International Journal of Research in Engineering and Innovation (IJREI)

journal home page: http://www.ijrei.com

ISSN (Online): 2456-6934

An analysis of association rule mining for classification based on artificial immune system

S. M. Zakariya

Electrical Engineering Section, University Polytechnic, Aligarh Muslim University, Aligarh, India

Abstract

One of the most important and effective data mining techniques is association rule mining. Associative classification uses the principle of rule finding and technique of classification to generate a classifier for prediction. The rule search method is also computationally expensive for small support threshold values, which are critical for designing an efficient classifier. The artificial immune system (AIS) employs the powerful informational capacities of the immune system. The population-based search model combined with evolutionary computation techniques allows the artificial immune system clonal selection methodology to manage a complex search space. This study calculated accuracy across a variety of clonal characteristics and generations to assess the efficacy of the artificial immune system-based categorization method. The output of these systems is shown on several benchmark datasets. Based on the accuracy of the different clonal factors (0.1 to 0.9) and generations (10, 20, 30, 40, 50, and 60), a comparison study is performed. The accuracy is computed using four standard datasets. It is observed that in every dataset for several generations, the approach provides the maximum accuracy with a clonal factor of 0.4.

Keywords: Artificial immune system, Association rule mining, Clonal selection algorithm, Accuracy rate

1. Introduction

As everyone knows, data mining is essential to making informed judgments. Finding some useful information hidden in the vast array of databases is the aim of data mining. As a type of data processing, classification entails predicting prospective data patterns or generating models to represent significant data categories. Associative classification in association rule mining is utilized in the rule discovery process to gather high-quality rules that correctly generalize the training dataset. This categorization approach has a high accuracy compared to other methods. Associative classification combines two well-known data mining techniques, association rule mining and classification, to create a model (classifier) for prediction. Classification and association rule mining are related jobs in data mining; however, the primary goal of classification is to forecast class markings, while the purpose of association rule mining is to explain links between items in a transactional database. As seen in Figure 1[2], the three main steps in associative classification are rule discovery, rule collection, and classification. It has previously been investigated to apply AIS algorithms for data mining tasks like grouping, clustering, and regular discovery of itemset. The clonal selection algorithm of AIS is a good approximation and searching algorithm, akin to mutation-based evolutionary algorithms. Therefore, the outcomes of a classification scheme based on AIS and using a clonal selection process are assessed [4]. This work investigates the effects of an AIS-based classification method to determine the optimal accuracy performance of the system for various clonal variables and generations. There are four benchmark databases to practice from: Gait, Codon Bean, and Car. These datasets are all available in the UCI machine learning library [3]. Finding the clonal factor and generation at which the best classification precision may be attained is the goal of the study. The structure of the paper is as follows. A brief analysis of associative categorization schemes is covered in section 2. In section 3, the tactics of artificial immune systems are covered. The categorization paradigm based on the artificial immune system is explained in Section 4. You may find the outcomes analysis in section 5. Lastly, the work's conclusion is given in section 6.

Corresponding author: S. M. Zakariya
Email Id: s.zakariya@gmail.com
https://doi.org/10.36037/IJREI.2017.1637

