ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

An Efficient Support Based Ant Colony Optimization Technique for Lung Cancer Data

Parag Deoskar[1], Dr. Divakar Singh[2], Dr. Anju Singh[3]

MTech Scholar CSE Deptt. BUIT,Barkatullah University, Bhopal[1]

HOD of CSE Deptt. BUIT,Barkatullah University, Bhopal[2]

Assistant Professor of IT Deptt. BUIT,Barkatullah University, Bhopal[3]

Abstract: Lung cancer is one of the most dangerous cancer's type in the world. Early detection can save the life and survivability of the patients. In this paper we want to propose a solution in the direction of lung cancer symptom detection. In this paper we proposed a new algorithm that is support based ant colony optimization technique (SPACO). Our algorithm is broadly divided into three parts, in the first part we accept the data set of cancer symptoms which is a generalized way for creating the patterns for Lung Cancer Framework, and in the second part we find the relevant data from the patterns. We can choose the frequent symptoms only by using the support count value. According to the support value we decide the ants and pheromone value. We initialize the pheromone value which is the support of the pattern of cancer symptoms. It is updated in each trial. By updating the pheromone value in each step we can check the symptom quality which either increases the prediction or decreases the prediction. Finally by result analysis we can prove the effectiveness of our algorithm.

Keywords: SPACO, ACO, data mining, rule pruning, Pheromone

1. INTRODUCTION

Decision classification is the most important task for mining any data set. A classification problem encompasses the assignment of an object to a predefined class according to its characteristics [7], [1]. There are several decision tasks which we observe in several fields of engineering, medical, and management related science can be considered as classification problems. Popular examples are pattern classification, speech recognition, character recognition, medical diagnosis and credit scoring.

But in our case classification alone is insufficient for classifying lung cancer dataset. If we consider data mining for frequent pattern classification then it is better tool for classifying relevant data from the raw dataset. Mining frequent item sets is the core problem of mining association rules; it determines the performance of association rules directly [4][13]. With the developing and more detailed of the research on frequent item sets mining, it is widely used in the field of data mining, for example, mining association rule, correlation analysis, classification, and clustering et al.[14]. The main aim of data mining is to extract knowledge from data. Data mining is an interdisciplinary field, whose core is at the intersection of machine learning, statistics, and databases. We emphasize to mine lung cancer data to discover knowledge that is not only accurate, but also comprehensible for the lung cancer detection [6], [8].

We provide here an overview of Image Compression Technique. The rest of this paper is arranged as follows: Section 2 introduces Ant Colony optimization; Section 3 describes about Literature Review; section 4 the shows the proposed approach; Result analysis is discusses in section5; Section 6 describes Conclusion and finally references are given.

2. ANT COLONY OPTIMIZATION

The Ant Colony Optimization (ACO) algorithm is a meta-heuristic that has a combination of distributed computation, autocatalysis (positive feedback), and constructive greediness to find an optimal solution for combinatorial optimization problems. This algorithm tries to mimic the ant's behaviour in the real world.

The ACO algorithm has been inspired by the experiments run by Goss et al. [2] using a colony of real ants. They observed that real ants were able to select the shortest path between their nest and food resource, in the existence of alternate paths between the two. The search is made possible by an indirect communication known as stigmergy amongst the ants. While traveling their way, ants deposit a chemical substance, called pheromone, on the ground. When they arrive at a decision point, they make a probabilistic choice, biased by the intensity of pheromone they smell. This behaviour has an autocatalytic effect because of the very fact that an ant choosing a path will increase the probability that the corresponding path will be chosen again by other ants in the future. When they return back, the probability of choosing the same path is higher (due to the increase of pheromone). New pheromone will be released on the chosen path, which makes it more attractive for future ants. Shortly, all ants will select the shortest path.

