Analysis of Evolutionary Prototype Selection by means of a Data Complexity Measure based on Class Separability

José-Ramón Cano
Dept. de Informática
Universidad de Jaén
EPS de Linares
23700 Linares (Jaén)
jrcano@ujaen.es

Salvador García, Francisco Herrera
Dept. de Cienc. de la Computac.n e Intel. Artificial
Universidad de Granada
ETS Ingeniería Informática
18071 Granada
(salvagl,herrera)@decsai.ugr.es

Ester Bernadó-Mansilla
Dept. de Ingeniería Informática
Universidad de Ramon Llull
Enginyeria i Arquitectura La Salle
08022 Barcelona
esterb@salleurl.edu

Abstract

In the literature there are several proposals of prototype selection algorithms. These algorithms follow different strategies or heuristics, being the evolutionary one of them. In this paper we analyze the behaviour of the evolutionary prototype selection strategy, considering a complexity data set measure based on class separability. The study has as objective the prediction of when the evolutionary prototype selection is effective for a particular problem, using the class separability measure.

1 Introduction

The nearest neighbour rule is a classical supervised classification method used in pattern recognition [15]. The nearest neighbour classifier try to predict the class of a new prototype computing the euclidean distance between it and every prototype previously stored, and considering as response the class of the nearest one (in the case on 1 nearest neighbour).

The prototype selection (PS) is a classic supervised learning problem where the objective consist on, using an input data set, find those prototypes which improve the accuracy of the nearest neighbour classifier [17]. More formally, let's assume that there is a training set T which consists of pairs (x_i, y_i), i = 1, ..., n, where x_i defines input vector of attributes and y_i defines the corresponding class label. T contains n samples, which have m input attributes each one and they should belong to one of the c classes. Let S T be the subset of selected samples resulted for the execution of an prototype selection algorithm algorithm.

There are many proposals of prototype selection algorithms [20, 7]. Those methods follow different strategies for the prototype selection problem, and offer different behaviour depending of the input data set.

One of those strategies or heuristics are the evolutionary algorithms. Evolutionary Algorithms (EAs) ([1, 6]) have been used to solve the PS problem in [3, 10, 16] with promising results. The EAs algorithms are very effective in their use but they are not very efficient in execution time, so it would be interesting to characterize the input data sets to improve their use.

In the literature there are studies of the properties of the data sets, where the authors present complexity measures to characterize them [9, 18]. Mollineda et al. in [12] presents a previous work where they analyze the complexity measures of overlap and nonparametric separability considering the Wilson's Edited Nearest Neighbor [19] and the Hart's Condensed Nearest Neighbour [8] as prototype selection algorithms.

In this study we are interested in the prediction of when the evolutionary prototype selection is effective for a particular problem, using the class separability measure suggested in

[12].

In order to do that, the paper is set out as follows. Section 2 is dedicated to describe the evolutionary prototype selection strategy and the algorithm which represents it in this study. In Section 3, we present the complexity measure considered. Section 4 explains the experimentation study and deals with the results and their analysis. Finally, in Section 5, we point out our conclusions.

2 Evolutionary Prototype Selection

Evolutionary Algorithms may be applied to the PS problem ([3]) because it can be considered as a search problem. The application of EAs to PS is accomplished by tackling two important issues: the specification of the representation of the solutions and the definition of the fitness function.

- Representation: Let's assume a data set denoted T with n instances. The search space associated with the instance selection is constituted by all the subsets of T. A chromosome consists on the sequence of n genes (one for each instance in T) with two possible states: 1 and 0, meaning that the instance is or not included in the subset selected respectively.

- Fitness function: Let S be a subset of instances of T to evaluate and be coded by a chromosome. We define the fitness function that combines two values: the classification performance ($clasper$) associated with S and the percentage of reduction ($percred$) of instances of S with regards to T:

$$Fitness(S) = \alpha \cdot clasper + (1-\alpha) \cdot percred. \quad (1)$$

The 1-Nearest Neighbour (1-NN) classifier is used for measuring the classification rate, $clasper$, associated with S. It denotes the percentage of correctly classified objects from T using only S to find the nearest neighbour. For each object y in T, the nearest neighbour is searched for amongst those in the set $S \backslash \{y\}$. Whereas, $percred$ is defined as:

$$percred = 100 \cdot (|TR| - |S|)/|TR|. \quad (2)$$

The objective of the EAs is to maximize the fitness function defined.

