Suppression Based Immune Mechanism to Find a Representative Training Set in Data Classification Tasks

Grazziela P. Figueredo
COPPE-UFRJ
Rio de Janeiro, Brazil
gpfigueredo@gmail.com

Nelson F. F. Ebecken
COPPE-UFRJ
Rio de Janeiro, Brazil
nelson@ntt.ufrj.br

Helio J. C. Barbosa
LNCC/MCT
Petrópolis, Brazil
hcbm@lncc.br

ABSTRACT

This article proposes a new classifier –inspired by a biological immune systems' characteristic– which also belongs to the class of k-nearest-neighbors algorithms. Its main feature is a suppression mechanism used to reduce the size of the training set –maintaining the most significative samples– without loosing much capability of generalization.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: [Design Methodology, Classifier Design and Evaluation]; H.2.8 [**Database Management**]: [Database Applications, Data mining]

General Terms

Design, Algorithms, Experimentation

Keywords

Artificial Immune Systems, Classification

1. THE PROPOSED ALGORITHM

The algorithm starts with the idea that the system's model must evolve to create antibodies that recognize the training set and be able to identify new presented antigens. Therefore, instead of the system generating and evolving B cells clones until the antibodies recognize the training set and establish a cellular memory, it is proposed that the training set itself constitutes the repertory of antibodies of the system. To proceed the system's learning, the database is divided into three subsets, training representing the antibodies, testing and validating as antigens. The initial proportion of samples adopted for each group, was empirically set as 60%, 20% and 20%, respectively. Both antigens and antibodies are represented by an array containing the attributes. The antigens, or test data, are classified according to the closest antibody, which is determined by a measure of distance. The reason why antigens have been split into two subsets is to provide the proposed suppression mechanism which is used in the training set, so that very similar antibodies are eliminated and the best ones are kept. Those antibodies able to recognize antigens from the test set remain while the others are eliminated from the population. In this artificial classifier, the signals for antibodies' survival are represented by a counter variable increased by one point for each antigen recognized, independently if the classification is correct or not.

2. EXPERIMENTS AND RESULTS

Experiments were performed using some databases extracted from the UCI machine learning repository. The metrics adopted to evaluate the efficiency of the classifier were extracted from [1].

2.1 Pima Indians Diabetes Database

Table 1: Results for the Pima Indians Diabetes Database. First line: predictor without suppression mechanism using 460 antibodies. Second line: predictor with suppression mechanism using 80 antibodies

acc	val	sens	spec	prec	Fmea	GSP	GSS
0.74	0.67	0.7	0.62	0.77	0.36	0.73	0.66
0.74	0.65	0.77	0.45	0.71	0.37	0.74	0.59

3. CONCLUSIONS

This article proposed a new k-nearest-neighbor deterministic data classifier method using a suppression mechanism inspired on the behavior of biological immune systems. Experiments were made using benchmarks and the results were satisfactory. There was just a little proportion of mistakes between test and validation set. The division between training and test data was set, at these first experiments, to 60% and 40%. As next steps of this work, it is intended to further investigate multi-class problems, databases with unbalanced examples, and to compare the obtained results with other classifiers found in the literature.

4. REFERENCES

[1] R. P. Espínola and N. F. F. Ebecken. On extending f-measure and g-mean metrics to multi-class problems. In *DATA MINING VI - Data Mining, Text Mining and Their Business Applications*, volume 1, pages 25–34, 2005.

GECCO'07 July 07/11, 2007, London, England, United Kingdom.
ACM 978-1-59593-697-4/07/0007.