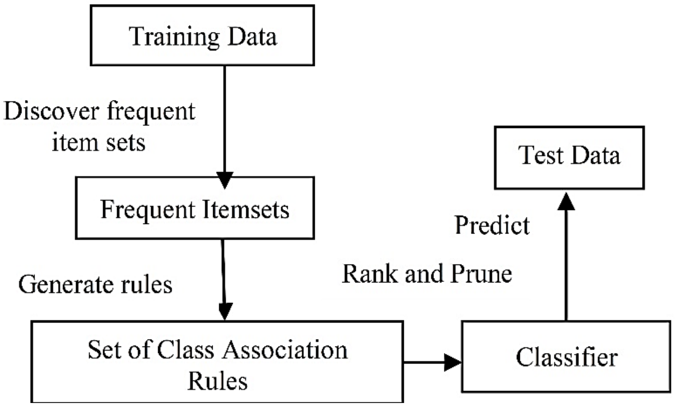

Figure 1: Steps in Associative Classification

2. An analysis of schemes for associative classification

Associative classification, or AC, is a subfield of data processing, which is a larger scientific domain. During the rule discovery phase, it employs association rule mining to gather high-quality rules that can accurately generalize the training dataset. Compared to alternative categorization techniques, this method has demonstrated a high degree of precision. Associative classification generates and analyzes association rules for use in classification. Some of the shortcomings of decision-tree induction, which only takes into account one characteristic at a time, can be resolved by association rules, which search for extremely confident links between various qualities. Associative classification performs better than the majority of conventional classification techniques in some tests. CPAR [8], CMAR [6], and CBA [12] are the three main strategies under analysis.

2.1 Association Based Classification (CBA)

It consists of a classifier function object (named CBA-CB) and a rule generator (called CBA-RG) that is based on the Apriori association rules discovery technique.

The main task of the CBA-RG is to find all rule items with support greater than minsup. The type is a ruleitem: <condset,y> where condset is a set of items y ∈ Y is a label of the class. The condset support count (called condsupCount) is the number of cases in D containing the condset. The number of cases in D that contain the condset and are labelled with Class y is the ruleitem support count (called rulesupCount). Each rule is a rule: condset → y, which supports (rulesupCount / |D|) *100%, where |D| is the dataset size, and whose trust is (rulesupCount/condsupCount)*100%. Rule items satisfying minsup are referred to as regular rule items, whereas the remainders are referred to as infrequent rule items. The rule item with the highest trust is chosen as the possible rule (PR) representing this collection of rule items for all rule items which have the same condset. If there is more than one rule item with the same maximum confidence, pick one ruleitem randomly Known as the rule is correct if the confidence is higher than minconf.

Thus, the collection of class association rules (CARs) comprises all regular and applicable PRs.

The CBA-RG algorithm iteratively processes the data, producing all of the common rules. It determines if it is common by calculating the contribution of each unique rule in the first run. In each subsequent pass, it starts with a seed collection of rule items that were found to be regular in the previous pass. It requires such a seed set to produce new rule items, called candidate rule items, which may be frequent. During the processing of the results, the actual support for these candidate rule items is determined. It generates the rules from this set of frequent rules (CARs) [12] and [13].

2.2 Multiple association rules-based classifications (CMAR)

The class name is defined by CMAR through a set of rules. In light of a fresh case for prediction, CMAR chooses a small group of highly confident, closely related rules and examines their relationship. A thorough performance analysis shows that CMAR outperforms CBA in terms of prediction accuracy overall. CMAR employs a novel data structure called CR-tree to efficiently store and retrieves a large number of classification rules, enhancing both performance and accuracy. A prefix tree architecture for investigating rule sharing that achieves a high degree of solidity is called the CR-tree. Another rule indexing method that can be used to retrieve rules is the CR-tree. CMAR uses a variant of the recently developed FP-growth methodology to accelerate the mining of an entire rule set. Compared to Apriori-like techniques employed in the prior association-based group, FP-growth is substantially faster when there are a lot of rules, big training data sets, and lengthy pattern rules [6].

2.3 Classification based on rules for predictive association (CPAR)

CPAR uses the following features to enhance its performance and accuracy: To prevent redundant computations during the rule-generation process, CPAR uses dynamic programming. All near-to-best literals are taken into account during rule-generation, preventing the omission of crucial rules. CPAR generates a smaller set of laws with better consistency and less redundancy than associative categorization. Because of this, CPAR maintains the same degree of accuracy as associative classification but saves time in both rule creation and prediction [8].

2.4 Steps in classification associative

There are four processes involved in creating an associative classification classifier [9].

- The list of every frequently occurring rule item.

- The output of every confidential CAR from the frequently occurring rule items that were removed in Step 1 and over the minconf level.
- Choosing a subset of CARs to build the classifier from those generated in Step 2.
- Assessing the generated classifier's accuracy using test data artifacts.

3. Artificial Immune System Techniques

The main function of a biological immune system is to protect the body from foreign substances called antigens. Individuality, autonomy, distributed detection, international identification, and noise tolerances are just a few of the characteristics of immune systems. Numerous applications, including pattern recognition, fault detection, computer protection, and many more, use the various AIS models [15] and [16].

