Unsupervised Feature Selection Using Multi-Objective Genetic Algorithms for Handwritten Word Recognition

M. Morita[1,2], R. Sabourin[1−3], F. Bortolozzi[3] and C. Y. Suen[2]
[1]École de Technologie Supérieure, Montreal, Canada
[2]Centre for Pattern Recognition and Machine Intelligence, Montreal, Canada
[3]Pontifícia Universidade Católica do Paraná, Curitiba, Brazil
e-mail: marisa@cenparmi.concordia.ca

Abstract

In this paper a methodology for feature selection in unsupervised learning is proposed. It makes use of a multi-objective genetic algorithm where the minimization of the number of features and a validity index that measures the quality of clusters have been used to guide the search towards the more discriminant features and the best number of clusters. The proposed strategy is evaluated using two synthetic data sets and then it is applied to handwritten month word recognition. Comprehensive experiments demonstrate the feasibility and efficiency of the proposed methodology.

1 Introduction

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. In this way, the selection of the best discriminative features plays an important role when constructing classifiers. However, this is not a trivial task especially when dealing with a lot of features. In order to choose a subset of the original features by reducing irrelevant and redundant ones automated feature selection algorithms have been used. The literature contains several studies on feature selection for supervised learning [7, 8]. But only recently, the feature selection for unsupervised learning has been investigated [3, 5].

The objective in unsupervised feature selection is to search for a subset of features that best uncovers "natural" groupings (clusters) from data according to some criterion. This is a difficult task because to find the subset of features that maximizes the performance criterion, the clusters have to be defined. The problem is made more difficult when the number of clusters is unknown beforehand which happens in most real-life situations. Hence, it is necessary to explore different numbers of clusters using traditional clustering methods such as the K-means algorithm [4] and its variants. In this light, clustering can become a trial-and-error work. Besides, its result may not be very promising especially when the number of clusters is large and not easy to estimate.

In the above context, feature selection presents a multi-criterion optimization function, e.g. the number of features and a validity index to measure the quality of the clusters. Genetic algorithm (GA) offers a particularly attractive approach to solve this kind of problems since they are generally quite effective in rapid global search of large, non-linear, and poorly understood spaces. In the last decade, GA has been largely applied to the feature selection problem. The approach often combines different optimization objectives into a single objective function. The main drawback of this kind of strategy lies in the difficulty of exploring different possibilities of trade-offs among objectives. In order to overcome this kind of problem, some authors [5] propose the use of a multi-objective genetic algorithm to perform feature selection.

In this paper we propose a methodology for feature selection in unsupervised learning for handwritten month word recognition (see Section 5). It makes use of the Non-dominated Sorting Genetic Algorithm (NSGA) proposed by Srinivas and Deb in [9] which deals with multi-objective optimization. The objective is to find a set of nondominant solutions which contain the more discriminant features and the more pertinent number of clusters. We have used two criteria to guide the search: minimization of the number of features and minimization of a validity index that measures the quality of clusters. A standard K-Means algorithm is applied to form the given number of clusters based on the selected features. The proposed strategy is assessed using two synthetic data sets where the significant features and the appropriate clusters in any given feature subspace are known. Afterwards, it is applied to handwritten month word recog-

nition in order to optimize the word classifiers. Experimental results show the efficiency of the proposed methodology.

2 Methodology for Feature Selection in Unsupervised Learning using NSGA

2.1 Objective Functions

As stated before, we have used two criteria: minimization of a validity index and minimization of the number of features.

In order to measure the quality of clusters, the within-cluster scatter and between-cluster separation have been widely used by various researchers. Kim et al in [5] make use of two objective functions to compute these measurements independently. Vesanto et al in [10] and Bandyopadhyay et al in [1] combine them in one objective function using the Davies-Bouldin (DB) index proposed by Davies et al in [2]. To make such indices suitable for our problem, they must be normalized by the number of selected features. This is due to the fact that they are based on geometric distance metrics and are therefore not directly applicable here because they are biased by the dimensionality of the space, which is variable in feature selection problems. In our experiments, we have considered the normalized DB index. Both criteria are described as follows.