Figure 1 shows the behavior of ants in a double bridge experiment [5]. In this case, because of the same pheromone laying mechanism, the shortest branch is most often selected. The first ants to arrive at the food source are those that took the two shortest branches. When these ants start their return trip, more pheromone is present on the short branch than the one on the Long Branch. This behavior was formulated as Ant System (AS) by Dorigo et al. [3]. Based on the AS algorithm, the Ant Colony Optimization (ACO) algorithm was proposed [15]. In ACO algorithm, the optimization problem is formulated as a graph G = (C; L), where C is the set of components of the problem, and L is the set of possible connections or transitions among the elements of C. The solution is expressed in terms of feasible paths on the graph G, with respect to a set of given constraints. The population of agents (ants) collectively solves the problem under consideration using the graph representation. Though each ant is capable of finding a (probably poor) solution, good quality solutions can emerge as a result of collective interaction amongst ants. Pheromone trails encode a long-term memory about the whole ant search process. Its value depends on the problem representation and the optimization objective.

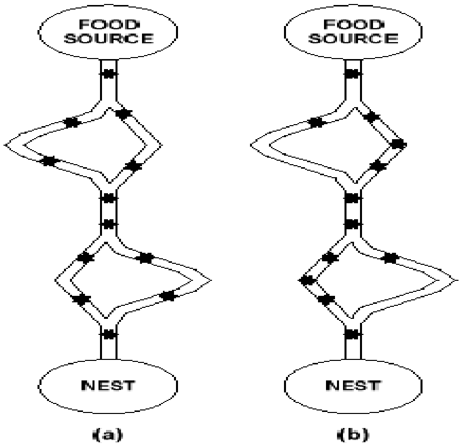


Figure 1: Double bridge experiment. (a) Ants start exploring the double bridge. (b) Eventually most of the ants choose the shortest path [5].

The algorithm presented by Dorigo et al. [15] was given below:

```
Algorithm ACO meta heuristic();
while (termination criterion not satisfied)
ant generation and activity();
pheromone evaporation();
daemon actions();
 "optional"
end while
end Algorithm
```

3. LITERATURE REVIEW

In 2011, Hnin Wint Khaing et al. [16] presented an efficient approach for the prediction of heart attack risk levels from the heart disease database. Firstly, the heart disease database is clustered using the K-means clustering algorithm, which will extract the data relevant to heart attack from the database. Their approach allows mastering the number of fragments through its k parameter. Subsequently the frequent patterns are mined from the extracted data, relevant to heart disease, using the MAFIA (Maximal Frequent Itemset Algorithm) algorithm. The machine learning algorithm is trained with the selected significant patterns for the effective prediction of heart attack. They have employed the ID3 algorithm as the training algorithm to show level of heart attack with the decision tree. The results showed that the designed prediction system is capable of predicting the heart attack effectively.

In 2008, S.H. Zainud-Deen et al. [17] proposed a hybrid technique based on Finite-difference frequency domain and particle swarm optimization techniques to reconstruct the breast cancer cell dimension and determine its position. Finite-difference frequency domain is formulated to calculate the scattered field after illuminating the breast by a microwave transmitter. Two-dimensional and three dimensional models for the breast are used. The models include randomly distributed fatty breast
tissue, glandular tissue, 2-mm thick skin, as well as chest wall tissue. The models are characterized by the dielectric properties of the normal breast tissue and malignant tissue at 800 MHz.

In 2012, M. H. Mehta et al. [18] observed that in engineering field, many problems are hard to solve in some definite interval of time. These problems known as "combinatorial optimisation problems" are of the category NP. These problems are easy to solve in some polynomial time when input size is small but as input size grows problems become toughest to solve in some definite interval of time. Long known conventional methods are not able to solve the problems and thus proper heuristics is necessary. Evolutionary algorithms based on behaviours of different animals and species have been invented and studied for this purpose. Particle swarm optimisation is a new evolutionary approach that copies behaviour of swarm in nature. However, neither traditional genetic algorithms nor particle swarm optimisation alone has been completely successful for solving combinatorial optimisation problems. So the authors present a hybrid algorithm in which strengths of both algorithms are merged and performance of proposed algorithm is compared with simple genetic algorithm.

