# Classification of Breast Cancer Samples Through Using the Artificial Bee Colony Algorithm

Mahsa Nazarian
Department of Computer Science and Engineering
Khouzestan Science and Research Branch,
Islamic Azad University
Ahwaz, Iran

Mashala Abbasi Dezfouli
Department of Computer Science and Engineering
Khouzestan Science and Research Branch,
Islamic Azad University
Ahwaz, Iran

Ali Haronabadi
Department of Computer
Islamic Azad University,
Central Tehran Branch
Tehran, Iran

**Abstract**: The algorithm of artificial bee colony has been widely applied in optimization issues. The pattern of this algorithm has been derived from the intelligent behavior of bees in the nature. The aim of this paper is to present a model for classification of binary problems through using data mining techniques, and on the basis of artificial bee colony algorithm. In this paper, breast cancer data that has been presented by Wisconsin University to investigate and evaluate the proposed method has been used and applied. With regard to the obtained results, it has been shown that this algorithm has a considerable and important function.

**Keywords**: breast cancer, data mining, ABC, artificial bee colony, classification

## 1. INTRODUCTION

By using data mining process, knowledge can be obtained via input data set through various steps and procedures. Using data mining techniques widely in todays different sciences and problems have made it possible for researchers to explore the new concepts and patterns. In the past, exploration of these problems and issues was impossible. Double problems in classification refer to those problems and issues in which data is classifies into two groups and classes. Data mining includes two steps, namely pre-processing and pattern diagnosis. In pre-processing step, the considered characteristics and specifications are obtained via data, and the result is data without any errors. Pattern diagnosis step involves various algorithms for classification of data obtained from the pre-processing step. The algorithm of artificial bee colony is an algorithm based on the crowd. This algorithm has been increasingly considered because understanding and applying this algorithm is easy. This algorithm has been used in different fields such as optimization. In this paper, a model based on the artificial bee colony algorithm has been presented for exploring knowledge via data base related to breast cancer. In fact, the algorithm of artificial bee colony has been used and applied in the field of data classification. In this paper, in pre-processing step, the related data is firstly normalized. In the second step, through using the model based on the artificial bee colony algorithm, data is classified into two pre-defined classes.

## 2. ARTIFICIAL BEE COLONY ALGORITHM

Artificial Bee Colony Algorithm has been presented in 2005. Since understanding and applying this algorithm is easy, it has been widely used in various fields of optimization. The results of using this algorithm has been compared with other widely used methods such as genetic algorithm [4], DE (differential Evolution) [5], and PSO (Particle Swarm Optimization) [6]. This algorithm involves three parts: food source, worker bee and non-worker bee. There are two kinds of non-worker bees, namely pioneer bees and supervisor bees. The worker bees refer to those bees finding the source of food, so the number of theses bees equals to the number of food sources. In addition, the number of food sources equals to the number of problem solutions. The pioneer and supervisor bees try to find new food sources. The basis of finding food sources is random exploration. The pioneer bees find the food sources randomly, while the supervisor bees evaluate and investigate the quality. This algorithm can be defined on the basis of the following equations:

$$V_{ij} = X_{ij} + \phi_{ij}(X_{ij} - X_{kj}) \qquad (1)$$

$$P_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \qquad (2)$$

In equation (1), $P_i$ is the possibility of selecting ith food source. $fit_i$ refers to the value of ith food source, and SN is the number of food sources. In fact, SN equals to the number of possible solutions [10].

In ABC algorithm, artificial bees find new solutions locally. This exploration has been shown in equation (2). $V_{ij}$ is the location of the new food source. Both $X_{kj}$ and $X_{ij}$ are old food sources. I this equation, $j \in [1,2,..., D]$ and $k \in [1,2,..., SN]$ are considered. K and j values are not equal, and are selected randomly. D equals to the number of optimization parameters. $\phi$ is a random number [-1 and 1]. The important point is that, in selecting food source, ABC algorithm selects the sources whose quality equals to the prior source or is more than it. This means greedy selection. After choosing the new source of food, that source is added to the memory. There are two other parameters in this algorithm, namely the limitation of food source location and the repeated number of food source exploration cycles [10].

## 3. LITERATURE

Osmar and his colleagues presented several methods for diagnosis of tumor in digital mammography in which two determining techniques, namely neural network and association rules, have been used for anomaly diagnosis and classification. In both methods, the precision rate of classification is more than 70 percent. 332 mammography images were used, and they were classified into three groups, namely normal tumor, benign tumor and malignant tumor. In addition, abnormal issues are classified to six groups, and it depends on the degree and kind of its growth. Their research involved several steps such as image reception, image correction, exploring symptoms and signs, and classification. They used apriori algorithm in order to derive association rules from the signs and symptoms [11].