As evolutive algorithm we have selected the CHC [5] considering its behaviour in [3]. During each generation the evolutionary instance selection CHC (EIS-CHC) develops the following steps:

1. It uses a parent population to generate an intermediate population of individuals, which are randomly paired and used to generate potential offspring.

2. Then, a survival competition is held where the best chromosomes from the parent and offspring populations are selected to form the next generation.

EIS-CHC also implements a form of heterogeneous recombination using HUX, a special recombination operator and a method of incest prevention. No mutation is applied during the recombination phase. Instead, when the population converges or the search stops making progress (i.e., the difference threshold has dropped to zero and no new offspring are being generated which are better than any members of the parent population) the population is reinitialized to introduce new diversity to the search. The chromosome representing the best solution found over the course of the search is used as a template to reseed the population. Reseeding of the population is accomplished by randomly changing 35% of the bits in the template chromosome to form each of the other new chromosomes in the population.

The fitness function (see expression 1)combines two values: the classification rate (using 1-NN) associated with S and the percentage of reduction of instances of S with regards to T.

3 Data Set Characterization Measure

The prediction capabilities of classifiers are strongly dependent on data complexity. That's the reason why various recent papers have introduced the use of measures to characterize the data and to relate those characteristics to classifier performance.

In [9], authors define some complexity measures for two classes. Singh in [14] offers a review of data complexity measures and proposes two new ones. Dong et al. in [4] propose a feature selection algorithm based in a complexity measure defined by Ho. In [11], Li et al. analyze some omnivariate decision trees using the measure of complexity based in data density proposed by Ho. Authors in [2] define specific measures for regularized linear classifiers, using the Ho measures as reference. Mollineda et al. in [12] extend some Ho's measure definitions for problems with two or more classes. They analyze these generalized measures in two classic PS algorithms and remark that the Fisher's discriminant ratio is the most effective for PS.

According to their conclusions, we have consider that measure to study the behaviour of evolutionary prototype selection. In this section we present the Fisher's discriminant ratio, which is based in class separability.

The plain version of Fisher's discriminant ratio offered by Ho et al. ([9]) computes how separated are two classes according to a specific feature. It compares the difference between class means with the sum of class variances. Fisher's discriminant ratio is defined as follows:

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{3}$$

where μ_1, μ_2, σ_1^2, σ_2^2 are the means and the variances of the two classes respectively.

A possible generalization for C classes is proposed by Mollineda et al. ([12]), and considers all feature dimensions. Its expression is the following:

$$F1 = \frac{\sum_{i=1}^{C} n_i \cdot \delta(m, m_i)}{\Sigma_{i=1}^{C} \Sigma_{j=1}^{n_i} \delta(x_j^i, m_i)} \tag{4}$$

where n_i denotes the number of samples in class i, δ is the metric, m is the overall mean, m_i is the mean of class i, and x_i^j represent the sample j belonging to class i. Smaller values of this measure indicates that classes present strong overlap.

4 Experimental Study

To analyze the EIS-CHC behaviour we include in the study two classical prototype selection algorithms, which will be described in Section 4.1. In the Subsection 4.2 we present the algorithm's parameters and data sets considered. Subsection 4.3 contains the results and Subsection 4.4 their analysis.

4.1 Prototype Selection Algorithms

The classical PS algorithms used in this study for the comparison are:

- Edited Nearest Neighbor (ENN) [19]. Wilson developed this algorithm which starts with $S = T$ and then each instance in S is removed if it does not agree with the majority of its k nearest neighbors. ENN filter is considered the standard noise filter and is usually employed as a noise filter previous stage of many algorithms.
- Condensed Nearest Neighbour (CNN) [8]. It begins by randomly selecting one instance belonging to each output class from T and putting them in S. Then, each instance in T is classified using only the instances in S. If an instance is misclassified, it is added to S, thus ensuring that it will be classified correctly. This process is repeated until there are no instances in T that are misclassified.

4.2 Data Sets and Parameters

The experimental study is defined in two aspects: Data sets and algorithm's parameters. They are as follows:

- Data Sets: The data sets used have been collected from the UCI repository [13] and their characteristics appear in Table 1. The features in the data sets conserve their original values, without normalization.

- Parameters: The parameters are chosen considering the authors suggestions in the literature. For each one of the algorithms are:

  - *CNN*: It hasn't parameters to be fixed.
  - *ENN*: Number of neighbours=3.
  - *EIS-CHC*: evaluations=10000, population=50 and α=0.5.

The algorithms have been executed one time for each partition in the ten fold cross validation. The measure F1 is calculated over the whole data sets in ten fold cross validation too.

