UDC 004-032.26

V.I. Lytvynenko

CLUSTERING ALGORITHM CLONAL SELECTION

Clustering algorithm based on clonal selection principle named clonal selection clustering algorithm (CSCA) is proposed in this paper. The new proposed algorithm is data driven and self-adaptive, it adjusts its parameters to the data to make the classification operation as fast as possible. The performance of CSCA is evaluated by comparing it with the well known K-means, EM, Cobweb DBSCAN algorithms using several real-life data sets. The experiments show that the proposed CSCA algorithm is more reliable and has high clustering precision comparing to traditional clustering methods such as K-means, EM, Cobweb DBSCAN.

Keywords: clonal selection algorithm, clustering, Optimization.

1. Introduction

Clustering technique is an effective tool for exploring the underlying structure of a given data set. Its main objective is to partition a given data set into homogeneous groups (called clusters) in such a way that patterns within a cluster are more similar to each other than patterns belonging to different clusters [1]. Clustering technique have been applied in a wide variety of engineering and scientific disciplines such as medicine, psychology, biology, sociology, pattern recognition, and image processing [2]. Many types of clustering algorithms have been developed. Among them, k-means is perhaps the most popular one, because it can be implemented easily and efficiently. This algorithm and its variant [3] have been successfully employed in many practical clustering problems. However, it easily converges to arbitrary local optima and is unable to deal with non-spherical shaped clusters. Recently, evolutionary algorithm (EA) inspired by biological evolution provides a new idea for clustering analysis, and a series of clustering algorithms based on evolutionary computation have been proposed [4]. Most of the available clustering algorithms are prone to get in local optima and cannot find the proper clustering partition. As an artificial immune system algorithm, CSA has been successfully applied to the study of cluster analysis [5–8]. In order to improve the performance of clustering algorithms based on CSA for clustering multi-class data sets, a modified clustering algorithm based on CSA is proposed in this paper.

2. Optimization approach to clustering

In accordance to [11] an optimization problem consists in maximization or minimization of some function (objective function) $f: S \to \Re$ in a set $S \subseteq \Re^n$. The feasible set S can be either finite or infinite, and can be described with the help of a finite or infinite number of equalities and inequalities or in the form of some topological structure in $\Re^n$. In case global maxima or minima (global extreme) are looked for, we have a global optimization problems, otherwise the problem is of local optimization. When the function f is continuous or piece-wise continuous and the set S is described with the help of functions (constraint functions) that have the same continuity property, we obtain a continuous optimization problem. The methods for solution of certain optimization problem depend mainly on the properties of the objective function and the feasible set. Thus, when we are looking for extrema of a linear function regarded on some polyhedral set, then the methods of linear programming can be applied; when f is a convex function and S is a convex set, we apply methods of convex programming; if the feasible set S is described with the help of an infinite number of equalities or inequalities, the methods of semi-infinite programming should be used, etc.. Let us describe the clustering problem formally. Assume that X is the given a finite set in the $n-$ dimensional space $\Re^n$: $X = \{x^1, x^2, \ldots, x^M\}$, where $x^i \in \Re^n, i = 1,2,\ldots,M$. Let us call the elements of the set X patterns. The aim of cluster analysis is to partition X into a finite number of clusters based on similarity between patterns. As a measure of similarity we use any distance function. Here for the sake of simplicity we consider Euclidean distance $\| \cdot \|_2$ Given a number $q \in N$, we are looking for q subsets C^i, $i = 1,2,\ldots,q$, such that the medium distance between the elements in each subset is minimal and the following conditions are satisfied:

$$C^i \neq 0, i = 1,2,\ldots,q, \tag{2.1}$$

$$X = \bigcup_{i=1}^{q} C^i \tag{2.2}$$

The set C^i, $i = 1,2,\ldots,q$ introduced above are called clusters and the problem of determination of clusters is the clustering problem. When the clusters can overlap, the clustering problem is fuzzy. If we request additionally

$$C^i \bigcap C^j = 0 \text{ if } i \neq j; \quad i,j = 1,2,\ldots,q \tag{2.3}$$

 ISSN 1562-9945

then we obtain a hard clustering problem. Let us assume that each cluster C^i, $i = 1, 2, \ldots, q$, can be

identified by its center or centroid, defined as [12] $c^i = \frac{1}{|C^i|} \sum_{x \in C^i} x$,

where $|C^i|$ denotes a cardinality of the cluster C^i. Then the clustering problem can be reduced to the following optimization problem, which is known as a minimum sum of squares clustering [13]:

$$\min \frac{1}{M} \sum_{i=1}^{q} \sum_{x \in C^i} \left\| c^i - x \right\|_2^2 , \qquad (2.4)$$

such that $C = \{C^1, C^2, \ldots, C^q\} \in \overline{C}$.

