

The SUPRAIC Algorithm: A Suppression Immune Based Mechanism to Find a Representative Training Set in Data Classification Tasks

Grazziela P. Figueredo¹, Nelson F.F. Ebecken¹, and Helio J.C. Barbosa²

¹ COPPE/UFRJ, Rio de Janeiro, RJ, Brazil

² LNCC/MCT, Av. Getúlio Vargas 333
25651 070 Petrópolis RJ, Brazil

Abstract. This article proposes a new classifier inspired on a biological immune systems' characteristic. This immune based predictor also belongs to the class of k-nearest-neighbors algorithms. Nevertheless, its main features, compared to other artificial immune classifiers, are the assumption that training set is the antibodies' population and a suppression mechanism that tries to reduce the training set into a smaller subset. This subset is supposed to contain the most significant samples, without losing much capability of generalization. It is known that in prediction problems, the choice of a good training set is crucial for the classification process. And this is the focus of this research. Experiments using some benchmarks and the analysis of the results of our ongoing work are presented.

1 Introduction

Classification [1] and pattern recognition are important tasks in all knowledge fields [2]. To classify means to categorize into a finite number of classes elements defined by a group of attributes. There are many types of statistical and artificially intelligent classifiers, as it can be seen in [3,4]. One of the main issues in classification problems involves the choice of good samples to train a classifier. A training set capable to represent well the characteristics of a class has better chances to establish a successful predictor.

This article proposes a new classifier inspired on biological immune systems' characteristics. The model belongs to the class of k-nearest-neighbors algorithms. Its main features, compared to other artificial immune classifiers [5,6,7,8,9,10,11,12,13,14,15,16,13,17,18,19,20,21,22], are the assumption that the training set constitutes the initial antibodies' population of the system, and a suppression mechanism that tries to reduce this training set into a smaller subset. This subset is supposed to contain the most significant samples, without losing much capability of generalization.

In this proposal, the inspiration came from the self-regulation mechanism from the biological immune systems. Clones of B cells that are no longer needed

in the organism suffer apoptosis. Therefore, another feature of the algorithm that distinguishes it from the others mentioned, is its simplicity. There are no mechanisms such as affinity maturation, clonal selection [23,24], its not based on immune network models or kohonen networks [19].

1.1 The Behavior of Biological Immune Systems

According to the IS self regulation mechanism, clones no longer needed by the organism, or those that self attack, do not receive signals to keep alive and suffer apoptosis. Those signals came from the lymph nodes or T helper cells [25]. This self-regulation characteristic allows the organism to save energy and keep only the repertory of lymphocytes really needed for self defense. These concepts are the inspiration of the presented model.

Basically, the algorithm presented here draws from a concept in which the model for the system should evolve to produce antibodies to recognize the training data and be capable to identify new presented antigens. Instead of working with a system that generates and evolves clones of B cells until the antibodies recognize the training group, this paper proposes the training data to be the very same antibodies' repertory of the system.

The suppression concept is employed in the training set to eliminate very similar antibodies. For suppression, the antigens, or test data, are divided into two subgroups. The first will be responsible for testing the model and eliminating redundant antibodies. The second subgroup is used to validate the efficiency of the remaining antibodies after suppression.

This article is organized as follows. Section 2 presents a detailed description of the algorithm proposed together with the suppression mechanism. Experiments and results obtained from standard databases found in the literature are presented. Finally, the last section presents the final remarks on the model and propose some suggestions for further research.

2 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. It means that the immunological memory represented by Artificial Recognition Balls (ARBs), which are mathematical metaphors for antibodies or B-cells [26,27,24], (see Figure 1), will be ready for generalization from the moment training data is placed in the system. In the presented figure, S is the shape space where \bullet are the antibodies, \times are the antigens and ε denotes the radius of each antibody [27,24].

The set of attributes is normalized to belong to the same scale of values. In this case, the data was transformed to fit in the interval $[0,1]$. At the present moment of this research, only real or integer data are considered.

