Weighted Clone Selection Algorithm based on Rough Set Theory

Jia Wu[1], Zhihua Cai*[1], Xiaolin Chen[1], Meng Li[1], Bin Guo[2]
1. School of Computer Science,
China University of Geosciences, Wuhan 430074, P.R.China,
2. School of Computer Science,
University of Sydney, NSW 2007, Australia
Email: wujiawb@126.com; zhcai@cug.edu.cn; cxlcug@126.com; mlcug@126.com; bguousyd@gmail.com

***Abstract*—Clone selection is a new artificial intelligence technology, with self-organization, self-learning, self-recognition, self-memory capacity. In the traditional clone selection algorithm for data classification, all the attributes for classification have the same influence, which affects its classification performance to some extent, given an appropriate weight for each attribute value can modify this imbalance. Accordingly this, proposes a weighted clone selection algorithm based on rough set to improve the performance of clone selection. In weighted clone selection algorithm attribute weights obtained directly from the training data using rough set theory, the attribute weights was used to test Data classification. Then verify the validity of the method by the experiments of UCI data sets.**

***Index Terms*—Clone selection, Attribute weight, Rough set Theory, Classification.**

I. INTRODUCTION

Classification is one of the data analysis methods in data mining. According to the attribute values of the test instances and other constraints, the target instance can be divided to a data type. Decision tress, Bayesian model and Instance Based method is regarded as traditional data mining technology. With more and more complex problem arises, the solution is not precise enough, so we need to look for a kind of high efficient method. Based on the demand, the artificial neural network (ANN) and genetic algorithm (GA) has been proposed, besides researcher has put forward the artificial intelligence (AI) concept to solve this kind of problem.

Clonal selection is a kind of new artificial intelligence technology, with the ability of self-organizing, learning, recognizing, making memory. It can be used in variety of problems such as space classification, virus detection, robot technology, optimization, etc. The research of Clonal selection that how to improve its performance and experimental result is not get much attention, compared with the wide application of Clonal selection. As we all know, in the real world the contribution of each attribute in process of classification may be different. Therefore, in order to correct the imbalance each attribute should be given the attribute weight to improve the classification accuracy. Attribute weighted method has a wide range of applications in patter recognition. Qin et al [1], put forward a weighted naïve Bayesian classification algorithm. Feng et al [2], propose an attribute weighted K-means algorithm based on mean square error. They both used the attribute weighted methods to improve the performance and classification accuracy of the corresponding algorithms.

Rough Set (RS) first described by a Polish computer scientist Zdzisław I. Pawlak, is a formal approximation of a crisp set in terms of a pair of sets which give the lower and the upper approximation of the original set. In the standard version of rough set theory, the lower- and upper-approximation sets are crisp sets, but in other variations, the approximating sets may be fuzzy sets. Rough set methods can be applied as a component of hybrid solutions in machine learning and data mining. They have been found to be particularly useful for rule induction and feature selection (semantics-preserving dimensionality reduction). Rough set-based data analysis methods have been successfully applied in bioinformatics, economics and finance, medicine, multimedia, web and text mining, signal and image processing, software engineering, robotics, and engineering. In this paper, we plan to calculate the importance of attributes based on the rough set theory and then to determine the weights of the different attributes in clonal selection algorithm. So, the quality of the performance of the clonal selection algorithm can be improved based on the characteristics of the data itself.

II. RELATED WORK

A. Clonal Selection Theory

Clonal selection is part of the primary immune response. A primary immune response is provoked when a new antigen invades the body. Traveling through the circulatory system, the antigen will inevitably meet up with the lymphocyte that has the correct antibody pattern. Clonal selection is an important immunological process

This work was supported by National Natural Science Foundation of China under Grant Nos.61075063, the Fund for Outstanding Doctoral Dissertation of CUG Nos. 2235122, Self-Determined and Innovative Research Funds of CUG Nos. 1210491B16.