Here, $\overline{C}$ is a set of all possible q — partitions of set X, $c = (c^1, c^2, \ldots, c^q) \in \Re^{n \times q}$

According to [14], clustering algorithms differ each from other depending on the input data representation, e.g., pattern-matrix or similarity matrix, or data type, e.g., numerical, categorical, or special data structures, such a rank data, strings, graphs, etc.;

- the output representation, e.g., a partition or hierarchy of partitions;
- optimization method chosen for solution of optimization model;
- clustering direction, e.g., agglomerative or divisive.

3. A Clonal Selection Algorithm for Clustering

3.1. Clonal Selection Theory

The clonal selection theory was originally proposed by Burnet, in order to explain the reinforcement learning of the immune system of mammals [15]. The theory of clonal selection [16], suggests that B and T lymphocytes that are able to recognize the antigen, will start to proliferate by cloning upon recognition of such antigen. When a B cell is activated by binding an antigen (and a second signal is received from T lymphocytes), many clones are produced in response, via a process called *clonal expansion*. The resulting cells can undergo *somatic hypermutation*, creating offspring B cells with mutated receptors. The higher the affinity of a B cell to the available antigens, the more likely it will clone. This results in a Darwinian process of variation and selection, called *affinity maturation*. The increase in size of these populations couples with the production of cells with longer than expected *lifetimes*, assuring the organism a higher specific responsiveness to that antigenic attack in the future. This gives rise to *immunological memory* which is demonstrated by the fact that, when the host is first exposed to the antigen, a *primary* response is initiated; in this phase the antigen is recognized and immune memory is developed. When the same antigen is encountered in the future, a *secondary* immune response is initiated. This results from the stimulation of cells already specialized and present as memory cells: a rapid and more abundant production of antibodies is observed. The secondary response can be elicited from any antigen that is similar, although not identical, to the original one that established the memory. This is known as cross-reactivity.

The main property of the clonal selection theory can be summarized as follows [17]: 1) *Negative selection:* elimination of self-antigens; 2) *Clonal expansion:* proliferation and differentiation; 3) *Monospecificity:* phenotypic restriction; 4) *Somatic hypermutation:* new random genetic changes; 5) *Autoimmunity*.

3.2. Clonal Selection Algorithm

The clonal selection algorithm, CSA, was first proposed by de Castro and Von Zuben in [18] and was later enhanced in their 2001 paper [19] and named CLONALG. The algorithm takes a population of antibodies and by repeated exposure to antigens, over a number of generations, develops a population more sensitive to the antigenic stimulus [20].

The artificial immune systems (including clonal selection algorithm) is composed of the following basic elements [9, 10]: (1) a representation for the components of the system; (2) a set of mechanisms to evaluate the interaction of individuals with the environment and each other. The environment usually simulated by a set of input stimuli or patterns, one or more fitness function(s) or other means; and (3) procedures of adaptation that govern the dynamics and metadynamics of the system, i.e. how its behavior varies over time. Under the affinity [17] we could understand the interaction measure (or the power connection) of the complementary areas of antigen and antibodies or of two antibodies, which formally can be represented as one of the metrics (e.g. Euclidean distance). This measure indicates the degree of similarity or differences between the appropriative attributes of lines such as $S^P \times S^P \to \Re^+$.Often

 ISSN 1562-9945

the type of attributes is defined by the subject area of AIS and is an important point in the degree of affinity determination.

Generalized function of affinity can be represented as the following expression [21]:

$$c_{ij} = f\left(x_i, x_j\right) , \tag{3.1}$$

where c_{ij} – is the affinity between molecules *i* and *j*; x_i and x_j – vectors that represent molecules in space forms, f – corresponding affinity function.