4.3 Results

The results are presented using the following table structure:

- In the first column we offer the name of the data sets, ordered considering the measure F1.

- The second column contains the measure F1 associated to the whole data set in increasing order.

- The third column contains the 1-NN test accuracy rate considering the whole data set, without PS. Between brackets appears the percentage of reduction over the original data set (obviously it conserves the 100% of the initial data set).

- Columns fourth, fifth and sixth present the 1-NN test accuracy using ENN, CNN and EIS-CHC respectively. In the same way that in the previous column, between brackets the percentage of reduction is offered.

In Table 2 we present the results, where the values in bold indicate the test accuracy rates equal or bigger than the offered by the 1-nearest neighbor using the whole data set (Without PS).

4.4 Analysis

Analyzing the Table 2 we can point out the following:

- The EIS-CHC algorithm outperforms the 1-NN (called Without PS in the table of results) when F1 is small. It presents the best accuracy rates in all data sets with the strongest overlaps.

- These behaviour is similar for those data sets with high F1 (weak overlap). EIS-CHC shows a interesting behaviour when F1 is small or high, for strong and weak overlaps.

- When F1 has a medium value, the EIS-CHC presents similar accuracy rates than the 1-NN, being its reduction capabilities the maximal ones.

- Paying attention to the relation between F1 and the behaviour of the EIS-CHC, we can point that the use of this measure can help us to decide when the use of EIS-CHC improves the accuracy rates of 1-NN classifier in a concrete data set, previously to its execution.

With these results on mind, we could analyze the F1 measure in a new data set and if small or high, we can use the EIS-CHC as PS algorithm to improve the accuracy rate and get the maximal reduction in the training set.

If F1 is medium, the accuracy rate offered by the EIS-CHC is similar than the Without PS and the classical PS algorithms, but it maintains its maximal reduction capabilities.

5 Concluding Remarks

This paper addresses the analysis of the evolutionary prototype selection considering a complexity data set measure based on class separability, with the objective of the prediction of

Table 1: Data Sets.

	Instances	Features	Classes
Bupa	345	6	2
Ecoli	336	7	2
Iris	150	4	3
Glass	214	9	6
Led24digit	200	24	10
Led7digit	500	7	10
Lymphography	148	18	4
Monks	432	6	2
Penbased	10992	16	10
Pima	768	8	2
Wine	178	13	3
Wisconsin	683	9	2
Satimage	6435	36	7
Thyroid	7200	21	3
Zoo	100	16	7

Table 2: Test accuracy rate and percentage of training instances sorted by the F1 measure.

	F1	Without PS	ENN	CNN	EIS-CHC
Thyroid	0.03	0.93(1)	**0.94**(0.93)	0.88(0.14)	**0.94**(0.01)
Lymphography	0.17	0.35(1)	**0.47**(0.38)	0.27(0.71)	**0.4**(0.04)
Bupa	0.17	0.68(1)	0.66(0.64)	0.63(0.42)	**0.69**(0.03)
Pima	0.22	0.70(1)	**0.71**(0.71)	0.6(0.37)	**0.75**(0.01)
Ecoli	0.24	0.82(1)	**0.88**(0.90)	**0.85**(0.12)	**0.91**(0.01)
Monks	0.36	0.95(1)	0.91(0.91)	0.84(0.23)	**0.99**(0.01)
Led24digit	0.47	0.15(1)	**0.15**(0.42)	**0.25**(0.67)	**0.3**(0.07)
Glass	0.74	**0.57**(1)	0.52(0.68)	0.52(0.41)	0.48(0.05)
Penbased	1.16	**0.99**(1)	**0.99**(0.99)	0.98(0.04)	0.96(0.01)
Led7digit	1.34	0.58(1)	**0.66**(0.74)	0.5(0.45)	**0.64**(0.03)
Wisconsin	1.35	0.94(1)	**0.96**(0.97)	**0.97**(0.07)	**0.94**(0.01)
Zoo	1.38	**0.99**(1)	**0.99**(0.92)	0.9(0.2)	**0.99**(0.09)
Satimage	1.47	**0.90**(1)	**0.90**(0.90)	0.88(0.15)	0.86(0.01)
Wine	1.82	**0.72**(1)	**0.72**(0.96)	**0.72**(0.15)	**0.72**(0.03)
Iris	2.66	0.93(1)	0.93(0.96)	0.93(0.09)	**0.99**(0.01)

when the evolutionary prototype selection is effective for a particular problem.