3.1 Methods that rely on clonal selection

Burnet put forward the hypothesis of clonal selection in 1959 [1]. This theory describes the adaptive immune system's antigenic stimulus-response mechanism. It gives rise to the notion those only cells that are able to recognize an antigen can multiply, even in the face of other cells being selected. Numerous artificial immune algorithms that imitate the clonal selection idea have been created [10]. Figure 2 displays the definition of the clonal selection algorithm.

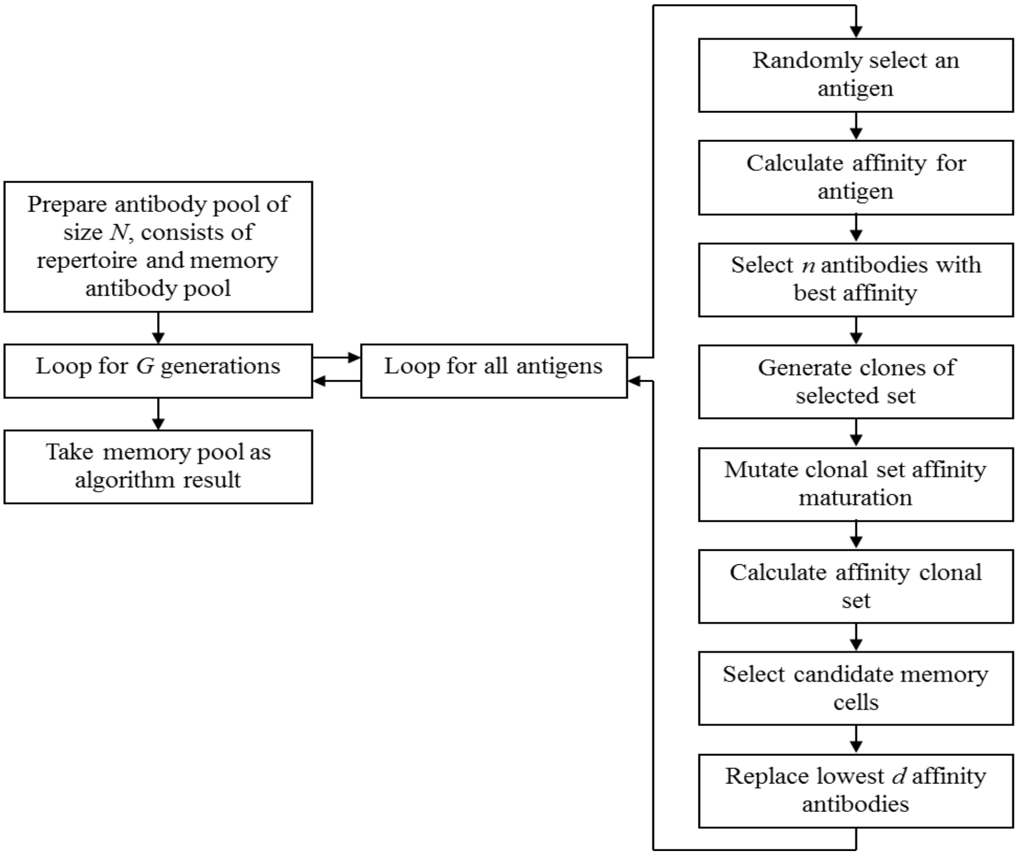
Prepare antibody pool of size N, consists of repertoire and memory antibody pool
Loop for G generations
Take memory pool as algorithm result
Loop for all antigens
Randomly select an antigen
Calculate affinity for antigen
Select n antibodies with best affinity
Generate clones of selected set
Mutate clonal set affinity maturation
Calculate affinity clonal set
Select candidate memory cells
Replace lowest d affinity antibodies

Figure 2: Clonal Selection Algorithm Description

3.2 Methods Utilizing Negative Selection

One of the natural immune system's mechanisms, negative selection, served as the impetus for the development of many contemporary artificial immune systems. When a T-cell in the thymus identifies any self-cell during the immune system's T-cell maturation phase, it is eliminated before it can be used for immunological activity. Any detector candidate that matches items from a collection of self-samples is eliminated by the negative selection method in order to produce a detector set. Negative selection-based algorithms have been employed in a variety of application areas, including anomaly detection. His algorithm's primary goal is to generate a large number of detectors by randomly selecting candidates and then eliminating those that recognize self-data training. Later on, an abnormality can be found with the use of these detectors. Artificial Negative Selection Classifier (ANSC) is a new negative selection technique for multi-class classification proposed by Igawa and Ohashi [11]. This features a cutting mechanism that lessens the sound's impact.

