GENIFER: A Nearest Neighbour based Classifier System using GA

Francesc Xavier Llorà i Fàbrega
Enginyeria i Arquitectura La Salle (URL)
Pg. Bonanova 8, 08022-Barcelona,
Catalonia, Spain, Europe
xevil@salleURL.edu

Josep Maria Garrell i Guiu
Enginyeria i Arquitectura La Salle (URL)
Pg. Bonanova 8, 08022-Barcelona,
Catalonia, Spain, Europe
josepmg@salleURL.edu

Summary

The work being summarized here deals with automatic classification problems. Our interest is focused in problems with the following characteristics: the problem's attributes are real valued, the classification classes are previously known and a set of positive and negative examples is given. The aim is to create a system that learn a set of classification rules.

Due to the real valued attributes the traditional GA based systems [1, 2] does not work properly. Our work is oriented to find an alternative way to represent the condition part of the rules in order to improve system's performance.

Our system (GENIFER) is based on the following concepts: *nearest neighbour policies* like the ones used in Case-Based Reasoning, *adaptive behaviour* of GAs, and reduntant chromosome information codification using *diploids*.

Our GA based classifier system instead of looking for rules that describes logical tests over attributes, looks for representative classification areas in the n-dimensional search space (defined by the real valued attributes) described by significant points. GENIFER uses a matching function based on nearest neighbour policies. Distance between one sample (m_i) and one point (x) is shown in equation 1. So, the system should obtain a set of points instead of a set of rules.

$$Dist(m_i, x) = \sqrt[n]{\sum_{j=1}^{Max\_attr} |m_{ij} - x_j|^n} \qquad (1)$$

When a new sample has to be classified by the system, the nearest significant point (r) is retrieved from the learned set of points (R) and the new sample is classified in the area described by the point:

$$Dist(m_i, r) \leq \min_{\forall x \in R} (Dist(m_i, x)) \qquad (2)$$

From this simple idea, the GENIFER system is improved with some adaptive behavior. As show in equation 1 several nearest neighbour functions can be used. GENIFER can automatically learn the best function from a discrete set of choices (GENIFER-MDAA). We also have been working with incremental approaches similar to the ones used in [1, 2]. Finally this incremental approach incorporates redundant information in chromosomes using *diploids* (GENIFER-DIA).

The system is tested using a real world noisy problem: the automatic diagnosis of breast cancer biopsies. The classes for this problem are cancerous or non cancerous. The systems achieves a mean prediction accuracy of 82% (correctly classified) over the complete testbed and all the GENIFER configurations. These results are better that the ones reported using traditional GA based classifier systems (72%).

Acknowledgments

This work was partially supported under grant number 1999FI-00719 by the Generalitat de Catalunya. We also thank Enginyeria i Arquitectura La Salle (Universitat Ramon Llull) for their support to our Research Group in Intelligent Systems.

References

[1] Kenneth A De Jong and William M. Spears. Learning Concept Classification Rules Using Genetic Algorithms. *Proc. of the 12th Int. Joint Conf. on Artificial Intelligence*, pages 651–656, 1991.

[2] William M. Spears and Kenneth A. De Jong. Using Genetic Algorithms For Supervised Concept Learning. *Machine Learning, 13*, pages 161–188, 1993.