2010 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010)
IPCSIT vol. 53 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V53.No.1.58

Associative Classification Based on Artificial Immune System

Zhang Lei[+] and Jiang Ji Min

School of Electronic Information Engineering
Henan University of Science and Technology
Luoyang, China

Abstract. Associative classification algorithms which are based on association rules have performed well compared with other classification approaches. However a fundamental limitation with these classification algorithms is that the search space of candidate rules is very large and the processes of rule discovery and rule selection are conducted separately. This paper proposes an approach called ARMBIS, which is based on the natural immune principle, for searching associative rules. The proposed algorithm has the capability of dealing with complex search space of association rules while still ensuring that the resultant set of association rules is appropriate for associative classification. The performance evaluation results have shown that the proposed algorithm has achieved good runtime and accuracy performance in comparison with conventional associative classification algorithms.

Keywords: Associative classification; association rules; artificial immune system

1. Introduction

As one of the most fundamental data mining tasks, classification has been extensively studied and various types of classification methods such as decision tree [1], neural network [2] and rough set theory [3] have been proposed to solve the classification problems in various fields. Since the introduction of association rule mining [4], many association-based (or related) classifiers have been proposed [5,6]. The Associative classification (AC) approach was first introduced in [6] with the classification based on associations (CBA) algorithm. Associative classification takes advantage of association rule mining in the rule discovery process in extracting high quality rules that can accurately generalize the training dataset. This approach has achieved high accuracy in comparison with other classification approaches. However, one of the main problems of using association rule mining in AC approaches is the exhaustive search in a very large search space of possible rules. Therefore, the rule search process is computationally expensive [7], especially with small support threshold values which are very important for building accurate associative classifiers from large training datasets. In addition, the two different processes on rule discovery and classification are poorly integrated in conventional AC approaches as these processes are conducted separately [8].

Artificial immune system (AIS) is inspired by the natural immune system. The powerful information processing capabilities of the natural immune system such as feature extraction, pattern recognition, learning, memory, and its distributive nature provide rich metaphors for its artificial counterpart. Specifically, three immunological principles are primarily used in AIS methods [9]. These include the immune network theory, negative selection mechanism, and clonal selection principle, etc. In this paper, we mainly apply the immune network theory and clonal selection principle. It follows the population-based search model of evolutionary algorithms that have the capability of dealing with a complex search space.

[+] Corresponding author.
E-mail address: leizhang87@163.com.

In this paper, we propose an AIS-based optimization algorithm for searching and optimizing association rules for AC. The optimization process for searching a set of high confidence association rules is mainly inspired by the clonal selection principle. In addition the diversity of the individuals in the population is also considered. Instead of searching for association rules using conventional association rule mining algorithms and then selecting the most suitable subset for the classification process, the proposed algorithm searches for the most suitable subset of association rules for the classification process directly in an evolutionary manner. We can avoid searching association rules exhaustively while still ensuring that the resultant set of association rules is appropriate for AC. Therefore, the computational complexity of rule search space can be reduced greatly. The performance evaluation results have shown that the proposed algorithm has achieved good runtime and accuracy performance in comparison with conventional AC algorithms.

The rest of the paper is organized as follows. We first review the related work on the AIS and introduce the related algorithms for AC. Next, we introduce some basic definitions and notations in section 3. The proposed AIS-based algorithm for AC is then presented in section 4. Substantial experimental results are shown in section 5. Finally, we conclude this paper in section 6.

2. Related Work

The natural immune system is seen as a complex of cells, molecules, and organs that protect organisms against infections [10]. One of the fundamental roles of an immune system is to recognize and eliminate disease-causing agents known as pathogens.

Artificial immune systems (AIS) are a new computational intelligence approach inspired by theoretical immunology, observed immune functions, principles and mechanisms [11]. Three immunological principles including the immune network theory, negative selection mechanism and clonal selection principle are most frequently adopted in AIS. AIS appear to offer powerful and robust information processing capabilities for solving complex problems. AIS have been applied to a wide variety of domain areas, such as pattern recognition and classification [12], optimization [13], computer security [14] and fault detection etc [15].