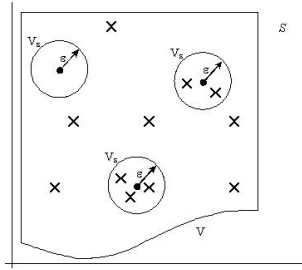


Fig. 1. The shape space model

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, respectively, was 60%, 20% and 20% and then, another experiment with the proportions 70%, 20% and 10%. These proportions were adopted empirically. Both antigens and antibodies are represented by an array containing the attributes. The antigens, or test data, are classified according to the closest antibody. It means that in this case, instead of a shape space with variant ε , the model uses the Voronoi diagram [28], that can be seen in Figure 2. In the figure, dots are the centroids of the ARBs and represent the antibodies that cover a certain group of nearest antigens.

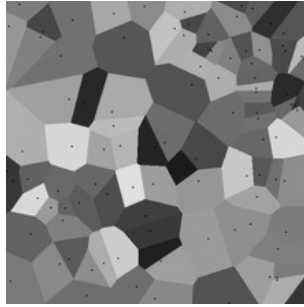


Fig. 2. Voronoi diagram: a representation of the classification process

The closest antibody is determined by a measure of distance. In the experiments, the adopted ones were the Euclidian and Manhattan distances, described respectively by the equations 1 and 2.

$$distance = \sqrt{\sum_{i=1}^{TotalAttributes} (Antibody[i] - Antigen[i])^2} \quad (1)$$

$$distance = \sum_{i=1}^{TotalAttributes} |Antibody[i] - Antigen[i]| \quad (2)$$

The reason why antigens have been split into two subsets is to provide the suppression mechanism. The suppression concept is used among the training set, so that very similar antibodies are eliminated and the best ones are kept, following the theories of self-regulation and affinity maturation of biological immune systems, described in section 1.1. In other words, those antibodies able to recognize antigens from the test set remain while the others are eliminated from the population.

In this artificial classifier, these signals for survival are represented by a counter variable for each antigen recognized by an antibody, independently if the classification is correct or not. It is known that some databases have limits on the percentage of correctly classified samples. This is the reason why antibodies that classify incorrectly are still maintained on the suppressed population. These antibodies could be viewed as cross-reactive ones.

The schematic algorithm can be seen in Algorithm 1, called SUPRAIC.¹

Algorithm 1. Algorithm SUPRAIC

- ◊ Read the database file;
 - ◊ Normalize data between the interval [0,1];
 - ◊ Determine antibodies and antigens by dividing the data (in a uniform distribution) into, for example, 60% antibodies and 40% antigens;
 - ◊ For each antibody, set its counter variable = 0;
 - ◊ Divide antigens into test and validation subgroups. Generally, it is used the same proportion for each one, 20%;
 - ◊ Determine the initial antibodies as the training set.
 - ◊ Train the model via test subset by finding the nearest antibody for each antigen. The nearest antibody is measured by the Euclidian distance among attributes;
 - ◊ For each antigen recognized by the antibody, increase its counter variable by one;
 - ◊ Suppress the remaining antibodies not capable to recognize (counter = 0) antigens. Eliminate them from the antibodies population.
 - ◊ Validate the final predictor, already suppressed by the previous step, by using the remaining antibodies to recognize the validation set;
 - ◊ Calculate the accuracy of the final predictor.
-

3 Experiments and Results

This section presents experiments using some benchmark examples of databases extracted from the UCI machine learning repository [29]. The metrics adopted to evaluate the efficiency of the classifier were extracted from [2].

For two classes problems, the metrics are based on confusion matrix, a tool which informs the sorts of hits and errors made by a classifier [2]. The classes are named positive and negative and the confusion matrix has four values computed in terms of real and predicted classes, namely:

- TP (true positives): the amount of positive elements predicted as positive;
- TN (true negatives): the amount of negative elements predicted as negative;

¹ Suppressor Artificial Immune Classifier.