In optimization problems, the generalized form of antibodies is a vector of arguments $Ab = (x_1, x_2, ..., x_l)$, and as antigens used optimality criteria y_j, expressed as functions $Ag = f(x_1, x_2, ..., x_l)$. Affinity values g_j calculated on the basis of criteria values y_j reflected in the set of nonnegative numbers such as:

$$f : X \to \Re, \quad F : \Re \to \Re^+ \tag{3.2}$$

Thus, there is some affinity function $g = F\big(f(x_1, x_2, ..., x_n)\big)$, that determines that determines the degree of conformity of individuals to each other. In such problems, we can not to operate the notion of distance, so that the best value criteria is previously unknown, and, therefore, we do not know the maximum possible extent to which individuals. Thus, the control dynamics of AIS is performed by the relative affinity values or by rank individuals set. This approach is very close to the concept of suitability (fitness) used in evolutionary algorithms that have some earlier theory of artificial immune systems.

Formally algorithm of clonal selection can be represented as [22]:

$$CLONALG = (P^l, G^k, l, k, m_{Ab}, \delta, f, I, \tau, AG, AB, S, C, M, n, d), \tag{3.3}$$

where P^l is space of search (space of forms); G^k is space representation; l is the length of vector of attributes (dimension of space of search); k is the length of antibody receptor; m_{Ab} is dimension of population of antibodies; δ is the expression function; f is the affinity function; I is the function of initialization of the initial population of antibodies; τ is the condition of completion of algorithm work; AG is the subset of antigenes; AB is population of antibodies; S is the operator of selection; C is the operator of cloning; M is the mutation operator; n is the number of the best antibodies selected for cloning; d is the number of the worst antibodies subjected to substitution for new ones.

Consider the shape- space $\left(P^l\right)$ phenotypes and their space images as antibodies $\left(G^k\right)$ or genotype space [22].

3.3. Clustering Algorithm Based on Clonal Selection

Traditional clustering algorithms find the data distribution in such a way that the cluster centers are placed in more or less the middle of the data subset, that constitutes the cluster. When used for classification, these cluster centers are sometimes not the best ones, especially when the number of clusters is too small [23]. In our study, we investigated the possibility of using clonal selection algorithm as stochastic optimization methods for clustering. This clustering algorithm based on clonal selection principle is proposed. The clonal selection clustering algorithm (CSCA) is data driven and self-adaptive, it adjusts its parameters to the data to make the classification operation as fast as possible. The proposed approach has been tested on several data sets and its performance is compared with the K-means algorithm. Experiments show that clonal selection clustering algorithm is more reliable and has high classification precision comparing to traditional classification methods such as K-means [24]. In this CSCA, clustering problem is considered as optimization problem and the objective is to find the optimal partitions of data where the resulting clusters tend to be compact as possible. A simple criterion which is the within cluster spread is used in CSCA, this criterion needs to be minimized for good clustering [24]. In difference from K-means which uses the square-error criterion to measure the within cluster spread, CSCA uses the sum of the Euclidean Distances of the points from their respective cluster centroids as clustering metric and uses clonal selection algorithm as clustering algorithm which warrants finding the global optima when most of others algorithms such as K-means stuck into local optima [24]. Clonal selection algorithm and a mathematical description of his basic steps are described in detail to our section 3.2. Figure 3.1. shows the basic structure for the CSCA.

Each antibody in P forms a string of real numbers representing the K cluster centers. For d-dimensional space, the length of the string is $d * K$ number, where the coordinates of the centers are localized in sequence[24].

$$p = [Ab_1, Ab_2, \ldots, Ab_n] \tag{3.4}$$

$$Ab_l = [m_{11}, m_{12}, \ldots, m_{1d}, m_{21}, \ldots, m_{Kd}], l = 1, \ldots, n \tag{3.5}$$