2.1.1 DB Index

This index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The scatter within the i_{th} cluster is computed as follows

$$S_i = \frac{1}{|C_i|} \sum_{X \in C_i} \{||X - Z_i||\} \tag{1}$$

and the distance between clusters C_i and C_j is defined as

$$d_{ij} = \{||Z_i - Z_j||\} \tag{2}$$

S_i is the average Euclidean distance of the vectors X in cluster C_i with respect to its centroid Z_i. d_{ij} is the Euclidean distance between the centroids Z_i and Z_j of the clusters C_i and C_j respectively. Subsequently, we compute

$$R_i = \max_{j, j \neq i} \{\frac{S_i + S_j}{d_{ij}}\} \tag{3}$$

The DB index is then defined as

$$I_{DB} = \frac{1}{D} \frac{1}{K} \sum_{i=1}^{K} R_i \tag{4}$$

where K corresponds to the number of selected clusters and D is the number of selected features. The objective is to achieve proper clustering by minimizing the DB index.

2.1.2 Number of Features

We have observed that the value of DB index decreases as the number of features increases (see Section 3.2). We correlated this effect by the normalization of such an index by D. In order to compensate this we have considered as objective the minimization of the number of features. In this case, one feature must be set at least.

2.2 Implementation of NSGA

In our experiments, NSGA is based on bit representation (binary codification), one point crossover, bit-flip mutation and roulette wheel selection (with elitism). In such cases, the parameters of NSGA were tuned based on experimentation.

The idea behind NSGA is that a ranking selection method is used to emphasize good points and a niche method is used to maintain stable subpopulations of good points. It varies from simple GA only in the way the selection operator works. The crossover and mutation remain as usual. Before the selection is performed, the population is ranked on the basis of an individual's nondomination. The nondominated individuals present in the population are first identified from the current population. Then, all these individuals are assumed to constitute the first nondominated front in the population and assigned a large dummy fitness value. The same fitness value is assigned to give an equal reproductive potential to all these nondominated individuals. In order to maintain the diversity in the population, these classified individuals are made to share their dummy fitness values. Sharing is achieved by performing selection operation using degraded fitness values obtained by dividing the original fitness value of an individual by a quantity proportional to the number of individuals around it. After sharing, these nondominated individuals are ignored temporarily to process the rest of population in the same way to identify individuals for the second nondominated front. These new set of points are then assigned a new dummy fitness value which is kept smaller than the minimum shared dummy fitness of the previous front. This process is continued until the entire population is classified into several fronts.

The population is then reproduced according to the dummy fitness values. Since individuals in the first front have the maximum fitness value, they get more copies than the rest of the population. The efficiency of NSGA lies in the way multiple objectives are reduced to a dummy fitness function using nondominated sorting procedures. More details about NSGA can be found in [9].

Since the proposed strategy tries to find a set of solutions with the more discriminant features and a proper value of the number of clusters, each chromosome in the population encodes these two types of information. While the first positions encode the features, the remaining is devoted to

the number of clusters. In order to find high-quality solutions, we have considered two objectives: minimization of the number of features and minimization of the DB index.

Computing the first objective is simple, i.e., the bits equal to 1 in the first part of the chromosome provide the number of selected features. The second one is evaluated after performing clustering. In this case, a standard K-means algorithm is applied to form the clusters based on the selected features and the number of selected clusters, which is obtained by computing the bits equal to 1 in the second part of the chromosome.

3 Evaluation of the Methodology on Synthetic Data Sets

It is not easy to evaluate the quality of an unsupervised learning algorithm especially performing feature selection at the same time due to the fact that the clusters depend on the dimensionality of the selected features. In order to assess the proposed methodology and improve our insight about it, we carried out two experiments using two synthetic data sets, where the distributions of their points, the significant features, and the appropriate clusters in any given feature subspace are known. In such cases, we can evaluate the solutions found in the Pareto-optimal front by examining the selected features and the number of selected clusters as well.

3.1 Experiment I

The first synthetic data set has 300 points, two significant features in which the points form three well defined clusters. All clusters are formed by generating points from a pseudo-Gaussian distribution with standard deviation equal to 0.04. Figure 1 illustrates this data set.