Since the introduction of association rule mining, many association-based (or related) classifiers have been proposed [5-6]. The Associative classification approach was first introduced in [6] with the classification based on associations (CBA) algorithm. This classifier builder uses a brute-force, exhaustive global search. CMAR is another typical association based algorithm [16]. These approaches adopt efficient association rule mining algorithms, such as Apriori and FP-growth [17] etc., to first mine a large number of high-confidence rules satisfying a user-specified minimum support and confidence thresholds and then use various sequential-covering-based schemes to select from them a set of high-quality rules to be used for classification. Generally, these kind of association-based methods are more accurate than traditional decision tree building algorithms because they are based on global knowledge. However, the drawback of these approaches is that the number of initial rules is usually extremely large, significantly increasing the rule discovery and selection time.

3. Notations and Definitions

3.1 Association classification rule

In this research, each dataset has already been separated into training and test sets. A training dataset TrDB is a set of training instances, where each training instance, denoted as a triple (Tn, X, C_n), where and Tn denote the unique training instance identifier, X contains a set of items ($X \in I$, I presents the complete set of items including in the whole dataset); C_n denote a class identifier which presents a possible class label in dataset. ($C_n \in \{C_1, C_2, \cdots C_k\}$). A training dataset TeDB is in the same form as the training dataset and is used to evaluate the performance of a classifier.

An association classification rule has the following form:

$$R_i : Y \rightarrow C_j \ : sup_Y^{C_j}, conf_Y^{C_j} \tag{1}$$

Where Y is itemset called the body or the condition of the rule R_i and C_j is the head or the consequence of the rule R_i which contains a single class label.

Two measures, the degree of support and the degree of confidence, are used to define an associative classification rule. $sup_Y^{C_j}$ and $con_Y^{C_j}$ are the degree of support and the degree of confidence of rule R_i respectively.

Let sup_Y represent the number of transactions in the dataset that contain the itemset Y. Also, let $sup_Y^{C_j}$ denote the number of transactions in the dataset containing both the itemset Y and associated with the class label C_j. The confidence of R_i $con_Y^{C_j}$ is defined as the following form:

$$con_Y^{C_j} = \frac{sup_Y^{C_j}}{sup_Y} \tag{2}$$

We say a transaction contains itemset Y, if the entire items in the itemset Y are included in this transaction.

3.2 Associative Classification

Generally, an associative classification approach consists of three major processes, namely rule discovery, rule selection and classification.

• *Rule discovery.* It mines the complete set of association rules from a training dataset. These association rules are called class association rules.

• *Rule selection.* It evaluates the association rules discovered by the rule discovery process based on the classification of the training dataset and selecting the subset, which gives the best accuracy, to form a classifier. Generally, AC approaches are based on the confidence measure to select rules. The idea is that rules with higher confidence would probably give higher prediction accuracy.

• *Classification.* It classifies data samples in a test dataset based on the set of association rules resulting from the rule selection process.

4. Association Rule Mining Based Immune System

To achieve AC efficiently, it needs not only to mine the highest association rules, but also needs to integrate these association rules into a whole effective classifier. It is not necessary and practical to mine all the possible highest confidence association rules and form a classification rule set. We need only a subset of them that can be used to form an accurate classifier. In the proposed algorithm a set of rules with highest confidence are searched in an evolutionary approach instead of using conventional association rule mining algorithms for rule discovery and selection. We can avoid searching association rules exhaustively while still ensuring that the resultant set of association rules is appropriate for AC. Therefore, the computational complexity of rule search space can be reduced greatly.

In the proposed algorithm, the problem of mining a set of association rules for AC is considered as an optimization process. The optimization process searches for association rules in an evolutionary manner which is based on the clonal selection principle and immune memory of the nature immune system. In this proposed algorithm, no distinction is made between a B cell and its receptor, known as an antibody, so that every element of our artificial immune system will be generically called an antibody. Each association rule is considered an antibody. The diversity of the individuals in the current population is also considered. This is reflected in the cloning process and memory cells selection process. The proposed algorithm can reach a diverse number of local optimal solutions, which are the candidate rules to form an effective association classifier. The resultant association rules will be extracted from the memory cell population and form the resultant association classification rule set.