*Corresponding author.

doi:10.4304/jsw.8.6.1333-1338

that determines which B and T lymphocytes, types of white blood cells, will be produced in large quantities [3]. It is through this process that our bodies combat antigens—substances it considers to be harmful to it. Niels Jerne, a Danish immunologist, provided the basis for the clonal selection theory in 1955. Prior to Jerne's theory, it was a commonly held belief that our bodies were stimulated to produce a specific antibody when a foreign substance entered it.

Clonal selection is corresponding to a process of affinity maturation. During this process, the lower affinity of individual will be improved gradually after the under the action of copy and variation operation in clonal selection mechanism. So, this affinity maturation in essences is a type of choice and variation of Darwin's process. And the clonal selection principle is realized by the use of variation, cross et al, genetic operator and the corresponding population control mechanism [4].

B. Clonal Selection Algorithm

Clonal selection algorithms are a class of algorithms inspired by the clonal selection theory of acquired immunity that explains how B and T lymphocytes improve their response to antigens over time called affinity maturation. These algorithms focus on the Darwinian attributes of the theory where selection is inspired by the affinity of antigen-antibody interactions, reproduction is inspired by cell division, and variation is inspired by somatic hyper mutation. Clonal selection algorithms are most commonly applied to optimization and pattern recognition domains, some of which resemble parallel hill climbing and the genetic algorithm without the recombination operator [5]. The steps of clonal selection algorithm used for classification is shown in Fig. 1.

- Training

For each training instance $Ag_i = (Ag_{i,1}, Ag_{i,2}, \cdots Ag_{i,k})$, the following steps should be repeated:

（1） Initialize the antibodies group P ,the number of antibody is set to N , antibody group can be divided into the memory antibody group P^s and non-memory antibody group P^m ,the memory antibody group is the clustering center that we want to get at the final step of the algorithm.

（2） Calculate the affinity between Ag_i and each antibody, sorting the antibodies according to affinity. Here the Euclidean distance is used as calculation function for affinity between Ag_i and Ab_j :

$$Affinity = \sqrt{\sum_{k=1}^{L} (Ab_{j,k} - Ag_{i,k})^2} \quad (1)$$

（3） Carry out clone operation using the n individuals with the best affinity from the antibody group P , then bring a new generation of antibodies P_1, of which adopts the cloning of proportion cloning. The best affinity of the antibody, the more times will be cloned.

（4） Do the mutation operation for the new generation antibody group P_1 , mutation probability and affinity is inversely proportional, so the mutate antibody set P_2 can be produced.

（5） Recalculate affinity between antigen antibody Ag_i and a new generation P_2 .

（6） Selection options. Choosing the individual Ab_{best} of the best affinity then compared the affinity between Ab_{best} and the original memory Ab_j ,if $Affinity(Ab_{best}) > Affinity(Ab_j)$,and the category of Ab_{best} is the same as antigen, then Ab_{best} will be into the memory antibodies group and be replaced by Ab_j .

（7） Replace individuals with lower affinity. In order to increase the diversity of antibodies, d antibodies are selected from the antibody group P_2 ,which can be called replacement antibodies. At last, we replace the d antibodies in non-memory antibodies P^s with the lowest affinity using the individuals in the replacement antibodies.

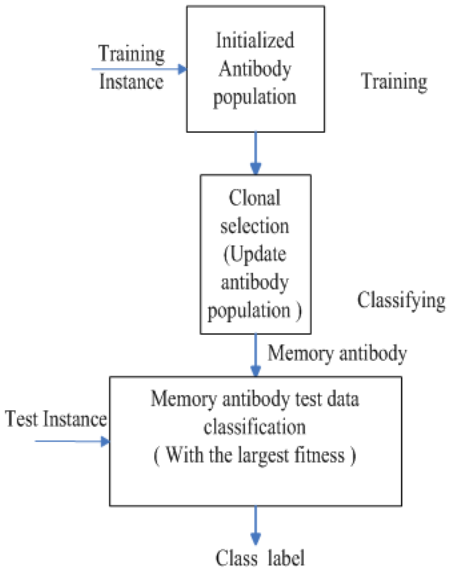

Figure 1. Block diagram of the clonal selection algorithm.