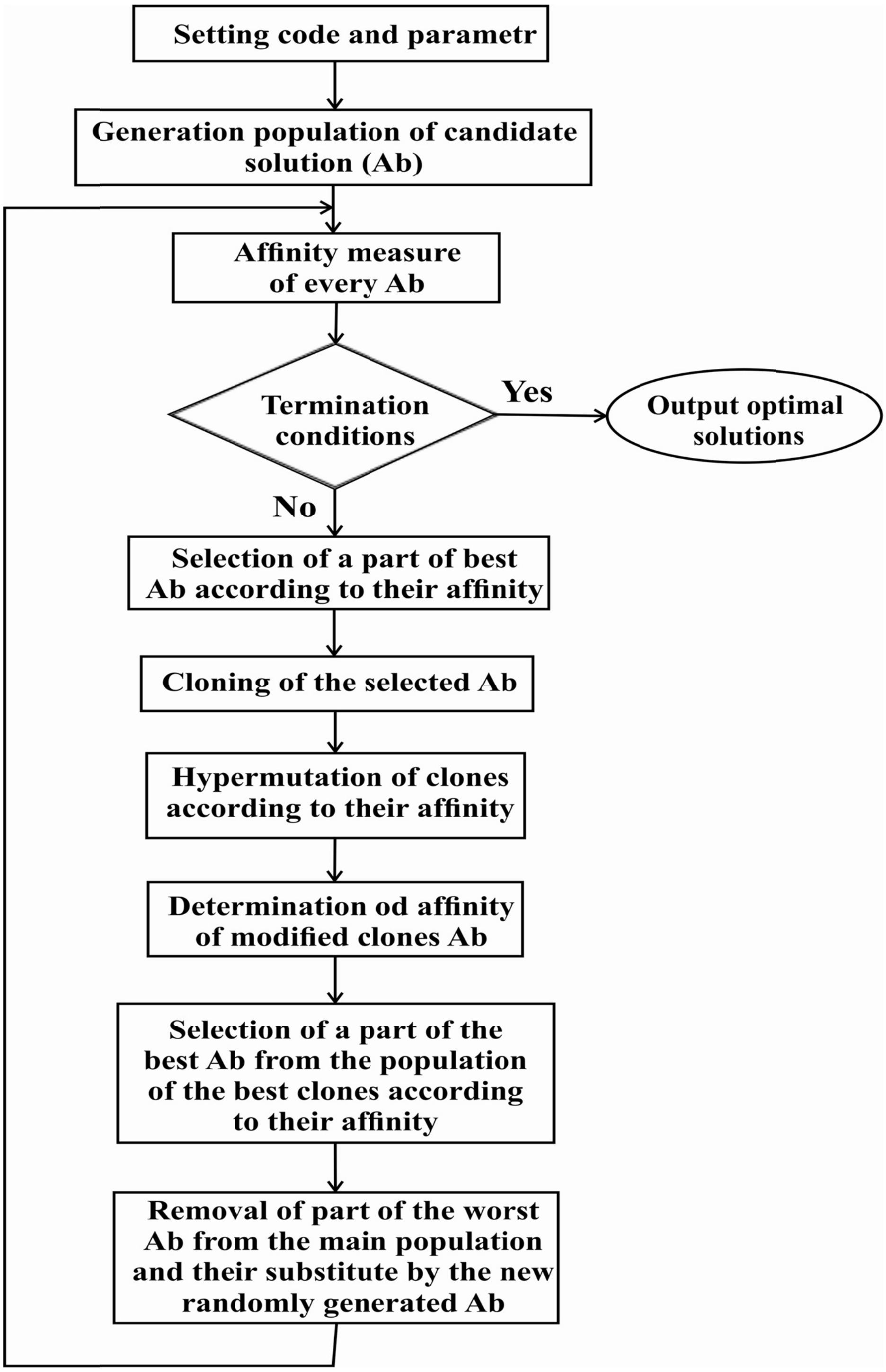

Fig. 3.1 Diagram of clonal selection clustering algorithm

The first d numbers represent the d dimensions of the first cluster center; the next d positions represent those of the second cluster center, and so on.

The number of clusters K is supposed to be known and the appropriate cluster centers $m_1, m_2, \ldots, m_k$ have to be found such that the clustering metric J is minimized. Mathematically, the clustering metric J for the K clusters $C = \{C_1, C_2, \ldots, C_K\}$ is given by the following equation:

$$J(\Gamma, M) = \sum_{i}^{K} \sum_{j}^{N} \gamma_{ij} \left\| x_j - m_i \right\| \tag{3.6}$$

where $x_j \in \Re^d, j = 1, \ldots, N$ are data points, $\Gamma = \{\gamma_{ij}\}$ is a partition matrix witch given by the eq. (2.4) , M is centroid matrix witch given by the eq. (2.5) and $m_i \in \Re^d, i = 1, \ldots, K$ is the mean for the C_i cluster with N_i data points.