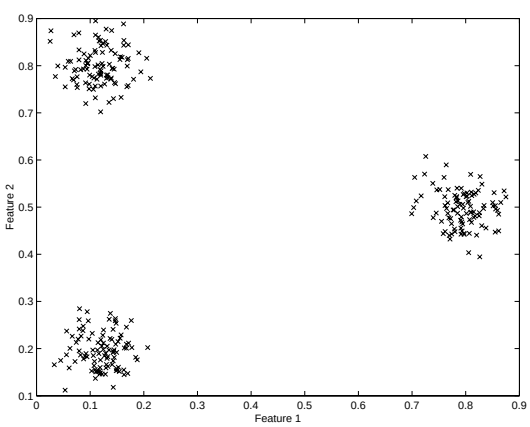


Figure 1. Data set of Experiment I

In this experiment, the chromosomes are composed of 12 bits, the first two bits encode the features, while the remaining (position 3 to 12) encode the number of clusters that can vary from 2 to 10. The cases of zero or one cluster are meaningless in this application. The NSGA parameters are: population size=20, number of generations=100, probability of crossover=0.8, probability of mutation=1/12, and the niche distance=0.4.

Table 1 shows the set of nondominant solutions found by NSGA. We can notice in this Table that both solutions describe the data set depicted in Figure 1 very well. The solutions S_1 and S_2 are extremely good along one of the two criteria. However, the solution with three clusters and one feature (feature 2) was not in the final population because it was dominated by the solution S_2 along DB index criterion.

Table 1. Solutions for Experiment I

Sol.	No. of Clusters	DB Index	No. of Features	Features
S_1	3	0.074246	2	1 and 2
S_2	2	0.086910	1	1

3.2 Experiment II

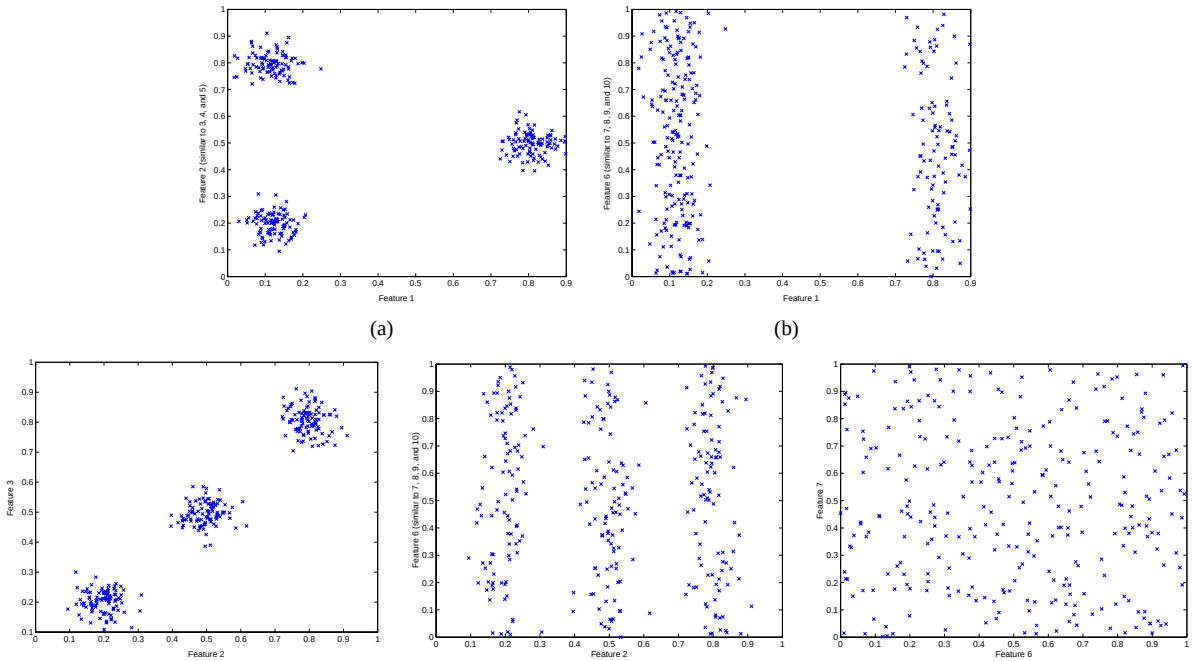
The second synthetic data set has 300 points and ten features and it is constructed as follows. Three clusters are formed along features 1 and 2 in the same way as the first data set. Features 3, 4, and 5 are similar to feature 2. Finally, for features 6, 7, 8, 9, and 10 the points are distributed uniformly. All the clusters are formed by generating points from a pseudo-Gaussian distribution with standard deviation equal to 0.04. Figure 2 illustrates this data set by projecting the points onto some of the feature subspaces with two dimensions.