- FP (false positives): the amount of negative elements predicted as positive;
- FN (false negatives): the amount of positive elements predicted as negative;

With the values above determined, the most common metrics that will be used to determine the efficiency of the proposed predictor are:

- Accuracy (*acc*) and Validation (*val*): they are the ratio of correct decisions made by a classifier. The difference between both is that the first one is used to train the predictor and the second one is to validate it:

$$acc(val) = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

- Sensitivity (*sens*): it measures how much a classifier can recognize positive examples:

$$sens = \frac{TP}{TP + FN} \quad (4)$$

- Specificity (*spec*): it measures how much a classifier can recognize negative examples:

$$spec = \frac{TN}{TN + FP} \quad (5)$$

- Precision (*prec*): it is the ratio of predicted positive examples which really are positives:

$$prec = \frac{TP}{TP + FP} \quad (6)$$

- F-measure (FM_{ea}): it is the harmonic mean of sensitivity and precision. In this study, the parameter β was set to zero:

$$prec = \frac{(\beta^2 + 1) \times sens \times prec}{sens + \beta \times prec} \quad (7)$$

- G-mean (GSP): it is the geometric mean of sensitivity and precision:

$$GSP = \sqrt{sens \times prec} \quad (8)$$

- G-mean2 (GSS): it is the geometric mean of sensitivity and specificity:

$$GSS = \sqrt{sens \times spec} \quad (9)$$

In all the experiments that follow, β was set to zero, which means that sensitivity and precision have the same importance.

Table 1. Results from the predictor without suppression mechanism for Breast Cancer Database with 419 antibodies

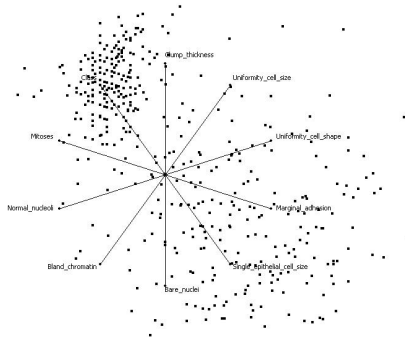
percentage split	distance	acc	val	sens	spec	prec	Fmea	GSP	GSS
60-20-20	Euclidian	0.96	0.97	0.94	0.99	0.98	0.48	0.96	0.96
60-20-20	Manhattan	0.96	0.98	0.94	1	1	0.47	0.95	0.95
70-20-10	Euclidian	0.97	0.97	0.93	1	1	0.48	0.96	0.96
70-20-10	Manhattan	0.98	0.94	0.85	1	1	0.46	0.92	0.92

Table 2. Results from the predictor with suppression mechanism for Breast Cancer Database with 56 antibodies

percentage split	distance	acc	val	sens	spec	prec	Fmea	GSP	GSS
60-20-20	Euclidian	0.96	0.94	0.85	0.99	0.98	0.45	0.91	0.91
60-20-20	Manhattan	0.98	0.96	0.88	1	1	0.47	0.94	0.94
70-20-10	Euclidian	0.97	0.96	0.89	1	1	0.47	0.94	0.94
70-20-10	Manhattan	0.98	0.96	0.88	1	1	0.47	0.94	0.94

Table 3. Results from other classifiers to cancer database

classifier	acc
J48 (Decision Tree)	0.95
Multi-Layer Perceptron	0.97
Naive Bayes	0.98

**Fig. 3.** The total amount of antibodies (training set) for the Cancer database

3.1 Breast Cancer Database

The purpose of this work was to reduce training sets without losing significantly the accuracy of the predictor. The first experiment of this new proposal was made on the breast cancer database. The breast cancer database is characterized for having 699 registers, 9 attributes and 2 classes.

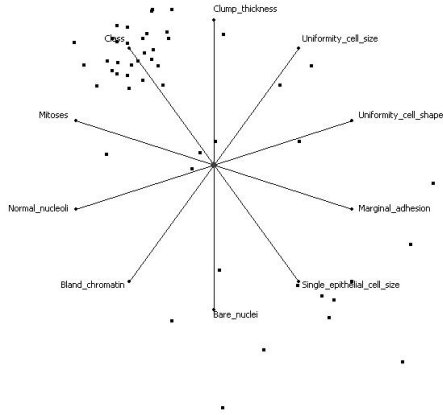


Fig. 4. Antibodies from Cancer database after suppression and 60-40% split

Table 1 shows the results with the previously mentioned metrics, but using the classifier without the suppression mechanism. In Table 2, the suppression mechanism was applied.

As it can be seen, comparing the two results in Tables 1 and 2, there was not too much difference between the metrics. Nevertheless, the amount of antibodies was reduced from 419 to 56. That means a reduction of almost 87% of training data. Figure 3 shows training data before suppression and Figure 4 shows the final suppressed classifier using star coordinates.