Dr. Rani classified the related data of medicine through using neural networks. He used the model of feed forward neural network and the algorithm of back propagation learning. He also considered the strategy of parallel neural network. In his research, 699 samples and 10 specs were used. According to the results of classification, it has been reported that success percentage of the proposed method was 96,6 [13].

Anuncias and his colleagues presented a data mining approach for detection of high-risk breast cancer groups [3].

Abdelghani Bellaachia and his colleagues tried to predict breast cancer survivability using three data mining techniques, namely Naïve Bayes, neural networks and decision-making tree. It has been reported that the success percentage of Naïve Bayes neural networks and decision-making tree were respectively 84,5 and 86,5 and 86,7 [7].

Chang and his colleagues tried to compare three data mining techniques with genetic algorithm, and they focused on the function of artificial intelligence and data mining. They presented a model for predicting breast cancer [12].

Amin Einipour and his colleagues presented a fuzzy method on the basis of ants colony for breast cancer diagnosis. The algorithm of ants colony has been derived from the behavior of ants searching the food in the shortest direction. In this research, the authors presented a method in order to classify the samples into two groups, namely cancroids and un-cancroids samples, and they used fuzzy rules and method as well as ants colony algorithm.[1]

Ping zhang and his colleagues explored Bayesian networks for breast cancer detection. These tools are used by radiologists [8].

Gupta and his colleagues investigated breast cancer diagnosis and prognosis. Also, they explored the medicine data and their classifications.[9]

## 4. USED DATA

In this paper, Wisconsin data set related to breast cancer samples has been used and applied in order to evaluate and investigate the proposed method. This set involves 699 records and 9 specs. In this data set, there are records having incomplete data. At first, incomplete records must be deleted. The number of these records is 16. When these records are deleted, 683 records remain.

## 5. THE PROPOSED METHOD

### 5.1 Pre-processing

As it was mentioned above, incomplete and non-numerical data must be deleted, and should not be considered in data set; therefore, in the first step, data records must be selected so that these data will enter the next step. Here, we consider 450 records as the training data. These records are used for exploring the rules as well as evaluating them. We consider remaining 233 records as the experimental records. After exploring and obtaining rules, all 683 records will be classified.

### 5.2 The Pattern of rules

Each classification rule involves two parts in classification, namely condition and result. The rule structure of this method is as follows:

```
Struct RuleSet{

        Double *lowb;

        Double *upb;

        String ClassName;

        Double *fitval;

}
```

lowb is the sign of minimum values of each spec. upb is the sign of maximum values of each spec. ClassName refers to the name of the class to which the rule is related, and fitval stands for the value of the rule fitness.

### 5.3 Evaluation Function

In order to evaluate the fitness of each rule, the following evaluation function has been proposed.

$$FV = \frac{TP}{TP+FN} \times \frac{TN}{TN+FP} \text{ [10]} \qquad (3)$$

In this equation, FV is the rule fitness value. TP refers to the number of records that have also been applied in rule condition, and their class is as same as the class rule. TN stands for number of records that have not been applied in rule condition, and their class is not as same as the class rule. FN is the number of records that are not used in rule condition, but their class is as same as the class rule.

### 5.4 Rule Exploration and Extraction

In this step, we explore the rules through using random numbers. Here, we create two random minimum and maximum values by using equations (4) and (5). The created random numbers neither should be too great nor too small. For instance, if the maximum value is too great, all records will be applied.

$$LB = f - \phi_1 (F_u - F_l) \text{ [2]} \qquad (4)$$

$$UB = f + \phi_2 (F_u - F_l) \text{ [2]} \qquad (5)$$

In these equations, LB is random minimum value, and UB is random maximum value. $\phi_1$ and $\phi_2$ are two random numbers, and their values range from 0 to 1. f stands for the current record value. $F_u$ is the maximum value of that spec, and $F_l$ refers to maximum value of the spec. since the ranges of specs in the investigated data set are between 0 and 10, we changed the above mentioned equations to the following equations:

$$LB = f - 9 \times \phi_1 \qquad (6)$$

$$UB = f + 9 \times \phi_2 \qquad (7)$$

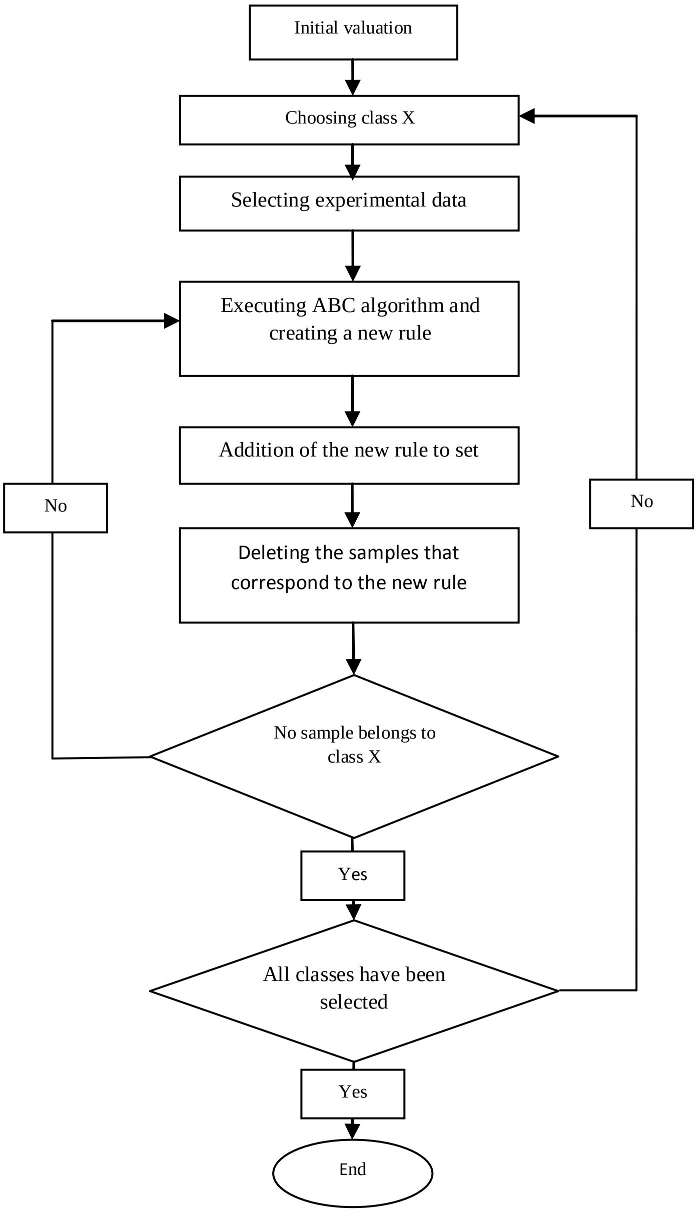
Initial valuation
Choosing class X
Selecting experimental data
Executing ABC algorithm and creating a new rule
Addition of the new rule to set
Deleting the samples that correspond to the new rule
No
No
No sample belongs to class X
Yes
All classes have been selected
Yes
End

Figure.1 flow chart of rule extracting

The flow chart of rule extraction has been shown in figure (1). The most important part in classification is rule extraction and exploration. since proper rules cannot be found, the efficiency of the method will be reduced. In this algorithm, rules are considered as the solutions or food sources, and they should be evaluated after extraction and exploration. In rule evaluation, the value of rule fitness is computed via equation (3), and the rule structure is stored in the memory. At the end of this step, there are many rules in the memory.

## 5.5 Rule Selection

As it was mentioned before, there are many rules in the memory. In order to use and apply these rules in classification, the best rule must be selected and considered. The criterion of selecting the best rule is fitness value. Through selecting the rule with maximum fitness value for each class, we consider data classification**.**

## 5.6 Selecting dominant Class

After selecting the best rule for each class, we select the class whose rule involves maximum fitness value. With regard to data available in WBCD set, the class whose distinguishing rule has the maximum fitness value is considered as benign class.

## 5.7 Classification

After selecting the rule related to dominant class, we try to classify data. In this step, each sample applied in rule condition in located in benign class. If a sample is not used in this rule, it will be located in malignant class, since there are two kinds of classes. This procedure will continue until all data are completed.

# 6. COMPARING THE METHOD WITH OTHER METHODS

In this paper, we compare the rules obtained from executing this algorithm in data related to breast cancer with other methods. In order to evaluate this method, the above mentioned algorithm is executed 10 times, and in each step, precision rate will be specified in diagnosis of samples. The results of this step have been demonstrated in table (1). The average precision rate of classification is 96,5% in the proposed method. The average precision rate of the proposed method and other methods has been shown in table (2). In order to use and apply the proposed method, visual studio, version 2005 and C# have been used. In addition, in order to compare this method with other methods, Weka software, version 3.6, has been used.