In 2012, Priyanka Dhasal et al. [19] proposed a feature sampling technique of image classification. Their sampling technique optimized the feature selection process and reduced the unclassified region in multi-class classification. For the process of optimization they used ant colony optimization algorithm for the proper selection of feature sub set selection Support Vector Machines are designed for binary classification. When dealing with several classes, as in object recognition and image classification, one needs an appropriate multi class method. They also discuss about the possibilities which include: Modify the design of the SVM, as in order to incorporate the multi-class learning directly in the quadratic solving algorithm. Combine several binary classifiers: "One-against- One" (OAO) applies pair wise comparisons between classes, while "One-against-All" (OAA) compares a given class with all the others put together. OAO and OAA classification based on SVM technique is efficient process, but this SVM based feature selection generate result on the unclassified of data. When the scale of data set increases the complexity of preprocessing is also increases, it is difficult to reduce noise and outlier of data set.

In 2011, Yao Liu et al. [12] implement a classifier using DPSO with new rule pruning procedure for detecting lung cancer and breast cancer, which are the most common cancer for men and women. Experiment shows the new pruning method further improves the classification accuracy, and the new approach is effective in making cancer prediction.

In 2009, Ping-Hung Tang et al. [9] study and observed all of features and optimal feature subsets with three features are investigated. For classification, crisp k-NN, fuzzy k-NN, and weighting fuzzy k-NN classifiers are compared. For weighting of features, two types of coding including Binary-coded genetic algorithms (BGA) and real-coded genetic algorithms (BGA) are evaluated by the authors. Experiments are conducted on the Wisconsin diagnosis breast cancer (WDBC) dataset and the Pima (PIMA) Indians diabetes dataset, and the classification accuracy, false negative, and computation time are reported by the authors.

In 2011, Shyi-Ching Liang et al. [10] suggest Classification rule is the most common representation of the rule in data mining. It belongs to the supervised learning process which generates rules from training data set. The goal of the classification rule mining is the prediction of the predefined class. Rafael S. Parpinelli etc. proposed the Ant-Miner algorithm. Based on ACO algorithm, Ant-Miner solved the classification rule problem. In its basic configuration, Ant-Miner shows good performance in many dataset. In this research author proposed, an extension of Ant-Miner is proposed to incorporate the concept of parallel processing and grouping. Intercommunication via pheromone among ants is a critical part in ant colony optimization's searching mechanism. Due to the algorithm design, Ant-Miner made a slight modification in this part which removes the parallel searching capability. Based on Ant-Miner, they propose an extension that modifies the algorithm design to incorporate parallel processing. The pheromone trail deposited by ants during searching affected each other. With the help of pheromone, ants can have better decision making while searching. They provide a possible direction for researches toward the classification rule problem.

In 2011, Arezoo Modiri et al. [11] uses particle swarm optimization (PSO) algorithm is used to estimate the permittivity of the tissue layers at microwave frequency band. According to the literature, microwave radiometry (MWR) is potentially a promising cancer detection technique. In addition, breast cancer is an appropriate candidate of MWR due to the breast's exclusive physiology. Authors potential of PSO in solving this problem is demonstrated at 1-2.25 GHz. Two distinct algorithms are developed for the two considered scenarios. In the first scenario, they assume no a priori knowledge of the tissue under the test, whereas, in the second scenario, a priori knowledge is assumed. The algorithm converges relatively fast, and, distinguishes between different tissues with an acceptable accuracy.

4. PROPOSED SOLUTION

In this paper we propose a new algorithm which is based on data mining frequent pattern analysis with ant colony optimization. In our approach we consider the lung cancer datset from UCI repository (http://archive.ics.uci.edu/ml/datasets.html). The ants are the symptoms of the lung cancer. First we determine the support of each ants(symptom), so that we initialize the pheromone value as the initial support. Then by the help of random initialization helper allocates a trail and initializes it to 0, 1, 2, ... randomly. Next, the method uses the Association Algorithm to randomize the weight[20][21].

Pheromones are chemicals ants place on their trails; they attract other ants. More ants will travel on a shorter trail to a food source and deposit more pheromones than on longer trails. The pheromones slowly evaporate over time. Here's method

InitPheromones:

```
static double[][] InitPheromones(int numCities)
{
  double[][] pheromones = new double[numCities][];
  for (int i = 0; i < numCities; ++i)
    pheromones[i] = new double[numCities];
  for (int i = 0; i < pheromones.Length; ++i)
    for (int j = 0; j < pheromones[i].Length; ++j)
      pheromones[i][j] = 0.01;
  return pheromones;
}
```

Pheromone information is stored in an array-of-arrays-style symmetric matrix where the row index i is the from on trial and the column index j is the second trail. All values are initially set to an arbitrary small value (0.01) to jump start the Update Ants-Update Pheromones cycle.