- Classification

The mature antibody sets which are regarded as a series of instance sets can be used to classify the test instance. For a instance being classified, the class mark of which is set to the one of the antigens with the best affinity. Relative to other intelligent algorithm, the encoding mechanism and the structure of the evaluation function is the same with those of clonal selection algorithm. However, the search strategy and the use of immune theory are different. Because clonal selection algorithm is using the immune memory mechanism, we can save each local optimal solution, which can quickly obtain the global optimal solution [6].

C. Attribute Weighting

The imbalance of attributes has great influence on the performance of algorithm [7-8]. So, if it is assigned higher weights to the attributes that are more important in determining one class and if these weights are used in calculation of distance, it can be prevented to make a misclassification of the two distant data according to the Euclidean norm in the same class [12]. Starting from this point, the proposed attribute weighting depends on the following base: if one attribute doesn't changing very much among the data of one class, this attribute is one of the characteristic attributes of related class and it must have a higher weight than others.

D. Weighted Clonal Selection Model

More attention in the existing research on clonal selection algorithm is its parameters control. This paper does the research about the optimization of the clonal selection algorithm from the other hand that is critical for the existing clonal selection algorithm to adapt to the role of the function. Currently, the Euclidean distance is selected as the affinity function. All the properties of the Euclidean distance for classification have the same influence, and its weight is set to 1. However, in practical applications the influence of each attribute is different. Therefore, each attribute are given to the correct attribute weight may correct this imbalance. In this paper, the rough set theory is used to weight the attribute then to classify the properties of the clonal selection.

The Euclidean distance is used to calculate the affinity in clonal selection algorithm:

$$Affinity = \sqrt{\sum_{k=1}^{L}(Ab_{j,k} - Ag_{i,k})^2} \quad (2)$$

Where, Ag_i and Ab_j are two vectors in the space, L is the length of the vector. According to the above formula, all the properties for the distance have the same influence. However, in the practical application the classification contribution of attributes is different. In the weighted clonal selection algorithm, we first calculate the importance of the attributes for the classification based on the rough set theory, and then use these weights for the affinity function in clonal selection algorithm. So, the affinity for weighted clonal selection algorithm will become the following form:

$$Affinity' = \sqrt{\sum_{k=1}^{L} w_k (Ab_{j,k} - Ag_{i,k})^2} \quad (3)$$

Where, $Ab_{j,k}$ and $Ag_{i,k}$ are separately respected the k th attribute of Ab_j and Ag_i, w_k is the weight of the k th attribute. The key problem of weighted clonal selection algorithm is how to determine the value of the weight. The system can be shown basically as in Fig. 2.

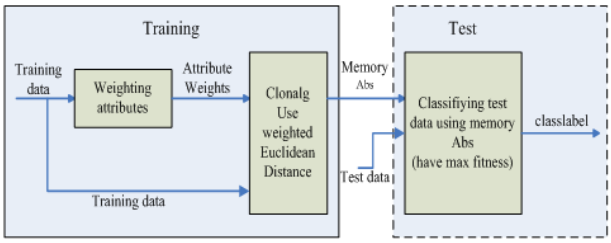

Figure 2. Weighted Clonal Selection Classificaton System

III. Weighted Clonal Selection based on Rough set throry

A. Rough Set Theory

S denotes information table, X is the non-empty subset of U, $B \subseteq A$ but $B \neq \varnothing$,For X, B-lower approximation and B-upper approximation are defined as follows:

$$\underline{B}(X) = \{x \in U : I_B(x) \subseteq X\} \quad (4)$$

$$\overline{B}(X) = \{x \in U : I_B(x) \cap X \neq \varnothing\} \quad (5)$$

$\underline{B}(X)$ in fact is the set which is composed by the objects belong to X. $\overline{B}(X)$ is the largest set made up by the objects belong to X. The set $X \subseteq U$ that has a collection of lower approximation and upper approximation of B is called a rough set [9].