4. Artificial Immune System based Classification

Associative classification uses association rule mining to find association rules in a database of transactions. The rule discovery process is very exhaustive due to the large search space. Artificial immune system algorithms have good features for problem search optimization. The schematic diagram of the AIS-based classification scheme is shown in figure 3. The cloning procedure is carried out in such a manner that a rule's clonal rate is directly proportional to its affinity, and the average value of each rule's clonal rate is equal to the clonal rate of the user [20].

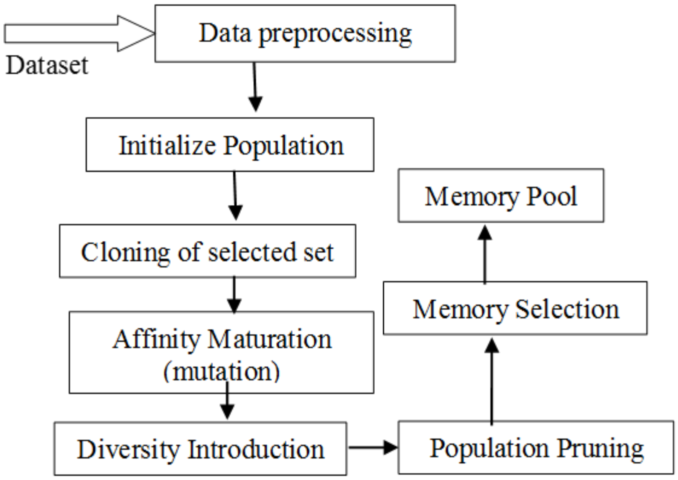

Figure 3: The design of the classification system based on the artificial immune system

1.2. Cloning of selected ruleset

The cloning procedure is done in such a way that the clonal rate of a rule is proportional to its confidence (i.e., affinity), and the average value of each rule's clonal rate is equal to the user's clonal Rate. Let denotes the clonal rate of a rule R as clone Rate(R) and R_1, R_2,R_3, …, R_n are rules selected at a certain generation. Since a rule's clonal rate is directly proportional to the confidence value, the Ri's clonal rate is equal to its confidence value, multiplied by a constant A.

$$cloneRate(R_i) = A \times confidence(R_i)(1)$$

Since the average value of each rule's clonal rate is equal to clonal Rate, put as:

$$clonalRate = \frac{1}{n} \times \sum_{i=1}^{n} cloneRate(R_i) \qquad (2)$$

or

$$clonalRate = \frac{1}{n} \times A \times \sum_{i=1}^{n} confidence(R_i) \qquad (3)$$

Thus,

$$A = \frac{n \times clonalRate}{\sum_{i=1}^{n} confidence(R_i)} \qquad (4)$$

2. Results and Discussion

In this paper, the Waikato Environment for Knowledge Analysis tool, or WEKA, is used to examine the results. A Java programming language workbench for machine learning, WEKA is developed. Because WEKA is now open-source, researchers and businesses can expand the framework by adding algorithm and tool plug-ins for the platform [5].

