Improving Nearest Neighbor Classification using Ensembles of Evolutionary Generated Prototype Subsets

Sarah Vluymans$^{(1,2,3)}$, **Nele Verbiest**$^{(1)}$, **Chris Cornelis**$^{(1,3)}$,
Nicolás García-Pedrajas$^{(4)}$, **Yvan Saeys**$^{(2,5)}$ Contact: sarah.vluymans@ugent.be

$^{(1)}$Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
$^{(2)}$Data Mining and Modeling for Biomedicine, VIB Inflammation Research Center, Ghent, Belgium
$^{(3)}$Department of Computer Science and Artificial Intelligence, University of Granada, Spain
$^{(4)}$Department of Computer Science and Numerical Analysis, University of Córdoba, Spain
$^{(5)}$Department of Internal Medicine, Ghent University, Belgium

Keywords: classification, evolutionary algorithms, prototype selection, ensembles

Abstract

Prototype selection reduces the dataset before the application of a classifier in order to achieve an improved accuracy and/or a considerable reduction in the number of instances. Among the proposed algorithms, evolutionary methods are the state-of-the-art. In (Verbiest et al., 2016), we developed a framework to further enhance the performance of these methods with minimal additional effort. We recall our findings here.

1. Introduction

Prototype selection (PS) is a type of preprocessing procedure that is executed on the dataset before training a classifier. It actively reduces the dataset by only selecting relevant and non-noisy elements from it. The goal of PS is twofold. On the one hand, by removing noisy samples, the accuracy of the posterior classifier can be improved. On the other, a high reduction of redundant elements can considerably decrease both storage requirements and runtime in the classification step. A classifier that is commonly combined with PS is the k nearest-neighbor method (kNN).

Evolutionary methods are the state-of-the-art among PS algorithms, as demonstrated in the study of (García et al., 2012), in terms of both the classification performance and reduction in the number of training instances. An evolutionary PS method evolves a population of candidate prototype subsets over a number of generations to finally obtain an optimal solution. All intermediate solutions are discarded. In (Verbiest et al., 2016), we posited that although these candidate subsets are not globally optimal, they may still perform well in a given subspace. Instances in particular regions of the feature space may be classified more accurately by candidates other than the final solution.

Based on this premise, we set up a diverse ensemble of well-performing prototype subsets that have been constructed during the execution of the evolutionary algorithm. When classifying a new instance, custom weights for each ensemble member are calculated. In this way, we determine the more appropriate prototype subsets and assign them a higher weight in the classification. Our framework is called *Ensembles of Evolutionary Generated Prototype Subsets (EEGPS)* and is described in Section 2. A summary of the experimental results can be found in Section 3.

2. The EEGPS framework

A visual overview of EEGPS is provided in Figure 1. For more detail, we also refer the reader to the original proposal. Let us consider a general evolutionary PS method M. During its execution, many candidate subsets are constructed. Rather than simply discarding these intermediate solutions, a percentage *pbest* is selected from among them. They represent the best-performing prototype subsets, based on the fitness measure used by M. As these sets will be used in an ensemble, it is important to guarantee a level of diversity. To this end, the selected group is further reduced by selecting the *pdiv*% most diverse from them.

When the ensemble members have been selected, the classification phase can commence. To classify a new

Appearing in *Proceedings of Benelearn 2016*.