Then we calculate the iteration which is based on the total number of symptoms qualified. So more number of symptoms must be justified with maximum iteration. Then the trail will be started according to the algorithm step 9 to 19. The algorithm for this approach is given below:

Algorithm: SPACO

1. Generate association rules
2. For each frequent itemset *X*,
3. For each proper nonempty subset *A* of *X*,
4. Let *B* = X - *A*
5. A → B is an association rules if
6. Confidence(A → B) ≥ minconf,
7. support (A → B) = support (A∪B) = support(X)
8. Create all symptoms as the ant for the feasible partial solution.
9. An ant k has a memory Mk that it can use to store information on the path it followed so far.
10. The stored information can be used to build feasible solutions, evaluate solutions and retrace the path b
11. An ant k can be assigned a start state and one termination conditions which depends on the i
12. Initialization is done by the support value.
13. Ants start from a start state and move to feasible neighbor states, building the solution in an incremental way. The procedure stops when at least one termination condition ek for ant k is satisfied.
14. An ant k located in node i can move to node j chosen in a feasible neighborhood Nki through probabilistic decision rules. This can be formulated as follows in 16.
15. An ant k can move to any node j in its feasible neighborhood with S is a set of all states.
16. A probabilistic rule is a function of the following.

a) The values stored in a node local data structure Ai = [aij] called ant routing table obtained from pheromone trails and heuristic values,

b) The ant's own memory from previous iteration, and

c) The problem constraints.

17. When moving from node i to neighbour node j, the ant can update the pheromone trails τij on the edge (i, j).
18. Once it has built a solution, an ant can retrace the same path backward, update the pheromone trails and die.

Weight updating is done according to the flowchart which is shown below. First it is initialize according to step 8. Then trail intensity is determined by step 9 –step 15 of our proposed algorithm. Then we determine the minimum value of each trail which is replaced by the pheromone trail (ptrail) if ptrail is higher than it is replaced otherwise no change. It will be continuous until the iteration is stopped.