3.2 Pima Indians Diabetes Database

Database diabetes is characterized by 768 registers, 8 attributes and 2 classes. As in the first database studied, it also can be seen that the results from Tables 4 and 5 are not very different, except for the value specificity. Even though, the reduction of the antibodies by the suppression was from 460 to 80. Figures 5 and 6 shows, respectively, the training data before suppression and data already suppressed with the most significant antibodies. Table 6 shows results obtained from other common classifiers: Naive Bayes, Multi-Layer Perceptron and Decision Tree [3].

Table 4. Results from the predictor without suppression mechanism for Pima Indians Diabetes Database with 460 antibodies

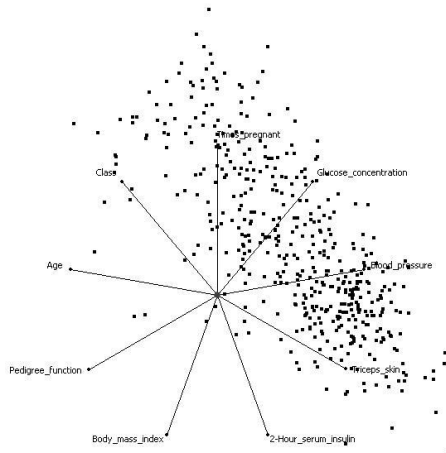
percentage split	distance	acc	val	sens	spec	prec	Fmea	GSP	GSS
60-20-20	Euclidian	0.74	0.67	0.7	0.62	0.77	0.36	0.73	0.66
60-20-20	Manhattan	0.71	0.69	0.71	0.65	0.79	0.37	0.75	0.68
70-20-10	Euclidian	0.70	0.69	0.74	0.61	0.74	0.37	0.74	0.67
70-20-10	Manhattan	0.73	0.71	0.74	0.68	0.77	0.38	0.76	0.71

Table 5. Results from the predictor with suppression mechanism for Pima Indians Diabetes Database with 80 antibodies

percentage split	distance	acc	val	sens	spec	prec	Fmea	GSP	GSS
60-20-20	Euclidian	0.74	0.65	0.77	0.45	0.71	0.37	0.74	0.59
60-20-20	Manhattan	0.71	0.65	0.8	0.4	0.7	0.37	0.74	0.56
70-20-10	Euclidian	0.70	0.65	0.72	0.55	0.70	0.35	0.71	0.63
70-20-10	Manhattan	0.73	0.66	0.80	0.45	0.68	0.37	0.74	0.60

Table 6. Results from other classifiers to diabetes database

classifier	acc
J48 (Decision Tree)	0.76
Multi-Layer Perceptron	0.75
Naive Bayes	0.70

**Fig. 5.** The total amount of antibodies (training set) for the Diabetes database

3.3 Iris Database

The following database is characterized by having more than two classes. Iris database has 150 registers, 4 attributes and 3 classes. Although there are some other metrics to evaluate multi-class problems [2], in this work it will be adopted only accuracy on test and validation sets. As a continuation of this work, it is intended to use metrics proposed by [2].

In Tables 7 and 8 are shown the results for the iris database. Although in the first case the accuracy and validation test are classified, respectively, as 97% and 100% correct and in Table 8 shows an inferior performance, the decrease of the antibodies is significant – from 90 to 10.

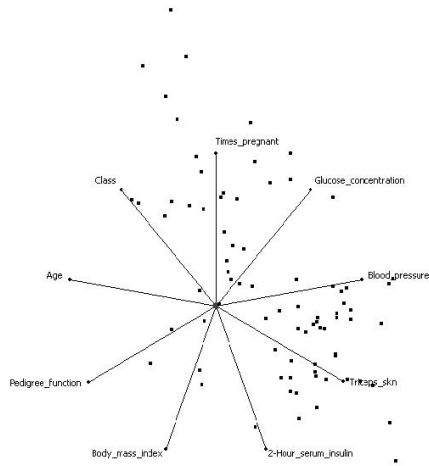


Fig. 6. Antibodies from Diabetes database after suppression and 60-40% split

Table 7. Results for the Iris Database with 90 antibodies without suppression

percentage split	distance	acc	val
60-20-20	Euclidian	0.97	1
60-20-20	Manhattan	0.97	1
70-20-10	Euclidian	1	1
70-20-10	Manhattan	1	1