5.1 Dataset used

The four reference datasets are used from the UCI machine learning repository [3], namely Gait Classification, Codon Usage, Dry Bean, and Car Evaluation. These datasets vary in the number of classes, samples, number of items, number of attributes, and number of training and test datasets. The Gait Classification and Car Evaluation datasets are very small datasets. The Gait dataset has 48 samples only with 34 samples as a training set and 14 samples as a test set. The Gait dataset has 4 different classes with 321 attributes and 24 items. The Car dataset has 1728 samples with 1210 samples as a training set and 518 samples as a test set. The car dataset has 4 different classes with 6 attributes and 21 items. The Codon Usage and Dry Bean datasets are big datasets. The Codon dataset has 13028 samples with 9120 samples as a training set and 3900 samples as a test set. The Codon dataset has 6 different classes with 69 attributes and 9 items. The Dry Bean dataset has 13611 samples with 9528 samples as a training set and 4083 samples as a test set. The Bean dataset has 7 different classes with 17 attributes and 32 items. The summary of these datasets is given in Table 1. The results are tested at varying clonal factors from 0.1 to 0.9 and at a different number of generations like 10, 20, 30, 40, 50, and 60.

Table 1. Summary of all four datasets used

Dataset name	Attributes in numbers	Items in Numbers	Classes in numbers	Instances in numbers	Training set	Test set
Gait Classification	321	24	4	48	34	14
Codon Usage	69	9	6	13028	9120	3900
Dry Bean	17	32	7	13611	9528	4083
Car Evaluation	6	21	4	1728	1210	518

5.2 Evaluation parameters

Table 2 shows a confusion matrix that includes details on the actual and predicted classifications produced by a classifier [13].

Table 2. Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	P	Q
	Positive	R	S

The percentage of cases in which the test data collection was correct is known as accuracy.

$$Accuracy= ((P+S))/((P+Q+R+S)) \qquad (5)$$

The proportion of correctly described positive cases is known as the True Positive Rate (TPR).

$$TPR=S/((R+S)) \qquad (6)$$

The False Positive Rate (FPR) is the percentage of negative cases that are reported as positive when they are not:

$$FPR=Q/((P+Q)) \qquad (7)$$

The proportion of correctly classified negative cases is known as the True-Negative Rate (TNR).

$$TNR=P/((P+Q)) \qquad (8)$$

The False Negative Rate (FNR) is the proportion of positive cases that are incorrectly classified as negative:

$$FNR=R/((R+S)) \qquad (9)$$

Where P represents the proportion of correctly predicted negative objects, Q represents the proportion of falsely predicted positive objects, R represents the proportion of incorrectly predicted negative objects, and S represents the proportion of positive cases.

Table 3. Average accuracy on four datasets with varying generation at a fixed clonal factor 0.4 by CLONALG

No. of Generation	Gait	Codon	Bean	Car
10	97.566	73.846	72.848	82.736
20	99.025	73.250	72.505	82.694
30	98.823	71.876	71.845	81.720
40	98.045	73.996	72.226	83.885
50	97.755	71.075	70.985	81.965
60	98.224	72.322	71.763	82.052

5.3 Results using the Car, Bean, Gait, and Codon Datasets

Table 3 displays the accuracy of the clonal approach using fixed clonal factor 0.4 on the Gait, Codon, Bean, and Car datasets with 3-fold cross-validation for various generations. The approach provides optimal accuracy on all four datasets at clonal factor 0.4, which is why it was selected.

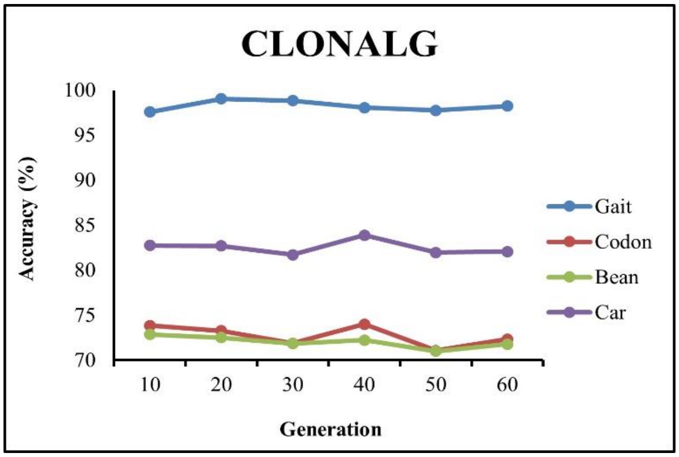

Figure 4: Accuracy on different generation for fixed Clonal factor 0. 4 on Gait, Codon, Bean, and Car Datasets by CLONALG

Figure 4 shows the graphical depiction of consistency in all four datasets using 3-fold cross-validation for different generations and at a constant clonal factor of 0.4. The figure below illustrates the greatest classification accuracy of generation 30 for the Gait Classification dataset at a fixed clonal factor of 0.4. For each of the four datasets, Table 4 displays the classification accuracy for a range of clonal variables at the highest accuracy possible across generations.