The chromosomes are represented by 20 bits, the first ten bits encode the features, while the remaining (position 11 to 20) encode the number of clusters that can vary from 2 to 10. The NSGA parameters are: population size=20, number of generations=100, probability of crossover=0.8, probability of mutation=1/20, and the niche distance=0.4.

Table 2 shows the set of nondominant solutions found by NSGA. It can observe from this Table that the features 6 through 10 which have no significance were not selected. Besides, the solutions describe the data set depicted in Figure 2 very well. In this experiment, the interaction between our two optimization criteria can be visualized as discussed before in Section 2.1.2.

4 Proposed Methodology Applied to Handwritten Month Word Recognition

Regarding the results achieved in the experiments using the synthetic data sets, it can be concluded that the proposed



(a) (b) (c) (d) (e)

Figure 2. Some 2-dimensional projections of the data set of Experiment II

Table 2. Solutions for Experiment II

Sol.	No. of Clusters	DB Index	No. of Features	Features
S_1	3	0.037460	5	1, 2, 3, 4 and 5
S_2	3	0.043207	4	1, 2, 3 and 4
S_3	3	0.051891	3	1, 3, and 4
S_4	3	0.073910	2	1 and 4
S_5	3	0.206899	1	3

strategy is capable of identifying the more significant features and the more relevant number of clusters. The next step lies in evaluating its effect on month word recognition.

The word classifier used in this experiment is based on the word verifier of the date recognition system presented in [6]. In such a case, a word image is segmented into segments (graphemes), each of which consists of a correctly segmented, under-segmented, or and over-segmented character. Then, two feature sets are extracted from the sequence of graphemes to feed the classifiers. However, in order to better assess our approach we have considered only one feature set which is based on a mixture of segmentation primitives and concavity and contour features. Thus, for each grapheme a concavity and contour feature vector of 34 components is extracted. Since we are working with the discrete Hidden Markov Models (HMMs) and the feature vectors contain real values (low-level features), we must convert them into symbols (high-level features) by using a clustering technique. Instead of using a traditional strategy, which considers the entire feature set and tries exhaustively various number of clusters, we propose to use the foregoing methodology to find automatically a proper number of clusters and the more discriminant features.

Basically, the proposed methodology works as follows: NSGA produces automatically a set of nondominant solutions called Pareto-optimal which corresponds to the best trade-offs between the number of features and quality of clusters. Nevertheless, when applying such a strategy on month word recognition, one solution from Pareto-optimal front must be chosen to be used in the system. In order to perform this task, firstly we train each solution of the Pareto-optimal front to validate the the best solutions found by NSGA. Thereafter, such classifiers are used in the system and the solution that supplies the best word recognition result on the validation set is chosen. To evaluate the results achieved by our approach, we compare them with the results obtained from the traditional strategy.

5 Evaluation of the Methodology on Month Word Recognition

This section is devoted to the experiments conducted on a database in which we do not have knowledge

about the clusters and relevant features. This database contains 2,000 isolated images of handwritten Brazilian month words ("Janeiro", "Fevereiro", "Março", "Abril", "Maio", "Junho", "Julho", "Agosto", "Setembro", "Outubro", "Novembro", "Dezembro") and it was divided into three sets: 1,200, 400, and 400 images for training, validation, and testing respectively. In order to increase the training and validation sets, we have also considered 8,300 and 1,900 word images respectively extracted from the legal amount database. This is possible because we are considering character models. For clustering we have used about 80,000 feature vectors extracted from the training set of 8,300 words.