4.1 ARMBIS algorithm

The proposed algorithm, which is given in Fig. 1, is called ARMBIS (Association Rule Mining Based on Immune System). Firstly, the minimum support threshold is used to filter out specific rules from the population. Next, the confidence values of the rules are used for affinity computation. The population is

cloned, mutated and diversified. Finally, the best rules in the population are moved to the memory based on the confidence constraint. At the end of each generation, the coverage measure is calculated based on the memory rules to decide whether the process should continue with another generation or be terminated. The process will be terminated when the coverage constraint is satisfied or the number of generations reaches a predefined maximum number of generations.

For the given training dataset TrDB, our approach first computes the frequent items by scanning TrDB once, and sorts them to get a list of frequent items (denoted by *f_list*).

1) Selection: firstly the support and the confidence values of the rules in the population P are computed. The selection process eliminates rules with support values below the support threshold. In addition, rules of low confidence values in will be eliminated. In the proposed algorithm, the confidence measure is considered as affinity.

Definition 1 (affinity/fitness). Let us denote the corresponding association rule for an antibody Ab_i in the population P is R_i. The affinity/fitness of antibody Ab_i is defined as the degree of confidence of association rule R_i.

2) Cloning: The cloning process is carried out such that the clone rate of a rule is directly proportional to its affinity and inversely proportional to the density of the association rule. The density of an association rule in the current population reflects whether the number of rules similar to this rule is many enough. The density of a association rule Ri density(Ri) is defined as follows.

Definition 2 (cover). The instance in the dataset which can be classified correctly by an association rule R_i. We call the instance is covered by the association rule R_i. The number of instances which is covered by the association rule Ri is denoted by Covern(R_i).

Definition 3 (*similarity degree)*. The *simdegree* between the association rule R_i and R_j in the population *P* is defined as follows.

$$simdegree(R_i, R_j) = \frac{SCovern}{Covern(R_i)} \times \frac{SCovern}{Covern(R_j)} \tag{3}$$

Where *SCovern* denotes the number of instances which is both covered simultaneously by the association rule R_i and R_j in the population P. When the *simdegree* between the rule R_i and another rule R_j is above the specified threshold we call the association rules R_i and R_j are similar to each other.

Definition 4 (*density*). Let us denote $simrn(R_i)$ as the number of association rules in the current population which is similar to R_i, the density of the association rules R_i is defined as follows.

$$density(R_i) = \begin{cases} 1 & \text{if } simrn(R_i) \geq th_d \\ 0 & \text{others} \end{cases} \tag{4}$$

When the density of an association rules R_i is 1, this means there have been many similar cells with this rule in the current population and we should control the number of clones in order to keep the diversity of rules in the current population. The density of rules was used to keep the diversity of the population P.

The cells with higher confidence value implies closer to the optimal association classification rule, thus produce more clone. The cells with higher density value implies that there have been many similar cells in the population, thus its number of clones should be smaller in order to reduce the possibility of sinking into local optima.

Let us denote the number of clones produced for a rule R_i as *CloneNum(R_i)*. When the density of an association rules R_i is 1, then its number of clones is directly set to the minimum otherwise the number of clones is proportional to its affinity. The number of clones for the rule R_i *CloneNum(R_i)* is computed as shown in (5).

$$CloneNum(R_i) = NumC_{\min} + (NumC_{\max} - NumC_{\min}) \times \frac{Affinity(R_i)}{Af_{\max}} \tag{5}$$

Where $NumC_{\max}$ and $NumC_{\min}$ are the maximum and minimum number of clones for an antibody respectively and $Af_{\max}$ is the maximum fitness of the current antibody population and $Affinty(R_i)$ is the affinity of the association rule R_i.

3) Maturation: In the clonal selection algorithm, the mutation rate of a cell is inversely proportional to the affinity of the cell. It gives the chance for each low affinity cell to "mutate" more in order to improve its affinity. In ARMBIS, the mutation rate is equal to "one item" for every rule. That is, when a rule is mutated, the newly produced rules will differ from the parent rule only by one item, either by adding one item or deleting one item.

When adding one new item, the item is randomly selected from the list of frequent items (*f_list*), and the C_i is determined according to the rule which covers more instances. When deleting one item, the item is in the same way randomly selected from the body of the rule, the class label keeps the same.