B. Dependence of the Attribute

An important issue of the data analysis is the dependence between attributes. If all the attributes of set D can be decided by those of the set C, then D is dependent on C, denoted as $C \to D$. In other words, the degree of sophistication in the attributes from C is not lower than that in D. So, you can use a set C to create a set U / I_D that belongs to D. Strictly speaking, only if $I_C \subseteq I_D$, D is totally dependent on C. The dependence of the attribute is generalized for entire dependence. The part dependence shows that only some of the attributes of D is decided by that in C.

$$\gamma(C,D) = \frac{\sum_{X \in U/D} |\underline{C}(X)|}{|U|} \quad (6)$$

$\gamma(C,D)$ that is called quality of classification represents the dependence between C and D, $C, D \subset A$. It shows that the proportion of the object which can be classified based on C in the entire system. $\underline{C} | X |$ in fact is the largest set composed of the objects which belong to X.

Suppose $k = \gamma(C,D)$,then D depends on C with a degree $k (0 \le k \le 1)$, denoted by $C \to D$.

If $k = 1$, then D totally depends on C, if $k < 1$,then D partly (with a degree k) depends on C.

C. Importance of the Attribute

Attribute importance of the Decision table $S = (U, C \cup D)$ can be tested by the classification ability for S when removing an attribute $a \in C$ from C.

As mentioned above, $\gamma(C,D)$ shows the dependence between C and D, therefore we can use the change of $\gamma(C,D)$ and $\gamma(C-\{a\},D)$ to evaluate the importance of an attribute, which can be defined as:

$$\sigma_{(C,D)}(a)=\frac{\gamma(C,D)-\gamma(C-\{a\},D)}{\gamma(C,D)} \quad (7)$$

The larger the value of $\sigma_{(C,D)}(a)$ is, the more important of the attribute a for the decision D is under the condition that a is known in advance. In clonal selection algorithm, the importance of each attribute is equal, which is set to 1. But during the weighted clonal selection, the importance of attributes should be re-distributed based on the mathematical expectation. The weight can be defined as under the algebraic concept [7].

$$w_i=\frac{\sigma_{(C,D)}(a_i)}{\frac{1}{m}\sum_{i=1}^{m}\sigma_{(C,D)}(a_i)} \quad (8)$$

Where, D is the decision attribute set, and has m condition attribute, w_i is the importance of attribute a_i. And the algebraic definition of attribute importance is based on the impact of that attribute to determine the classification subset.

TABLE 1

THE PSEUDO-CODE TO CALCULATE THE ATTRIBUTE WEIGHTS BASED ON ROUGH SET THEORY

Algorithm : Fixed Weight, Calculate the attribute weights

Input : Training Data Set;

Output : Weight array W_L, L denotes the number of the attributes

1) For the training data set, calculate the number of $Ag\_class_j(j=1,2,.....n, n: number\ of\ class)$;

2) Calculate the dependence $\gamma(C,D)$ of D relative to C;

   a). Get $\underline{C}(X)$ based on Ag_class_L, Ag_class_L contains all the training data belonging to the class, where L denotes the attribute number;

   b). Compute $\gamma(C,D)$

$$\gamma(C,D)=\frac{\sum_{X\in U/D}|\underline{C}(X)|}{|U|}$$

3) for each attribute a

   a). Calculate the importance of a, $\sigma_{(C,D)}(a)$:

$$\sigma_{(C,D)}(a)=\frac{\gamma(C,D)-\gamma(C-\{a\},D)}{\gamma(C,D)}$$

   Where $C-\{a\}$ denotes the set C removing the attribute a, $\gamma(C-\{a\},D)$ represents the dependence of decision attribute D relative to $C-\{a\}$;

   b). Calculate the weight:

$$w_i=\frac{\sigma_{(C,D)}(a_i)}{\frac{1}{m}\sum_{i=1}^{m}\sigma_{(C,D)}(a_i)}$$

D. An Example for the Attribute Weighted Method based on Rough Set

As shown in Table 2, (a, b, c) is the condition attributes, while d is the decision attribute.