$$\gamma_{ij} = \begin{cases} 1 \ if \ x_j \in C_i \ with \sum_{i=1}^{K} \gamma_{ij} = 1 \ \forall j \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{3.7}$$

$$M = [m_1, m_2, \ldots, m_K] \ where \ m_i = \frac{1}{N_i} \sum_{j=1}^{N} \gamma_{ij} x_j, \ i = 1, \ldots, K \tag{3.8}$$

In general, the task of clonal selection algorithm is to search for the appropriate cluster centers wherefore J is minimized. Based on clustering criterion, CSCA is supposed to give right results if the clusters are compact and hyperspherical in shape. To measure the affinity of an antibody, the clusters are formed according to the centers encoded in the antibody under consideration, this is done by assigning each point $x_j \in \Re^d, j = 1, \ldots, N$ to one of the clusters C_i whose center are the closest to the point. After the clustering is done, the new cluster centroids are calculated by finding the mean points of the respective clusters, then clustering criterion J is calculated by eq. (3.6). The affinity is defined as:

$$affinity = \frac{1}{J} \tag{3.9}$$

The maximum value of the affinity standing for the minimum value of J . Zero is assigned to the affinity if any cluster becomes empty. Antibodies will be cloned proportionally to their affinities, the higher the affinity the higher the number of clones generated for the antibody. The antibodies were sorted in descending order according to their affinity and then the amount of clones generated for the antibodies was given by:

 ISSN 1562-9945

$$nc = round(\beta \cdot n) \tag{3.22}$$

where nc is number of clones and β is clonal factor, n is the total amount of antibodies and $round(\cdot)$ is the operator that round its argument towards the closed integer.

Every antibody in *PC* is submitted to a mutation that is inversely proportional to the affinity and this is done according to the following equations: $Ab' = Ab + \alpha N(0,1)$, where $N(0,1)$ is Gaussian number generated by Gaussian function given as $f_{Gaussian}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and α control the learning imposed on the antibodies.

4. Experiments

The CSCA was tested using several artificial and real-life data sets, then compared with the well-known *K*-means, EM, Cobweb DBSCAN, [25]. **K-means clustering algorithm**. In data mining, *k*-means clustering [26] is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster.

EM algorithm. EM algorithm [27] is also an important algorithm of data mining. We used this algorithm when we are satisfied the result of k-means methods. An expectation– maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori estimates of parameters in statistical models, where the model depends on unobserved latent variables.

COBWEB algorithm. The COBWEB algorithm yields a clustering dendrogram called classification tree that characterizes each cluster with a probabilistic description. Cobweb generates hierarchical clustering [29], where clusters are described probabilistically. COBWEB uses a heuristic evaluation measure called category utility to guide construction of the tree. **DBSCAN** is a data clustering algorithm for density-based spatial clustering of applications with noise. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes [30]. The key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts).

3.1 Artificial Data Sets

The experimental results comparing the clonal clustering algorithm with the K-means, EM, COBWEB and DBSCAN algorithms are provided for three real-life data sets (*Vowel*, *Iris* and *Crude Oil*), respectively. The data sets are described below:

• *Vowel data*: This data consists of 871 Indian Telugu vowel sounds [31]. These were uttered in a consonant-vowel-consonant context by three male speakers in the age group of 30-35 years. The data set has three features F_1, F_2 and F_3, corresponding to the first, second and third vowel formant frequencies, and six overlapping classes $\{\delta, a, i, u, e, o\}$. The value of K is therefore chosen to be 6 for this data.

• *Iris data*: This data represents different categories of irises having four feature values. The four feature values represent the sepal length, sepal width, petal length and the petal width in centimeters [32]. It has three classes (with some overlap between classes 2 and 3) with 50 samples per class. The value of K is therefore chosen to be 3 for this data.

• *Crude oil data*: This overlapping data [33] has 56 data points, 5 features and 3 classes. Hence the value of K is chosen to be 3 for this data set.