Table 4. Average accuracy on four datasets on varying clonal factor 0.1 to 0.9

Clonal Factor	Gait	Codon	Bean	Car
0.1	96.258	70.243	70.626	80.737
0.2	96.975	71.042	71.729	81.830
0.3	96.158	73.014	70.380	80.491
0.4	99.025	73.996	72.848	83.885
0.5	95.168	72.094	71.583	81.694
0.6	97.052	70.945	72.266	82.377
0.7	96.992	71.318	72.027	82.130
0.8	91.247	72.442	71.173	81.284
0.9	96.152	70.745	71.724	81.830

Figure 5. Accuracy on varying Clonal Factor at fixed generations 20, 40, 10, 40, 30, and 20 for Gait, Codon, Bean and Car datasets respectively. The graphical depiction of classification accuracy on four distinct datasets with different clonal factors—Gait, Codon, Bean, and Car—is shown in Figure 5. This graph shows that it gets the greatest classification accuracy on all datasets with a clonal factor of 0.4. The greatest accuracy using 3-fold cross-validation on all four datasets is displayed in Table 5 at a fixed clonal factor of 0.4. In each of the four datasets, maximum accuracy is attained in different generations. The graphical representation of Table 5 is shown in Figure 6.

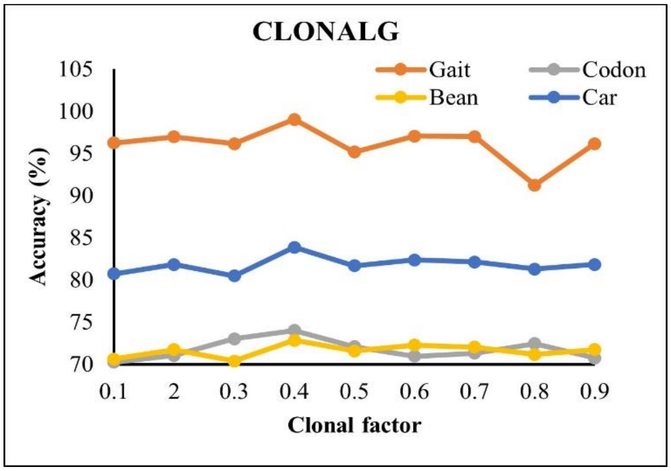

Figure 5: Accuracy on varying Clonal Factor at fixed generations 20, 40, 10, 40, 30, and 20 for Gait, Codon, Bean and Car datasets respectively

Table 5. Comparisons of result at maximum accuracy for all four datasets

Dataset	Training Set	Test Set	Generation	Accuracy (%) at 0.4 clonal factor
Gait	34	14	20	99.025
Codon	9120	3900	40	73.996
Bean	9528	4083	10	72.848
Car	1210	518	40	83.885

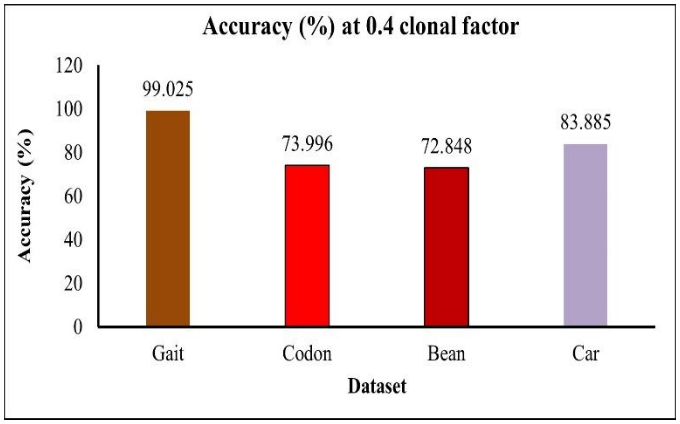

Figure 6: Maximum accuracy at fixed clonal factor 0.4 on all four datasets with 3-fold cross-validation

Based on this, it was determined that the Gait dataset has the highest classification accuracy of the four datasets.

