PERGAMON

Pattern Recognition 34 (2001) 523–525

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Comparison of genetic algorithm based prototype selection schemes

T. Ravindra Babu, M. Narasimha Murty*

Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India

Received 26 April 2000

1. Introduction

Prototype selection is the process of finding representative patterns from the data. Representative patterns help in reducing the data on which further operations such as data mining can be carried out. The current work discusses computation of prototypes using medoids [1], leaders [2] and distance based thresholds. After finding the initial set of prototypes, the optimal set is found by means of genetic algorithms (GAs). A comparison of stochastic search algorithms is carried out by Susheela Devi and Narasimha Murty [3]. They conclude that performance of genetic algorithms is the best among the search algorithms. Chang and Lipmann [4] suggest the use of genetic algorithms for pattern classification.

In the following sections, we discuss and compare various prototype selection methods under consideration. Comparison of results are based on nearest neighbor classifier (NNC). Subsequently, considering those prototype sets which provided good classification accuracy, GAs are used for optimal prototype selection. Based on the nature of the data characteristics a number of experiments based on GAs are carried out. A summary of results is presented.

2. Description of data

Handwritten digit data [5] is used for the comparison exercises. The training data consists of 667 patterns for each class of digits 0–9, totalling to 6670 patterns. The test data consists of 3333 patterns. While carrying out experiments using GAs, validation data is drawn from the training data itself.

3. Initial prototype selection

The prototypes are selected based on medoids, leaders and Euclidean distance based thresholds.

3.1. Medoids

The k-medoid method or partition around medoids (PAM) method [1] is based on search for k-representative objects in the input data set. The medoids are selected based on the average minimum dissimilarity. CLARA [1] is an efficient method developed based on PAM. All the exercises of computing medoids are carried out based on CLARA. Since the smallest number of medoids which provides high classification accuracy (CA) cannot be pre-determined, the number of medoids is varied between 20 and 400 per class. Further, the number of medoids is taken as same for each class. With higher number of medoids selected, it is likely that the prototype sets contain redundant medoids.

3.2. Leaders

The leader algorithm [2] is based on a pre-defined dissimilarity threshold. Initially, a random pattern among the input patterns is selected as leader. Subsequently, distance of every other pattern is compared with that of selected leaders. If the distance of new pattern is less than the threshold, the corresponding pattern falls in the cluster with the initial leader. Otherwise, the pattern is identified as a new leader. The computation of leaders is continued till all the patterns are considered. It should be noted that the required distance threshold cannot be predetermined. The number of leaders is

* Corresponding author. Tel.: + 91-80-309-2779; fax: + 91-80-360-2911.

E-mail addresses: trbabu@csa.iisc.ernet.in (T.R. Babu), mnm@csa.iisc.ernet.in (M.N. Murty).

PII: S0031-3203(00)00094-7

inversely proportional to the selected threshold. Thus, a smaller threshold provides a good classification accuracy, but it also results in a redundancy in prototypes.

3.3. Distance-based threshold method

The number of prototypes selected by both the above methods is based on distance. The current method considers initial prototype randomly from the input patterns. The next prototype is computed as the farthest one from the initial prototype. Subsequently, third prototype is selected as the one that is farthest from both of the previously selected prototypes. The procedure continues till no further prototypes can be found or the required number of prototypes is found. Table 1 provides a summary of CA of prototype selection methods.

Table 1
CA for selected prototypes using various methods

Type of prototypes	No. of prototypes	CA(%)
Medoids	2000	90.8
Medoids	3000	91.5
Medoids	4000	91.8
Leaders	5345	92.9
Distance-based	3000	91.5

4. Optimal prototype sections using GAs

Each of the handwritten digit classes is subjected to a preliminary statistical analysis. This results in obtaining a range of distances among all the patterns of each class. For example, the overall range of distances in the current data is between 1 and 12 and this range may vary for each class. During prototype selection the patterns should be found such that they capture all representative patterns with least redundancy. But in each of the previously discussed methods, in the absence of knowledge of optimal number, a large number of prototypes is selected such that they maximize the classification accuracy. In the current section, from such large and possibly redundant set of prototypes, steady-state genetic algorithm (SSGA) is used to obtain the optimal set. The experiments can be classified into three types.

- Search for optimal dissimilarity for medoid reduction.
- Search for optimal dissimilarities for computing optimal leaders.
- Selection of optimal medoid subset by treating each chromosome as a subset of medoids. The length of a chromosome is equal to the size of the initial set of medoids.

These are based on varying the control parameters of the GA: cross-over and mutation probabilities, number of generations, population size and random seed.

4.1. Prototype selection by computing optimal dissimilarity limits

In this case, the optimal lower and upper limits on distances for all the classes are obtained. One set of medoids out of the medoid sets which provided a high CA is considered for reduction. Here, medoids are eliminated if their distance is below the lower limit or above the upper limit from the selected medoids. The results are tabulated in Table 2. The best result provided a classification accuracy of 91.8%.

4.2. Prototype selection by computing optimal thresholds for "leaders"

The leaders have been computed based on selected threshold. Initially, experiments are carried out by searching for optimal distance threshold for each class between lower threshold of 0.0 and upper threshold of 7.0. The best CA obtained is 92.65%. Some of the best results are provided in Table 2. The previous best result obtained on the same data was reported by Prakash and Narasimha Murty [5], using a subspace pattern recognition method as 92%.

Table 2
Summary of CA of optimal prototypes

Type of prototype	No. of optimal prototypes	CA with validation data (%)	CA with test data (%)	Remarks
Medoids	1592	99.27	91.18	Optimal thresholds (2.54, 13.49)
Leaders	5404	100.00	92.65	Threshold range (0.0, 7.0)
Leaders	3432	99.85	92.05	Threshold range (2.0, 5.0)
Vector of medoids	1534	97.84	90.25	Initial set (3000)

4.3. Prototype selection by computing optimal medoid set

The SSGA experiments are designed to search for optimal set of medoids directly. Here, each position in the chromosome takes a value of 0 or 1, indicating absence or presence of the corresponding medoid. This results in obtaining about 1500 out of 3000 medoids with best CA of 90.25%. Table 2 provides summary of best CAs obtained using above methods.

5. Summary and conclusions

The current study enlists three prototype selection methods and demonstrates experimentally their merits and demerits based on CA and the number of selected prototypes. Considering the set of prototypes obtained in each of the above methods, the prototype reduction is carried out using GAs. The best CA, of 92.65% is found to be better than previously reported result [5] on the data in the literature.

References

[1] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data — An Introduction to Cluster Analysis, Wiley, NY, 1989.

[2] H. Spath, Cluster Analysis — Algorithms for Data Reduction and Classification of Objects Ellis Horwood Limited, West Sussex, UK, 1980.

[3] V. Susheela Devi, M. Narasimha Murty, in: Sankar K. Pal, A. Ghosh, M.K. Kundu (Eds.), Handwritten Digit Recognition Using Soft Computing, Physica-Verlag, Berlin, 2000.

[4] E.I. Chang, R.P. Lippmann, Using genetic algorithms to improve pattern classification performance, Advances in Neural Information Processing Systems 3, Morgan Kaufman, Los Attos, CA, 1990, pp. 797–803.

[5] M. Prakash, M. Narasimha Murty, Growing subspace pattern recognition methods and their neural network models, IEEE Trans. Neural Network 8 (1) (1997) 161–168.