operates on a single string and generally changes a bit at random. Thus, a string 11010 may, as a consequence of random mutation gets changed to 11110. Crossover on two parent strings to produce two offsprings. With a randomly chosen crossover position 4, the two strings 01101 and 11000 yield the offspring 01100 and 11001 as a result of crossover.

Our experiments used the following parameter settings:

Population size: 30

Number of generation: 1000

Probability of crossover: 0.8

Probability of mutation: 0.007

The parameter settings were based on results of several preliminary runs [11].

The selection mechanism emulates the survival-of-the-fittest mechanism in nature. It is expected that a fitter chromosome receives a higher number of offsprings and thus has a higher chance of surviving on the subsequent evolution while the weaker chromosomes will eventually die.

In this work we are using the roulette wheel selection [12] which is one of the most common and easy-to-implement selection mechanism.

The fitness evaluation is a mechanism used to determine the confidence level of the optimized solutions to the problem. Usually, there is a fitness value associated with each chromosome, e.g., in a minimization problem, a lower fitness value means that the chromosome or solution is more optimized to the problem while a higher value of fitness indicates a less optimized chromosome.

Our problem consists of optimizing two objectives: minimization of the number of features and minimization of the error rate of the classifier. Therefore, we are dealing with a multi-objective optimization problem. While in single-objective optimization the optimal solution is usually clearly defined, this does not hold for multi-objective optimization problem. Instead of a single optimum, there is rather a set of alternative trade-offs, generally known as Pareto-optimal solutions.

In order to generate the Pareto-optimal set, we are using a classical approach proposed by Hajela and Lin in [13], called weighting method, which aggregates the objectives into a single and parameterized objective. Such an aggregation is performed through a linear combination of the objectives

$$f(x) = f_1(x) \times \omega_1 + f_2(x) \times \omega_2 \qquad (1)$$

where ω_i are called weights and, without loss of generality, normalized such that $\sum \omega_i = 1$. $f_1(x)$ is the error rate produced by the classifier for a given feature subset (represented by the chromosome x) and $f_2(x)$ is the number of features selected in the chromosome x. Therefore, the fitness of a chromosome is represented by a single and parameterized objective function $f(x)$ [11].

5 Principal component analysis

PCA has been called one of the most valuable results from applied linear algebra. It is used abundantly in many forms of analysis from neuroscience to computer graphics. PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data, ie. by reducing the number of dimensions, without much loss of information. All the mathematical details on PCA could be found in [14].

SVM is one of best classifiers; however one of its problems is the dimension of its training data. Here we have used PCA to reduce dimension of features for training. The initial

number of features is 132 and we have projected it with PCA and selected 10 prior of them for training. In Figure 4 the initial data, its projection, and projection on GA output have been presented.

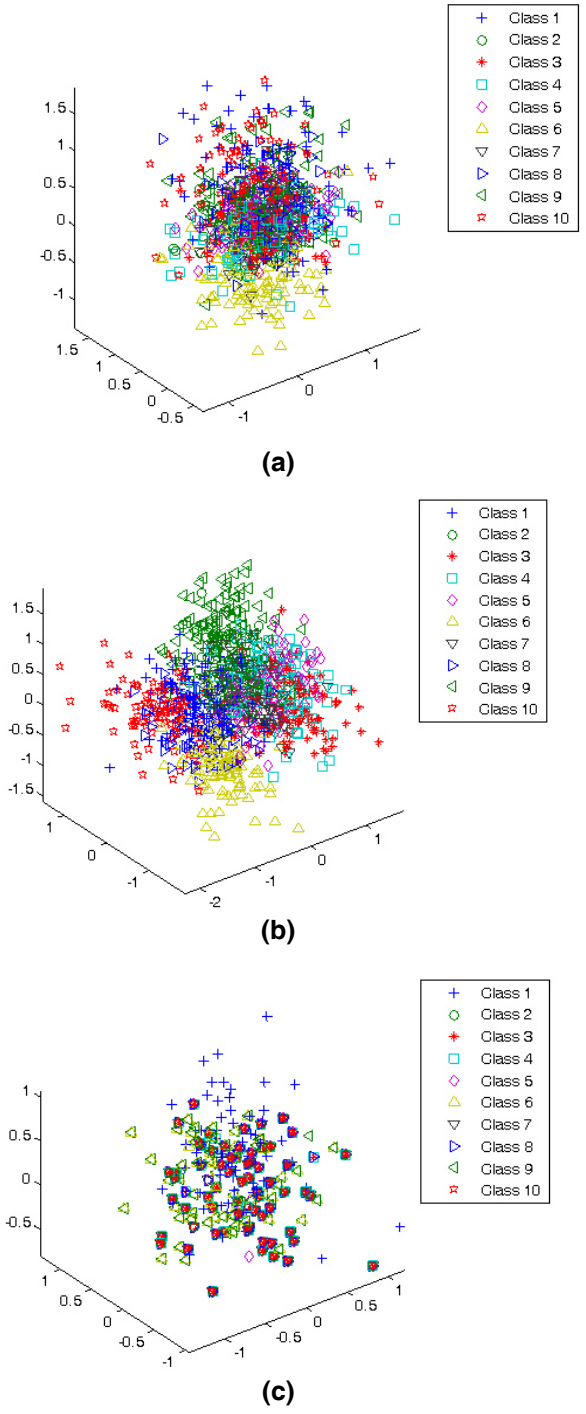

Fig.4. (a) Projection of initial data, (b)Projected data with PCA, (c) PCA projection of genetic algorithm's selected data.