The was tested with the following parameters of CSCA algorithm: n=70,

m_{AB}=100; d =30 and number of generations *gen*=100. For testing the algorithms K-means, EM, COBWEB and DBSCAN was used WEKA. For K-means algorithm [34] 100 as a maximum number of iterations was used in case it does not terminate normally. For EM algorithm: the number of folds to use when cross-validating to find the best number of clusters — 10; the minimum improvement in log likelihood required to perform another iteration of the E and M steps — 0.000001; the number of folds to use when cross-validating to find the best number of clusters — 10; number of iterations was used-100. For COBWEB algorithm: the random number seed to be used— 42; set the category utility threshold by which to prune nodes— 0.00282; set the minimum standard deviation for numeric attributes — 0.7. For DBSCAN algorithm: minimun number of DataObjects required in an epsilon-range-query — 6; radius of the epsilon-range-queries— 0.9. At every experiment the algorithms were run for 100 times with different random initial configurations To provide statistical evaluation of the performance.

The experiments show that the proposed CSCA algorithm is more reliable because it finds the best solution all the time unlike K-means, EM, COBWEB and DBSCAN which got stuck at sub-optimal solutions.

Table 1

Classification result for dataset

Dataset	CSCA	K-means	EM algorithm	COBWEB	DBSCAN
Vowel data	91%	82%	85%	87%	88%
Iris data	98%	86%	85%	88%	91%
Crude oil data	98%	88%	83%	95%	96%

CSCA algorithm has high classification precision comparing to K-means, EM, COBWEB and DBSCAN algorithms. The clonal algorithm is data driven and self-adaptive, it adjusts its parameters to the data to make the classification operation as fast as possible.

5. Conclusion

In this work, clustering algorithm based on clonal selection principle is designed to find the optimal partition between the data. This algorithm uses within cluster spread criterion as a clustering criterion. The criterion is based on Euclidean distance between the data in the clusters. CSCA is data driven and self-adaptive, it adjusts its parameters to the data to make the clustering operation as fast as possible. CSCA is tested on several data sets and its performance is compared with the well known K-means, EM, COBWEB and DBSCAN algorithms. The experiments show that CSCA algorithm has classification precision higher than K-means, EM, COBWEB and DBSCAN algorithms which got stuck at sub-optimal solutions even for simple data sets. Developed by us CSCA gives good results if the clusters are compact and hyperspherical in shape

REFERENCE

1. R. Liu, X. Zhang, N. Yang, Q. Lei, L. Jiao Immunodomaince based Clonal Selection Clustering Algorithm/ Applied Soft Computing 12 (2012) 302–312
2. A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988, ISBN 0-13-022278-X.

3. J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, vol. 1, University of California Press, 2007, pp. 281–297.
4. I.A. Sarafis, P.W. Trindera, A.M.S. Zalzala, NOCEA: a rule-based evolutionary algorithm for efficient and effective clustering of massive high-dimensional databases, Applied Soft Computing 7 (3) (2007) 668–710.
5. L.N. De Castro, F.J. Von Zuben, An evolutionary immune network for data clustering, in: Proceedings of the IEEE SBRN'00 (Brazilian Symposium on Artificial Neural Networks), Rio de Janerio, Brazil, 2000, pp. 84–89.
6. L.N. De Castro, F.J. Von Zuben, Immune network models: theoretical empirical comparisons, International Journal of Computational Intelligence and Applications 1 (3) (2001) 239–257.
7. J. Li, X.B. Gao, L.C. Jiao, A-CSA-based clustering algorithm for large data sets with mixed numeric and categorical values, in: Proceeding of the World Congress on Intelligent Control and Automation (WCICA), Hang Zhou, China, 2004, pp.2003–2007.
8. F.O. de Francë a, G.P. Coelho, P.A.D. Castro, F.J. Von Zuben, Conceptual and practical aspects of the aiNet family of algorithms, International Journal of Natural Computing Research (IJNCR) 1 (1) (2010) 1–35.
9. L.N. De Castro, F.J. Von Zuben, Learning and optimization using the clonal selection principle, IEEE Transactions on Evolutionary Computation 6 (3) (2002) 239–251.
10. R.C. Liu, Z.C. Shen, L.C. Jiao, W. Zhang, Immunodomaince based clonal selection clustering algorithm, in: Proceedings of the 2010 IEEE Congress on Evolutionary Computation, CEC2010, Barcelona, Spain, July 18–23, 2010, pp. 2912–2918.
11. Tatiana Tchemisova, Başak Akteke-Цztьk, Gerhard Wilhelm Weber On Continuous Optimization Methods in Data Mining — Cluster Analysis, Classification and Regression — Provided for Decision Support and Other Applications/ : Journal of Mathematical Sciences, vol. 120, no. 1, pp. 1016-1031, 2004