Algorithm: ARMBIS（Association Rule Mining Based on Immune System） for Associative Classification

Input: C_i-a class in the dataset; I-the set of items in the dataset
minsup, minconf -minimum thresholds for the support and confidence,
itern-parameters for control the number of generation of iteration
Output: M-memory set containing a set of association rules
Method:

Initialize P= $\{\{item\} \Rightarrow C_i \mid item \in I\}$； Initialize M= $\varnothing$， iternum=0

do // begin a generation
  for all rule R_i in P **do** // delete those rules below the determined threshold
    count sup(R_i) and conf(R_i)
    if sup(R_i)<minsup **then** remove R_i from P **endif**
  endfor
  clone P with teding clonrate→P′ // clone operation
  mutate P′ →P " // mutation operation
  prune P " // pruning some inefficient rules
  select some number of high confidence rules into the M: conf(R_i)>minconf //memory population selection
  P←P "； iternum=iternum+1 // set the parameters for next generation
While covernum(M)<mincovernum and iternum<itern
Return M // high confidence rules produced for AC

Figure 1. The proposed ACBAIS algorithm

*4) Diversity Introduction:*In the proposed algorithm two ways are used to maintain the diversity of the evolution population. One way is that the clone selection process updates the population by replacing some existing cells with new ones. The other way is to adopt the new different mutation and clone strategy.

5) Memory Selection: The memory selection process adds high confidence rules into the memory pool. In our approach, when new rules produced have higher confidence, these rules can replace those rules which have lower confidence and the similarity degree between them surpass the threshold.

*6) Termination test :*The stopping condition in this algorithm is the redefined maximum number of iteration and the coverage constraints. When the algorithm terminates, the resultant association rules will be extracted from the memory cell population and form the resultant association classification rule set.

Definition 5 (coverage). The definition coverage reflects the number of instance in the dataset which can be classified correctly by an association rule set denoted as *RS*.

The coverage constraints are that the number of instances covered by current CARs in the memory population surpasses the predetermined threshold.

5. Empirical Results

We evaluated the proposed algorithm on the UCI databases in comparison with FOIL [19] and HARMONY [20], which are two well-known algorithms for classifying categorical data. Foil is a rule induction-based algorithm using the foil gain to determine how each rule would be extended. HARMONY follows an instance-centric framework and mines the covering rules with the highest confidence for each instance. All the experiments were performed on a 2.0GHz machine with 1GB memory. In the implementation of the ARMBIS algorithm, the coverage threshold was set to 95%.

The datasets used in the experiment are obtained from the UCI Machine Learning Repository [18]. Table 1 lists the properties of the five attribute datasets. Here, each item corresponds to a value of a nominal attribute, or an interval of value of a continuous attribute.

Table 2 shows the performance comparison results in terms of the 10-fold cross validation accuracy for the five UCI datasets.

All the results are the best accuracy achieved in each dataset for the three algorithms. From these results we can see that ARMBIS has better accuracy than both FOIL and HARMONY for most of the five UCI datasets. And only in the penDigits dataset the FOIL and HARMONY algorithms performed • better than our algorithm. In the experiment the total minimal support threshold was fixed at 1%.

Table 1. Attribute datasets

Data set	**Attribute**		
	# instances	***# items***	***# classes***
Adult	48842	131	2
Chess	28056	66	18
letter	20000	106	26
nursery	12960	32	5
penDigits	10992	90	10

Table 2. Comparison results of classification accuracy for different algorithms

Data set	**Attribute**		
	FOIL	***HARMONY***	***ARMBIS***
Adult	82.5%	81.9%	85.9%
Chess	42.6%	44.8%	58.2%
letter	57.5%	76.8%	77.6%
nursery	91.3%	92.8%	96.9%
penDigits	88.0%	96.2%	87.8%

In addition, Table 3 shows the performance comparison results of runtime on the three algorithms using the five datasets. Similarly all the results are the shortest runtime (in seconds) achieved in each dataset for the three algorithms. We can see that the runtime of ARMBIS have a distinct advantage over FOIL and are comparable to HARMONY. In the nursery and chess dataset the runtime of HARMONY are faster than the proposed algorithm. Because the results of the FOIL and HARMONY algorithms were achieved by implementing in JAVA and testing on different machines, these times only provide a relative computational requirement of the different schemes. The reason is that ARMBIS is a stochastic optimization algorithm and if the stopping condition is not satisfied the algorithm will continue the iteration until the maximum number of iteration. However the proposed algorithm usually can achieve the better classification accuracy in such datasets.