TABLE 2

DECISION TABLE

U	a	b	c	d
1	3	2	1	2
2	2	1	1	1
3	2	1	1	2
4	1	1	1	1
5	2	2	2	1
6	3	1	2	2

According to the rough set theory, we can get the weight of each attribute. For example, if we want to calculate the weight of attribute a, first $\underline{C}|X|=4$, then we can get the value of $\gamma(C,D)$, which is $2/3$. We can get C-{a}(X)=3. $\gamma(C-\{a\},D)$ can be computed by the function (6), the value of which is $1/2$. At last, $\sigma_{(C,D)}(a)=1/4$ under the function (7), when using the function (8), the weight of a, w_a is 3.

IV. EXPERIMENTAL METHODS AND RESULTS

A. Experimental Data

In order to verify the effect of the weighted clonal selection algorithm based on rough set, we choose the 8 data set in UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets)to test [10]. The selected standard is based on the fact that the data set with small attribute number is suit for the weighted clonal selection. Therefore, we set the attribute number less than 10, the details of the data set is described as follows:

TABLE 3

DETAILED INFORMATION FOR THE EXPERIMENTAL DATA

Dataset	Instances	Attributes	Class	Missing	Numer
breast-cancer	286	10	2	N	N
car-evaluation	1727	7	4	N	N
contact-lenses	24	5	3	N	N
haberman	306	4	2	N	N
hayes-roth	132	5	4	N	N
iris	150	5	3	N	N
postoperative-p	90	9	3	N	N
shuttle-land-c	15	7	2	N	N

B. Analysis of Experimental Results

Clonal selection and weighted clonal selection algorithm are tested in this section on Weka [11]. In order to obtain independently reliable results, each of the original data set should be divided into the ratio of 9 to 1 randomly respectively used for training and testing data [12]. The weight and the clonal selection model are trained on the training set, and then use it to classify the test data set. The two algorithms run on the same training data set and the same testing data set. In view of the randomness of clonal selection algorithm, we do the experiment 10 times, and get the average classification results. The classification results are shown in Table 4.

TABLE 4

THE COMPASSION VIA CLASSIFICATION ACCURACY AND STANDARD DEVIATION(%)

Dataset	Clonal	Weighted-Clonal
breast-cancer	*66.13±9.83*	***66.67±6.55***
car-evaluation	*68.63±3.37*	***70.90±2.97***
contact-lenses	*65.67±30.87*	***76.17±27.75***
haberman	71.96±6.27	71.70±8.22
hayes-roth	*56.37±13.83*	***67.23±12.81***
iris	*85.33±9.89*	***92.00±8.1***
postoperative-p	*56.33±15.97*	***57.33±15.69***
shuttle-land-c	95.00±15.08	95.00±15.08
Average	70.68	74.63

From the specific information shown in Table 4, the experiment results of which show that the weighted clonal selection has significantly improved (win 6 data sets) compared to the traditional clonal selection based on the classification accuracy: *breast-cancer* (***66.67±6.55***, *66.13±9.83*), *car-evaluation* (***70.90±2.97***, *68.63±3.37*), *contact-lenses* (***76.17±27.75***, *65.67±30.87*), *hayes-roth*(***67.23±12.81***, *56.37±13.83*), *iris*(***92.00±8.1***, *85.33±9.89*), *postoperative-p* (***57.33±15.69***, *56.33±15.97*). And the average accuracy is weighted clonal selection is ***74.63%***, which is higher than clonal selection *70.68%*.

Figure 3. Weighted Clonal Selection Classificaton System

The same result can be seen in Fig. 3, compared with clonal selection, the weighed methods win 6 data sets, lose 1 data sets

V. Conclusions and Future Work

In this paper, we propose a new clonal selection algorithm based on rough set theory. This new method takes full account of the contribution of each attribute in the classification process, so it has better performance compared to the traditional method. Firstly, it learns the best weight using rough set method, and then classifies the test instance using the weighted clonal selection based on the best weight we have learned. We experimentally tested our new algorithm using the 8 UCI data sets selected by Weka, and compared it to standard clonal selection. The experimental results show that the accuracy of weighted clonal selection is much higher than that of standard clonal selection.