**Table 1. The Results of the Proposed Algorithm Repetition in Breast Cancer Diagnosis**

| Repetition Stage | Percent of diagnosis precision | The number of correct samples | The number of incorrect samples |
|---|---|---|---|
| 1 | 95.91 | 655 | 28 |
| 2 | 96.12 | 656 | 27 |
| 3 | 96.58 | 660 | 23 |
| 4 | 96.59 | 660 | 23 |
| 5 | 96.69 | 660 | 23 |
| 6 | 96.33 | 658 | 25 |
| 7 | 97.25 | 664 | 19 |
| 8 | 95.92 | 655 | 28 |
| 9 | 96.84 | 661 | 22 |
| 10 | 96.79 | 661 | 22 |
| Average | 96.5 | | |

**Table 2. Comparing the Proposed Algorithm with Other Methods**

| algorithm | Percent of diagnosis precision | The number of correct samples | The number of incorrect samples |
|---|---|---|---|
| The proposed method | 96.5 | 659 | 24 |
| Naive Bayes | 96.48 | 659 | 24 |
| Neural Network | 95.17 | 650 | 33 |
| decision tree-C4.5 | 95.75 | 654 | 29 |
| Support Vector Machine | 97.07 | 663 | 20 |

## 7. CONCLUSION

In this paper, the algorithm of artificial bee colony has been used for classification because this algorithm is a new, simple and efficient algorithm in terms of optimization. Classification is very important for data processing in data mining. In addition, breast cancer is one of the most prevalent cancers, and its growth can be prevented through on- time diagnosis and prognosis. In this paper, we have proposed a method on the basis of artificial bee colony algorithm; therefore, diagnosis of this illness will be possible through using data related to breast cancer samples. The results of this method have been compared with other methods of breast cancer diagnosis. It should be mentioned that diagnosis precision of this method has been considered as medium in comparison with other methods. This method is simpler than other methods, and its concepts are more understandable. In the future, we will try to increase the success percentage of the proposed method.

## 8. REFERENCES

[1] Amin E. 2011 A Fuzzy-ACO Method for Detect Breast Cancer Global Journal of Health Science : Vol. 3, No. 2 (October 2011)

[2] Afizi Mohd, Mujahid Ahmad, Zaidi Ahmad, "Artificial Bee Colony based Data Mining Algorithms for Classification Task", Modern Applied Science,Vol 5, No 4, 2011

[3] Anunciac Orlando ¸ Bruno C. Gomes, Susana Vinga, Gaspar Jorge, "A Data Mining Approach for the detection of High-Risk Breast Cancer Groups", 2004

[4] B. Basturk, D. Karaboga, An Artificial Bee Colony (ABC) algorithm for numeric function optimization, in: IEEE Swarm Intelligence Symposium, Indiana, USA,2006.

[5] B. Basturk ,D. Karaboga, Artificial Bee Colony (ABC) optimization algorithm for solving constrained optimization problems, LNCS: Advances in Soft Computing:Foundations of Fuzzy Logic and Soft Computing, vol. 4529, Springer Verlag, p. 789–798, 2007

[6] B. Basturk, D. Karaboga, A powerful and efficient algorithm for numerical function optimization: Artificial Bee Colony (ABC) algorithm, J. Global Optim. 39 ,171–459, 2007

[7] Bellaachia Abdelghani,Erhan guven, "Predicting Breast cancer survivability using Data MiningTechniques.", 2005

[8] Gadewadikar Jyotirmay, Ognjen Kuljaca, Kwabena Agyepong, Erol Sarigul, Yufeng Zheng, Ping Zhang, "Exploring Bayesian networks for medical decision support in breast cancer detection", 2010

[9] Gupta Shelly, Kumar Dharminder Dean, Anand Sharma, "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis", 2011

[10] D. Karaboga, "An idea based on honey bee swarm for numerical optimization",Technical Report-TR06, Erciyes University, Engineering Faculty, ComputerEngineering Department, 2005.

[11] Luiza Antonie Maria, Osmar R. Za¨ıane, Alexandru Coma, "Application of Data Mining Techniques for Medical Image Classification", 2001

[12] Wei Pin, Chang, Der-Ming, Liou , " Comparison of Three Data Mining Techniques with Genetic Algorithm in the Analysis of Breast Cancer Data ",2007

[13] Rani K. Usha, "Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique", 2002