6 Support vector machines

We have used a set of 10,000 images for training phase. It is mentioned that SVMs with three different kernels have been tested according to all the proven mathematical and statistical equalities in [15]. RBF model has 98% of precision on original data which a better performance is seen using affected features by GA. The polynomial SV machine however, does slightly better with 98.20% in the case of a polynomial kernel with degree 4 which also get a few improvements on reduced features by GA. The result with Gaussian kernel demonstrates about 98.70%, again better than the result of 97.60% on original data. It is worth to mention that the kernel parameters have been chosen in an empirical way in order to ensure good classification with respect to each kernel [16].

7 Experimental results

Performance of methods is evaluated on Persian digits database. The dimensionality of the training data is reduced to 10 prior for recognition.

Table .1 Classification results for Original Digits Database

Method	Mean of Margin	Mean of SV	Mean of Correctness
Gaussians	0.4390	245	97.60
Polynomial	0.4390	256	98.20
RBF	0.4390	248	98

Table 2 demonstrates the effect of feature reduction by GA on the classification results.

Table .2 Classification results for Projected Digits Database with GA

Method	Mean of Margin	Mean of SV	Mean of Correctness
Gaussians	0.4450	235	98.70
Polynomial	0.4450	243	98.80
RBF	0.4450	236	98.90

8 Conclusions and remarks

In this paper, we applied a methodology for handwritten character recognition employing feature subset selection to reduce the number of extracted features. The new method has the advantage of optimal selection of the features. A genetic algorithm based feature selection proceeds by providing of the eigenvectors of PCA method to be applied as initial data for training SVMs with three different kernels. Applying data fusion techniques, specifically OWA, is subject for the future studies according to our previous researches [17]

9 References

[1] L. S. Oliveira, R. Sabourin, F. Bortolozzi, C. Y. Suen, "A Methodology For Feature Selection Using Multiobjective Genetic Algorithms for Handwritten Digit String Recognition", International Journal of Pattern Recognition and Artificial Intelligence, World Scientific Publishing Company, 17, 6, 2003, pp. 903-929.

[2] F. J. Ferri, P. Pudil, M. Hatef and J. Kittler, "Comparative Study of Technique for Large-scale Feature Selection," Pattern Recognition in Practice, eds. J. E. Moody, S. J. Hanson and R. L. Lippmann, 4, Elsevier, 1994, pp. 403-413.

[3] M. Kudo and J. Sklansky, "Comparison of Algorithms That Select Features for Pattern Classifiers," Pattern Recognition. 33, 1, 2000, pp. 25-41.

[4] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large scale on feature selection," Pattern Recognition Letter 10 , 1989, pp. 335-347.

[5] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proceedings of the European Conference on Machine Learning, Springer, 1998.

[6] P. Michel and R. E. Kaliouby, "Real Time Facial Expresion Recognition in Video Using Support Vector Machines," Proceedings of ICMI'03, 2003, pp. 258-264.

[7] H. Miyao, M. Maruyama, Y. Nakano, T. Hananoi, "Off-line handwritten character recognition by SVM based on the virtual examples synthesized from on-line characters," Document Analysis and Recognition, 2005, pp. 494- 498.

[8] C. Hu, Y. Zhao, J. Wang, and Z. Yang, "An Improved Method for the Character Recognition based on SVM," Artificial Intelligence and Applications (AIA 2006), 2006 Innsbruck, Austria.

[9] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 2nd Edition, 1998.

[10] K.F.Man, K.S.Tang, and S.Kwong, *Genetic Algorithms:Concepts and Designs*, Springer-Verlag, London-UK, 1999.

[11] L. S. Oliveira, N. Benahmed, R. Sabourin, F. Bortolozzi, C. Y. Suen, "Feature Subset Selection Using Genetic Algorithms for Handwritten Digit Recognition", Computer Graphics and Image Processing, IEEE Computer Society, Washington DC USA, 2001, pp. 362-369.

[12] L. Davis., *Handbook on Genetic Algorithms*, Van Nostrand Reinhold, 1991.

[13] P.Hajela and C.Y.Lin, "Genetic search strategies in multicriterion optimal design," Structural and Multidisciplinary Optimization, 4, 2, 1992, pp. 99–107.

[14] I. T. Jolliffe, *Principal Component Analysis*, Springer, 2002.

[15] N. Cristianini, J. Shawe -Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.

[16] H. Abrishami Moghaddam and M. Ghayoumi, "Facial Image Feature Extraction Using Support Vector Machines," Proceedings of VISAPP, Setubal, Portugal, 2006.

[17] M. Shayganfar, B. Moshiri, C. Lucas, "Applying OWA Operator in a Multi-Agent Architecture of an Emotional Robot," International Summit of AI50, Switzerland, Monte Verita, 2006.