The chromosomes are represented by 184 bits, the first thirty four bits encode the features, while the remaining (position 35 to 184) encode the number of clusters that can vary from 2 to 150. The NSGA parameters are: population size=96, number of generations=1,000, probability of crossover=0.8, probability of mutation=1/184, and the niche distance=0.3.

Figure 3 illustrates the Pareto-optimal front found by NSGA and the selected solution S_2 (29 features and 36 clusters) which supplied the best word recognition result on the validation set (88.6%). In such a case, the recognition rate (zero-rejection level) on the test set was 86.0%. Besides, we can visualize in this graphic that the value of the DB index decreases as the number of features increases. The above recognition rates are very similar to the results reached using the traditional strategy that tries empirically various number of clusters without performing feature selection. In that case, the solution that brought better results was composed of 34 features and 80 clusters. It achieved 88.3% and 86.2% on validation and test sets respectively.

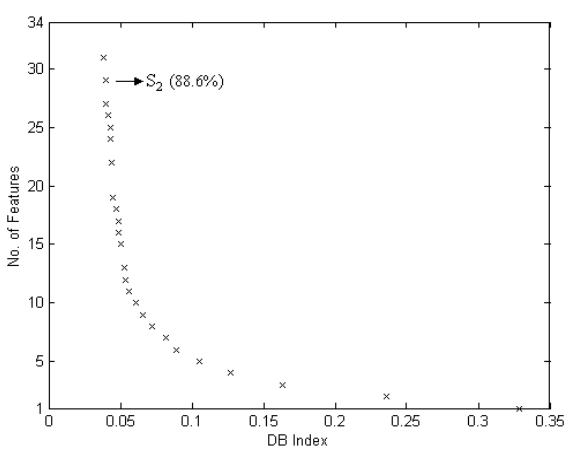


Figure 3. Pareto-optimal front

This confirms the efficiency of the proposed methodology in selecting a powerful subset of features and a proper value of the number of clusters. Besides, it reduced the number of features (from 34 to 29) and number of clusters (from 80 to 36) while keeping the recognition rates at the same level as the traditional strategy. Moreover, the time required for training the HMMs was decreased once the number of clusters was significantly reduced.

6 Conclusion

In this paper a methodology for feature selection in unsupervised learning based on multi-objective optimization has been presented. It generates a set of nondominant solutions called Pareto-optimal which corresponds to the best trade-offs between the number of features and quality of clusters. The proposed strategy was evaluated using two synthetic data sets and then applied to handwritten month word recognition in order to optimize the word classifiers. The results achieved show the efficiency of the proposed methodology where the number of features and clusters were reduced and the recognition rates were kept at the same level as the traditional strategy. Therefore, it can be successfully applied to the problem of feature selection in unsupervised learning.

References

[1] S. Bandyopadhyay and U. Maulik. Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recognition*, 35:1197–1208, 2002.

[2] D. L. Davies and W. Bouldin. A cluster separation measure. *IEEE PAMI*, 1:224–227, 1979.

[3] J. G. Gy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In *Proc. 17th International Conference on Machine Learning*, 2000.

[4] X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh Univ. Press, 1990.

[5] Y. S. Kim, W. N. Street, and F. Menczer. Feature selection in unsupervised learning via evolutionary search. In *Proc. 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 365–369, 2000.

[6] M. Morita, R. Sabourin, F. Bortolozzi, and C. Suen. Segmentation and recognition of handwritten dates. In *Proc. 8th IWFHR*, pages 105–110, 2002.

[7] L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C. Y. Suen. Feature selection using multi-objective genetic algorithms for handwritten digit recognition. In 16th *ICPR*, pages 568–571, 2002.

[8] W. Siedlecki and J. Sklansky. A note on genetic algorithms for large scale on feature selection. *Pattern Recognition Letters*, 10:335–347, 1989.

[9] N. Srinivas and K. Deb. Multiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation*, 2(3):221–248, 1995.

[10] J. Vesanto and E. Alhoniemi. Clustering of the self-organization map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.