 ISSN 1562-9945

12. Bagirov, A.M., and Yearwood, J., A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems, EJOR 170, 2 (2006) 578-596.
13. Bock, H.H., Automatische Klassifikation, Vandenhoeck and Ruprecht, GËOTTINGEN (1974).
14. Jain, A.K., Topchy, A., Law, M.H.C., and Buhmann J.M., Landscape of clustering algorithms, in Proc. IAPR International conference on pattern recagnition, Cambridge, UK (2004) 260-263.
15. F. M. Burnet, The Clonal Selection Theory of Acquired Immunity. Cambridge, U.K.: Cambridge Univ. Press, 1959.
16. Cutello,V., Nicosia,G., Pavone,M. and Timmis,J. (2007) An immune algorithm for protein structure prediction on lattice models. IEEE Trans. Evol. Comput., 11, 101–117.
17. L. N. de Castro and J. Timmis, Artificial Immune Systems: A New Computational Intelligence Paradigm. London, U.K.: Springer-Verlag,2002.
18. Leandro Nunes de Castro and Fernando J. Von Zuben. The clonal selection algorithm with engineering applications. In Workshop Proceedings of GECCO’00, Workshop on Artificial Immune Systems and their Applications, pages 36–37, Las Vegas, USA, July 2000.
19. Leandro Nunes de Castro and Fernando J. Von Zuben. Learning and optimization using clonal selection principle. IEEE Transactions on Evolutionary Computation, Special Issue on Artificial Immune Systems, 6(3):239–251, 2001.
20. Jennifer A. White, Simon M. Garrett Improved Pattern Recognition with Artificial Clonal Selection? Artificial Immune Systems Lecture Notes in Computer Science Volume 2787, 2003, pp 181-193
21. J. Brownlee, "Clonal Selection Algorithms", Technical Report 070209A, Complex Intelligent Systems Laboratory (CIS), Centre for Information Technology Research (CITR), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology, 2007.
22. Бідюк П.І. Литвиненко В.І. Фефелов А.О. Формалізація методів побудови штучних імунних систем// Наукові вісті Національного технічного університету України “Київський політехнічний інститут”/ 2007, №1, - С.29-41.

23. M. Bereta and T. Burczynski ґ (2006), Immune K-means: A novel immune algorithm for data clustering and multiple–class discrimination., in Evolutionary Computation and Global Optimization 2006. 2006. Prace Naukowe, Elektronika, Warsaw Univ. of Technology Publishing House, pp. 49–60
24. M.T. Al-Muallim, R. El-Kouatly Unsupervised Classification using Immune Algorithm International Journal of Computer Applications (0975 – 8887)Volume 2 – No.7, June 2010 pp 44-48
25. Yuni Xia, Bowei Xi ―Conceptual Clustering Categorical Data with Uncertainty‖ Indiana University – Purdue University Indianapolis Indianapolis, IN 46202, USA
26. Jinxin Gao, David B. Hitchcock ―James-Stein Shrinkage to Improve K-meansCluster Analysis‖ University of South Carolina,Department of Statistics November 30, 2009
27. A. P. Dempster; N. M. Laird; D. B. Rubin ―Maximum Likelihood from Incomplete Data via the EM Algorithm‖ Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp.1-38.
28. Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. Computational statistics and data analysis, 14:315–332
29. Sanjoy Dasgupta ―Performance guarantees for hierarchical clustering‖ Department of Computer Science and Engineering University of California, San Diego
30. Timonthy C. Havens. "Clustering in relational data and ontologies" July 2010
31. S.K. Pal, D.D. Majumder, Fuzzy sets and decision making approaches in vowel and speaker recognition, IEEE Trans. Systems, Man Cybernet. SMC-7 (1977) 625}629.
32. R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 3 (1936) 179-188.
33. R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice-Hall, Englewood Cli!s, NJ, 1982
34. R. Xu, et al., Clustering. Hoboken, New Jersey: John Wiley & Sons, Inc, 2009.

 ISSN 1562-9945