3. Conclusions

Four benchmark datasets are used in this analysis to assess the system's performance: gait classification, codon usage, dry bean, and car evaluation. Accuracy is used as an evaluation criterion to assess the performance over a range of clonal variables and generations. The highest accuracy is attained on each dataset at a clonal factor of 0.4. It is clear from the results that the accuracy fluctuates randomly with a variable number of generations with a fixed clonal factor. Additionally, it has been noted that as dataset sizes increase, categorization accuracy declines. Because of its tiny size, the Gait classification dataset exhibits the highest accuracy of categorization. Because the car dataset is the second lowest in size, it has the second greatest classification accuracy. It might be tried on more datasets in the future.

References

[1] Burnet F.M., "The Clonal Selection Theory of Acquired Immunity", Cambridge University Press, 1959.

[2] B. Liu, H. Hsu, and Y. Ma, "Integrating classification and association rule mining", in Proc. 4th Int. Conf. Knowledge Discovery Data Mining, pp. 80–86, 1998.

[3] D. J. Newman, S. Hettich, C. Blake, and C. Merz, "UCI Repository of Machine Learning Databases", Berkeley, CA: Dept. Information Comput. Sci., University of California, 1998.

[4] L. N. d. Castro and F. J. V. Zuben, "The clonal selection algorithm with engineering applications," in Proceeding Workshop Artificial Immune System and Their Application (GECCO'00), pp. 36–37, 2000.

[5] Ian H. Witten and Eibe Frank, "Data Mining: Practical machine learning tools with Java implementations", San Francisco: Morgan Kaufmann, 2000.

[6] W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules", in Proc.Of IEEE Int. Conference on Data Mining (ICDM '01), pp.369 – 376, 2001.

[7] L. N. d. Castro and F. J. V. Zuben, "Learning and optimization using the clonal selection principle", IEEE Transaction on Evolutionary Computation, vol. 6, no. 3, pp.239–251, 2002.

[8] X. Yin and J. Han, "CPAR: Classification based on predictive association rules," in Proc. 2003 SIAM Int. Conf. Data Min. (SDM'03), 2003.

[9] Brownlee J., "Clonal Selection Theory & CLONALG Clonal selection Classification Algorithm (CSCA)", Technical report No. 2-02, January 2005.

[10] F. Campelo, F.G. Guimarães, H. Igarashi, J.A. Ramírez, "A clonal selection algorithm for optimization in electromagnetic", IEEE Trans. Magn., 41, pp. 1736-1739, 2005.

[11] Igawa K. and Ohashi H., "A Negative Selection Algorithm for Classification and Reduction of the Noise Effect", Appl. Soft Comput. Journal, 2008.

[12] T. D. Do, S.C. Hui, A.C.M. Fong, and Bernard Fong, "Associative Classification With Artificial Immune System", IEEE Transactions on Evolutionary Computation, vol.13, No. 2, pp. 217-228, 2009.

[13] AftabYaseen, Rashid Ali, M. QasimRafiq, and S. M. Zakariya, "Effect of Varying Clonal Factor and Number of Generation on AIS based Classification," Proc. of the 2011 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 545-548, December 2011.

[14] Sharma A., Sharma D., "Clonal Selection Algorithm for Classification", In: Liò P., Nicosia G., Stibor T. (eds) Artificial Immune Systems (ICARIS 2011), Lecture Notes in Computer Science, vol 6825. Springer, Berlin, 2011.

[15] M. Pavone, G. Narzisi, G. Nicosia, "Clonal selection: an immunological algorithm for global optimization over continuous spaces", J. of Global Optim., 53, pp. 769 – 808, 2012.

[16] G.C. Silva, D. Dasgupta, "A survey of recent works in artificial immune systems", Handbook on Computational Intelligence: Volume 2: Evolutionary Computation, Hybrid Systems, and Applications, World Scientific, pp. 547-586, 2016.