Table 8. Results for the Iris Database with 10 antibodies with suppression

percentage split	distance	acc	val
60-20-20	Euclidian	0.96	0.97
60-20-20	Manhattan	0.97	0.97
70-20-10	Euclidian	1	1
70-20-10	Manhattan	1	1

Table 9. Results from other classifiers to iris database

classifier	acc
J48 (Decision Tree)	0.95
Multi-Layer Perceptron	0.97
Naive Bayes	0.95

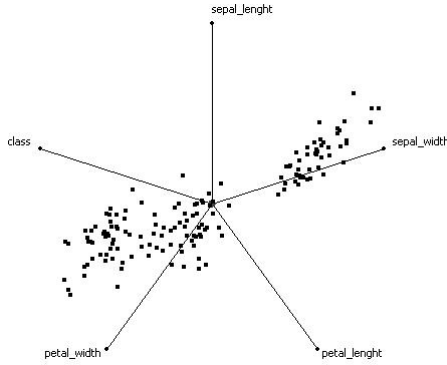


Fig. 7. The total amount of antibodies (training set) for the Iris database

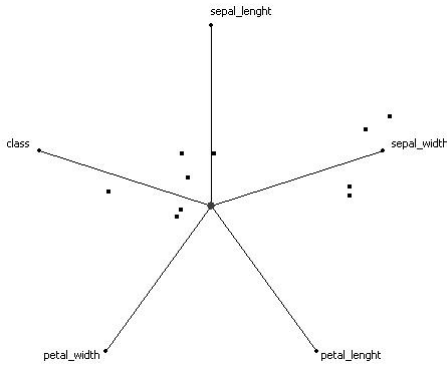


Fig. 8. Antibodies from Iris database after suppression and 60-40% split

4 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. Particularly, the idea came from the way lymph nodes and T-helper cells behave towards lymphocytes no longer needed in the organism or those who self attack.

Basically, the predictor works with a database divided into three parts. The first one represents the antibodies and have the great proportion of the studied database, so that it can cover all the space. The second is the test set, responsible to help the system to eliminate those antibodies that were not used. The principle is, basically, those antibodies who recognized presented antigens remain on the population. The other ones are suppressed – taken away from the population. The third set is the validation one. Its role is to make sure that the suppression mechanism worked, retesting the system.

Experiments were made using benchmarks with two or more classes and the results were satisfactory. There was just a little proportion of mistakes between test and validation set. As next steps of this work, it would be necessary to further investigate multi-class problems and databases with unbalanced examples. Also, apply some hybrid techniques, such as fuzzy logic [30] or support vector machines [31] to try to reduce cross-reaction and improve performance.

References

1. Gordon, A.: Classification. Chapman and Hall, London (1981)
2. Espínola, R.P., Ebecken, N.F.F.: On extending f-measure and g-mean metrics to multi-class problems. *DATA MINING VI - Data Mining, Text Mining and Their Business Applications* 1, 25–34 (2005)
3. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco (2001)
4. Mitchell, T.: *Machine Learning*. McGraw-Hill, New York (1997)
5. Watkins, A.: AIRS: A resource limited immune classifier. Master's thesis, Mississippi State University (2001)
6. Watkins, A.B., Boggess, L.C.: A resource limited artificial immune classifier. In: *Congress on Evolutionary Computation, IEEE World Congress on Computational Intelligence*, Honolulu, HI, USA, pp. 926–931. IEEE Computer Society Press, Los Alamitos (2002)
7. Watkins, A.B., Boggess, L.C.: A new classifier based on resource limited artificial immune systems. In: *Congress on Evolutionary Computation, IEEE World Congress on Computational Intelligence*, Honolulu, HI, USA, pp. 1546–1551. IEEE Computer Society Press, Los Alamitos (2002)
8. Watkins, A., Timmis, J.: Artificial immune recognition system AIRS: Revisions and refinements. In: Timmis, J., Bentley, P. (eds.) *1st Intl. Conference on Artificial Immune Systems (ICARIS2002)*, pp. 173–181. University of Kent (2002)
9. Goodman, D., Boggess, L., Watkins, A.: An investigation into the source of power for airs, an artificial immune classification system. In: *Intl. Joint Conference on Neural Networks (IJCNN'03)*, Portland, OR, USA, pp. 1678–1683 (2003)
10. Watkins, A., Timmis, J., Boggess, L.: In: *Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm*, vol. 5, pp. 291–317. Springer, Netherlands (2004)
11. Watkins, A., Timmis, J.: Exploiting parallelism inherent in AIRS, an artificial immune classifier. In: Nicosia, G., Cutello, V., Bentley, P.J., Timmis, J. (eds.) *ICARIS 2004*. LNCS, vol. 3239, pp. 427–438. Springer, Heidelberg (2004)
12. Watkins, A.: *Exploiting Immunological Metaphors in the Development of Serial, Parallel, and Distributed Learning Algorithms*. PhD thesis, University of Kent, Canterbury, UK (2005)
13. Marwah, G., Boggess, L.: *Artificial immune systems for classification: Some issues* (2002)
14. Carter, J.H.: The immune system as a model for pattern recognition and classification. *Journal of the American Medical Informatics Association* 7, 28–41 (2000)
15. Timmis, J., Neal, M., Hunt, J.: An artificial immune system for data analysis. *Biosystems* 55, 143–150 (2000)
16. Timmis, J., Neal, M.: A resource limited artificial immune system for data analysis. *Knowledge Based Systems* 14, 121–130 (2001)