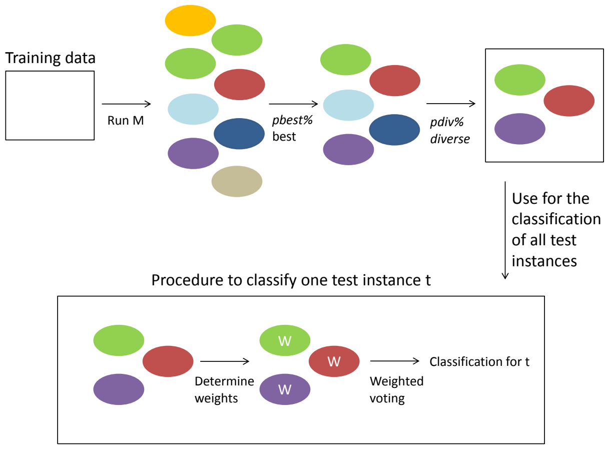

Figure 1. Overview of the EEGPS framework. An ellipse represents a prototype subset.

instance t, custom weights are assigned to the prototype subsets in the ensemble. Subsets that are estimated to lead to a higher accuracy in the region of t are assigned higher weights. Finally, 1NN is applied to classify t with each ensemble member. The predictions are aggregated with a weighted vote, based on the weights computed in the previous step.

3. Experimental results

In (Verbiest et al., 2016), we selected four different evolutionary PS methods to incorporate in our framework: GGA, SSGA, CHC and SSMA. In a first phase, we conducted an extensive analysis of the influence of the framework parameters on the performance and provided guidelines for proper EEGPS settings for each of the four PS methods.

Secondly, we showed the performance gain of EEGPS over the traditional PS setting, as presented in Figure 2. All selected PS methods exhibit a better performance within EEGPS compared to the original PS. Moreover, we were able to show that, within EEGPS, a smaller number of generations of the evolutionary are often sufficient to obtain these improvements.

Finally, it was also shown that the integration of the PS methods in EEGPS leads to a negligible increase in computational cost, since the bulk of the work is done by the genetic algorithm itself.

4. Conclusion

In (Verbiest et al., 2016), we proposed a framework that enhances the performance of evolutionary PS algorithms at a minimal increase in computational cost. A diverse ensemble of well-performing prototype sets generated by the PS method is set up. The experimental work clearly showed the accuracy improvement of the EEGPS framework over traditional PS. These improvements are already obtained after fewer generations of the evolutionary methods.

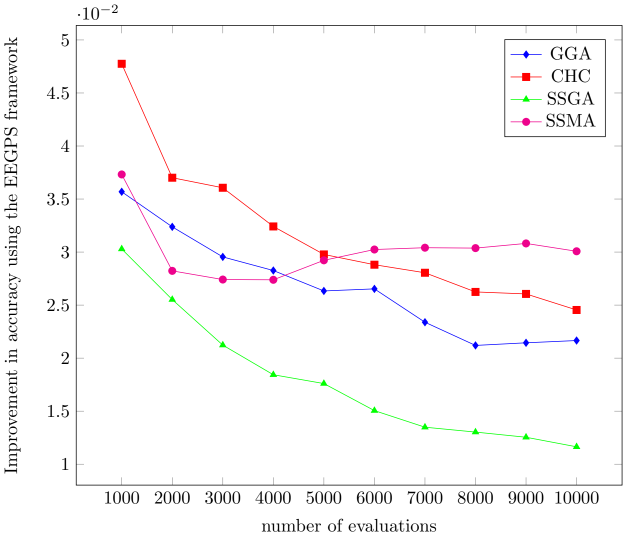

Figure 2. Absolute accuracy improvement of EEGPS over traditional PS.

Acknowledgments

S. Vluymans is supported by the Special Research Fund (BOF) of Ghent University. C. Cornelis was partially supported by the Spanish Ministry of Science and Technology under the project TIN2014-57251-P and the Andalusian Research Plans P11-TIC-7765, P10-TIC-6858 and P12-TIC-2958, and by project PYR-2014-8 of the Genil Program of CEI BioTic GRANADA.

References

García, S., Derrac, J., Cano, J., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*, 417–435.

Verbiest, N., Vluymans, S., Cornelis, C., García-Pedrajas, N., & Saeys, Y. (2016). Improving nearest neighbor classification using ensembles of evolutionary generated prototype subsets. *Applied Soft Computing*, *44*, 75–88.