Table 3. Comparison results of runtime (in seconds) for different algorithms

Data set	**Attribute**		
	FOIL	***HARMONY***	***ARMBIS***
Adult	10251.0	1395.5	972.6
Chess	10122.8	11.34	698.6
letter	4365.6	778.9	512.2
nursery	73.1	6.2	298.7
penDigits	821.1	82.6	152.9

6. Conclusion

In this paper, we have proposed an approach called ARMBIS, which is based on the natural immune principle, for searching association rules for AC. Following the population-based search model of evolutionary algorithms, the proposed algorithm has the capability of dealing with complex search space of association rules. Moreover, we can avoid searching association rules exhaustively while still ensuring that the resultant set of association rules is appropriate for AC. The computational complexity of rule search space can be reduced greatly. It enables the integration between rule discovery and rule selection processes which makes this approach adjustable to the data in comparison with the conventional AC approaches. The performance

evaluation results have shown that the proposed algorithm has achieved good runtime and accuracy performance in comparison with conventional AC algorithms.

7. References

[1] J. R. Quinlan, "C45: Programs for machine learning," San Mateo: Morgan Kaufmann, 1993.

[2] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," Vol. 1, D. E. Rumelhart and J. L. Mcclelland , Eds. MA: MIT Press, 1986.

[3] Z. Pawlak, "Rough sets: Theoretical aspects of reasoning about data. Boston," MA: Kluwer Academic, 1991.

[4] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," SIGMOD'93, ACM Press, 1993, pp.207-216, doi:10.1145/SCIS.170035.170072.

[5] K. Ali, S Manganaris and R. Srikant, "Partial Calssification Using Association Rules," KDD'97, Press, 1997, pp. 69-87.

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

[7] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," Proc. ACM SIGMOD Intl. Conf. Manage. Data, Press, 2000, pp. 1–12.

[8] D. R. Carvalho, A. A. Freitas, and N. F. Ebecken, "A critical review of rule surprisingness measures," Proc. 4th Int. Conf. Data Mining, Press, 2003, pp. 545–556.

[9] D. Dasgupta and J. Zhou, "Reviewing the development of AIS in last five years," Proc. 2003 IEEE Congr. Evol. Comput., IEEE Press, 2003, pp .

[10] D. Dasgupta, "Artificial Immune Systems and Their Applications," Berlin: Springer-Verlag, 1999.

[11] L.N. de Castro, J.I. Timmis, "Artificial Immune Systems: A New Computational Intelligence Approach," London: Springer-Verlag, 2002.

[12] J. Timmis and M. Neal, "A resource limited artificial immune system for data analysis," Knowl. Based Syst., vol. 14, Jun. 2001, pp. 121–130, doi:/science.1065555.

[13] L. N. d. Castro and F. J. V. Zuben, "Learning and optimization using the clonal selection principle," IEEE Trans. Evol. Comput., vol. 6, Jun. 2002, pp. 239–251, doi:/science.1065555.

[14] S. Forrest, "Self–nonself discrimination in a computer," Proc. IEEE Symp. Res. Security Privacy, IEEE Press, 1994, pp. 202–212.

[15] D. W. Bradley and A. M. Tyrrell, "Immunotronics: Hardware fault tolerance inspired by the immune system," Proc. 3rd Int. Conf. Evolvable Syst. (ICES 2000), Press, 2000, pp. 11–20.

[16] W. Li, J. Han and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules". Proceedings of. IEEE-ICDM, IEEE Press, 2001, pp. 369−376.

[17] J. Han, J. Pei, Y. Yin. Mining Frequent Patterns without Candidate Generation, SIGMOD'00.

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

[19] J. Quinlan, R. Cameron-Jones, FOIL: A Midterm Report, ECML, 2004.

[20] J. Y. Wang and G. Karypis, "HARMONY: Efficiently Mining the Best Rules for Classification," SIAM International Conference on Data Mining, Press, 2005, pp. 205--216.