17. de Castro, L.N., Zuben, F.J.V.: Ainet: An artificial immune network for data analysis. In: Hussein, A., Abbass, R.A.S., Newton, C.S. (eds.) *Data Mining: A Heuristic Approach*. Idea Group Publishing, USA (2001)
18. Timmis, J., Neal, M.J.: A Resource Limited Artificial Immune System for Data Analysis. In: Timmis, J., Neal, M.J. (eds.) *Research and Development in Intelligent Systems XVII. Proceedings of ES2000*, Cambridge, UK, pp. 19–32 (2000)
19. Knight, T., Timmis, J.: Aine: An immunological approach to data mining. In: Cercone, N., Lin, T., Wu, X. (eds.) *IEEE International Conference on Data Mining*, San Jose, CA. USA, pp. 297–304. IEEE, New York (2001)
20. Timmis, J., Neal, M.: A resource limited artificial immune system for data analysis. *Knowledge Based Systems* 14, 121–130 (2001)
21. Knight, T., Timmis, J.: Assessing the performance of the resource limited artificial immune system AINE. Technical Report 3-01, Canterbury, Kent. CT2 7NF (2001)
22. Timmis, J., Knight, T.: Artificial immune systems: Using the immune system as inspiration for data mining. In: Abbass, H.A., Sarker, R.A., Newton, C.S. (eds.) *Data Mining: A Heuristic Approach*, pp. 209–230. Group Idea Publishing (2001)
23. de Castro, L.N.: *Immune Engineering: Development of Computational Tools Inspired by the Artificial Immune Systems* (in Portuguese). PhD thesis, DCA FEEC/UNICAMP, Campinas/SP, Brazil (August 1998 to May 2001)
24. de Castro, L.N., Timmis, J.: *Artificial Immune Systems: A New Computational Intelligence Approach*, 1st edn. vol. 1 (2002)
25. Janeway, C.A., Travers, P., Walport, M., Shlomchik, M.: *Immunobiologia: O sistema imune na saúde e na doença* (in Portuguese), 5th edn. Artes Médicas (2001)
26. Perelson, A.S., Oster, G.F.: The shape space model. *Journal of Theoretical Biology* 81, 645–670 (1979)
27. Perelson, A.S., Weisbuch, G.: Immunology for physicists. *Reviews of Modern Physics* 69, 1219 (1997)
28. Voronoi, G.: Nouvelles applications des paramètres continus á la théorie des formes quadratiques. *J. für die Reine und Angewandte Mathematik* 133, 97–178 (1907)
29. Newman, D.J., Hettich, S., Merz, C.B.C.: *UCI repository of machine learning databases* (1998)
30. Ross, T.J.: *Fuzzy Logic with Engineering Applications*. McGraw Hill, New York (1995)
31. Cristianini, N., Shawe-Taylor, J.: *An Introduction to support vector machines (and other kernel-based learning methods)*. Cambridge University Press, Cambridge (2000)