As future work, we plan to investigate some issues. It is aimed in further studies which will be conducted in parallel with this study that better results will be obtained using rough set weighting schemes for other weighted classification systems.

References

[1] F. Qin, S. L. Ren, Z. K. Cheng and H. Luo, "Attribute weighted Naïve Bayes classification," *Computer Engineering and Applications*, vol. 44, no. 6, pp. 107–109, 2008.

[2] R. Y. Feng, T. H. Shang and H. C. Liu, "An attribute weighting K-means alforithm based on mean-square-deviation," *Information Technology*, vol. 3, pp. 55–57, 2010.

[3] L.N. De Castro and J. Timmis, *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer-Verlag, 2002.

[4] J. Brownlee, *Clonal Selection Theory &Clonalg the Clonal Selection Classification Algorithm*. Swinburne University of Technology, 2005.

[5] G.L. Yuan and J.Z. Liu, "The Design for Feed Water System of Boiler Based on Fuzzy Immune Smith Control," *Journal of Computers*, vol. 7, no. 1, pp. 278-283, 2012.

[6] L. N. De Castro and F. J.Von Zuben, "Learning and Optimization Using the Clonal Selection Principle," *IEEE Transactions on Evolutionary Computation, Special Issue on Artificial Immune Systems*, vol. 6, pp. 239-251, 2002.

[7] J.Wu, Z. Cai and S. Ao, "Hybrid Dynamic K-Nearest-Neighbor and Distance and Attribute Weighted Method for Classification," *International Journal of Computer Applications in Technology*, vol. 43, no. 4, pp. 378-384, 2012.

[8] J.Wu and Z. Cai, "Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes (WNB)," *Journal of Computational Information Systems*, vol. 7, no. 5, pp.1672-1679, 2011.

[9] D.H. Wang, X.W. Liu, L.X. Jiang, X.T. Zhang and Y. G. Zhao, "Rough Set Approach to Multivariate Decision Trees Inducing," *Journal of Computers*, vol. 7, no. 4, pp.870-879, 2012.

[10] C. Merz, P. Murphy and D. Aha, UCI repository of machine learning databases. Department of ICS, University of California, Irvine,1997.
http://www.ics.uci.edu/mlearn/MLRepository.html.

[11] H. W. Ian and F. Eibe, "Data mining: practical machine learning tools and techniques, 2nd edn," Morgan Kaufmann, San Francisco, 1996.
http://prdownloads.sourceforge.net/weka/datasets-UCI.jar

[12] L.Jiang and C. Li, "Scaling Up the Accuracy of Decision-Tree Classifiers: A Naive-Bayes Combination," *Journal of Computers*, vol. 6, pp. 1325-1331, 2011.

Jia Wu is currently studying in CUG for his PhD degree in Geology Information Technology. His current research interests include data mining and knowledge discovery, machine learning, evolutionary computation, and hyperspectral remote sensing.

Zhihua Cai received the PhD degree from China University of Geosciences, in 2003. Dr. Cai is currently a faculty member at School of Computer Science, China University of Geosciences, Wuhan, China. His main research areas include data mining, machine learning, evolutionary computation, and their applications.

Xiaolin Chen received the B. Eng degree in computer science from China University of Geosciences (CUG), Wuhan, China, in 2010. She is currently studying in CUG for her Msc degree in Computer Science. Her current research interests include data mining and machine learning, evolutionary computation.

Meng Li is currently studying in CUG for his B. Eng degree in computer science from China University of Geosciences (CUG), Wuhan, China. Her current research interests include data mining and machine learning.

Bin Guo received the B. Eng degree in computer science from Zhongnan University of Economics and Law (ZUEL), in 2009. He is currently studying in USYD for his Msc degree in Computer Science. His current research interests include data mining and knowledge discovery.