An experimental study has been carried out using data sets from different domains and comparing the results with classical PS algorithms, having the F1 measure as reference. The main conclusions reached are the following:

- The EIS-CHC presents the best accuracy rate when the input data set has strong or weak overlapping. In addition, the EIS-CHC is the one with the smallest subsets selected among the PS algorithms.
- When the overlapping of the data set is medium, the EIS-CHC has associated similar accuracy rates than the 1-NN and the ENN (the classic PS algorithm with the best test accuracy rate), but its reduction rates are kept (higher than 91% of the original data set).

As we have indicated in the analysis section, the use of this measure can help us to evaluate a data set previously to the evolutionary PS process and decide if it is adequate or not to improve the classification capabilities of the 1-nearest neighbour.

Acknowledgments.

This work was supported by Projects TIN2005-08386-C05-01 and TIN2005-08386-C05-03.

References

[1] Back, T., Fogel, D., Michalewicz, Z., *Handbook of evolutionary computation*, Oxford University Press, 1997.

[2] Baumgartner, R., Somorjai, R.L., *Data complexity assessment in undersampled classification of high-dimensional biomedical data*, Pattern Recognition Letters 27:12 (2006) 1383-1389.

[3] Cano, J.-R., Herrera, F., Lozano, M., *Using Evolutionary Computation as Instance Selection for Data Reduction in KDD: An Experimental Study*, IEEE Transactions on Evolutionary Computation 7:6 (2003) 561-575.

[4] Dong, M., Kothari, R., *Feature subset selection using a new definition of classifiability*, Pattern Recognition Letters 24:9-10 (2003) 1215-1225.

[5] Eshelman, L.J., T*he CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination*, in: Foundations of Genetic Algorithms 1, Rawlins, G.J.E. (Eds.), Morgan Kauffman, 1991, 265-283.

[6] Goldberg, D.E., *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley, 1989.

[7] Grochowski, M., Jankowski, N., *Comparison of instance selection algorithms II. Results and Comments*, in: Proceedings of the 7th International Conference on Artificial Intelligence and Soft Computing, LNCS 3070, 2004, 58-585.

[8] Hart, P. E., *The Condensed Nearest Neighbour Rule*, IEEE Trans. on Information Theory 14 (1968) 515-516.

[9] Ho, T. K., Basu, M., *Complexity Measures of Supervised Classification Problems*, IEEE Transactions on Pattern Analysis and Machine Intelligence 24:3 (2002) 289-300.

[10] Kuncheva, L., *Editing for the k-nearest neighbors rule by a genetic algorithm*, Pattern Recognition Letters 16 (1995) 809-814.

[11] Li, Y.-H., Dong, M., Kothari, R., *Classifiability-based omnivariate decision trees*, IEEE Transactions On Neural Networks 16:6 (2005) 1547-1560.

[12] Mollineda, R. A., Sánchez, J. S., Sotoca, J. M.: *Data Characterization for Effective Prototype Selection*, in: Proceedings of the IbPRIA 2005, LNCS 3523, 2005, 27-34.

[13] Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases. [http://www.ics.uci.edu/ mlearn/ ML-Repository.html]. Irvine, CA: University of California, Department of Information and Computer Science (1998).

[14] Singh, S., *Multiresolution estimates of classification complexity*, IEEE Transactions On Pattern Analysis And Machine Intelligence 25:12 (2003) 1534-1539.

[15] Shakhnarovich, G., Darrell, T., Indyk, P. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*, MIT Press, 2006.

[16] Shinn-Ying, H, Chia-Cheng, L., Soundy, L., , *Design of an optimal nearest neighbour classifier using an intelligent genetic algorithm*, Pattern Recognition Letter 23:13 (2002) 1495-1503.

[17] Skalak, D. B., *Prototype and feature selection by sampling and random mutation hill climbing algorithms*, in: Proceedings of teh 11th International Conference on Machine Learning, 1994, 293-301.

[18] Sotoca, J. M., Mollineda, R. A., Sánchez, J. S., A meta-learning framework for pattern classification by means of data complexity measures, Revista Iberoamericana de Inteligencia Artificial 29 (2006) 31-38.

[19] Wilson, D. L., *Asymptotic Properties of Nearest Neighbor Rules Using Edited Data*, IEEE Transactions on Systems, Man and Cybernetics 2:3 (1972) 408-421.

[20] Wilson, D. R., Martinez, T. R., *Reduction techniques for instance-based learning algorithms*, Machine Learning 38 (2000) 257-268.