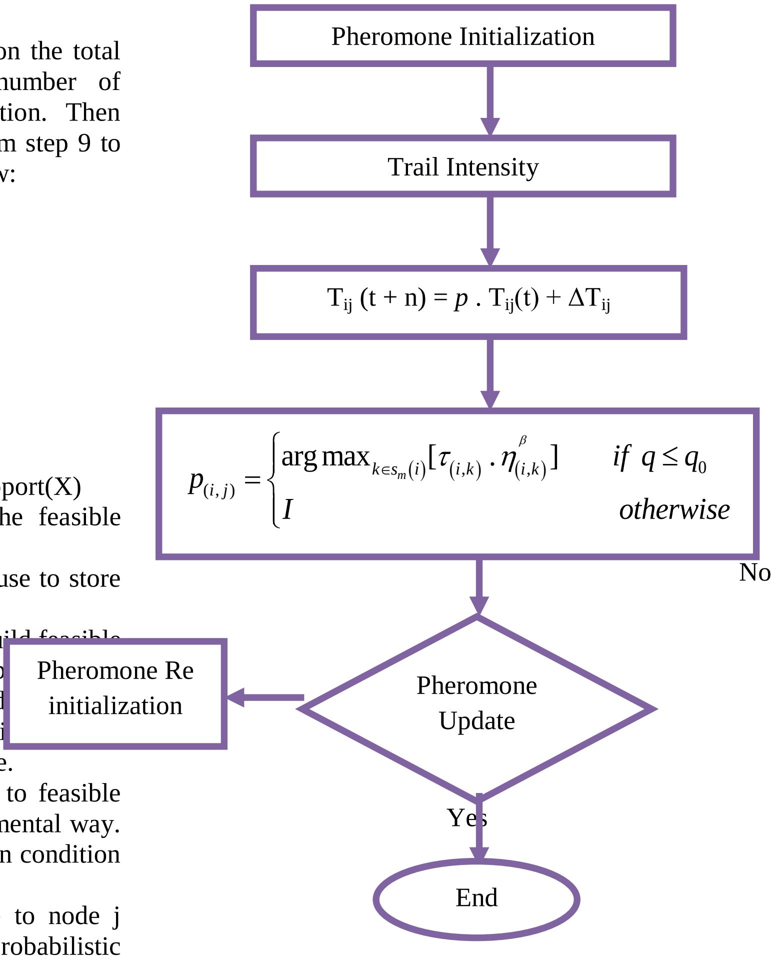

Pheromone Initialization
Trail Intensity
$T_{ij}(t+n) = p \,.\, T_{ij}(t) + \Delta T_{ij}$
$p_{(i,j)} = \begin{cases} \arg\max_{k\in s_m(i)} [\tau_{(i,k)} \,.\, \eta^{\beta}_{(i,k)}] & if\ q \le q_0 \\ I & otherwise \end{cases}$
No
Pheromone Re initialization
Pheromone Update
Yes
End

Figure 2: Weight Updating

Ants change the pheromone level on the paths between sessions using the following updating rule:

$$\tau_{(i,j)} \leftarrow \rho \cdot \tau_{(i,j)} + \Delta\tau_{(i,j)} \quad (2)$$

Where

ρ : the trail evaporation parameter.
$\Delta\tau_{(i,j)}$: the pheromone level.

The amount of deposited pheromone is the mechanism by which ants communicate to share information about good paths. Ant Colony System (ACS) differs from the other ACO instances due to its strategy of constructing an observation schedule.

This strategy can be categorized in three step. An ant positioned on session *i* selects the session *j* to observe by applying the following equation:

$$p_{(i,j)} = \begin{cases} \arg\max_{k \in s_m(i)}[\tau_{(i,k)} \, . \, \eta^{\beta}_{(i,k)}] & if \; q \le q_0 \\ I & otherwise \end{cases}$$

where
I : a random variable selected according to the probability given by Equation 1.

q : a uniformly distributed random number to determine the relative importance of exploitation versus exploration q∈[0,..,1].

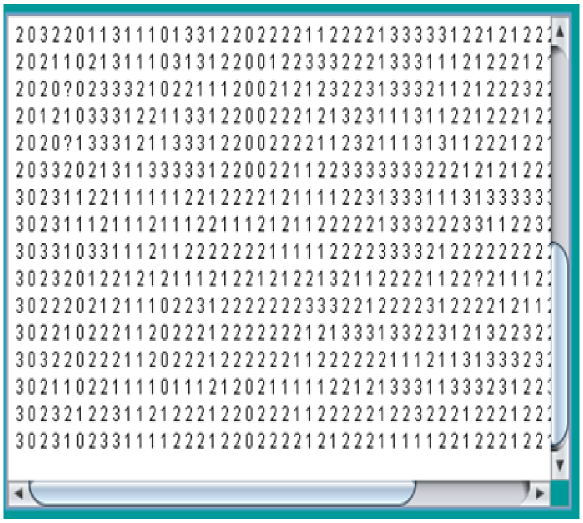
q_0 : a threshold parameter and the smaller q_0 the higher the probability to make a random choice ($0 \le q_0 \le 1$).

After applying the above phenomena we can reach to the ptrail which is better and help in detecting lung cancer and improves the accuracy.

5. RESULT ANALYSIS

In this paper we took the data set from UCI repository (http://archive.ics.uci.edu/ml/datasets/Lung+Cancer). This data was used by Hong and Young to illustrate the power of the optimal discriminant plane even in ill-posed settings. In the dataset 1,2,3 shows the different types of cancer. 0 means the cancer symptom is not found. ? means the data is unknown, which signifies that the symptom can be presented or not is not clear.

We first find the patterns from the lung cancer dataset. The dataset is shown in figure 3.



```
20322011311101331220222211222213333122121222
202110213111031312200122333222133311121222212
2020?02333210221112002121232231333211212222322
201210333122113312200222121323111311221222122
2020?13331211333122002222112321113131122212221
203320213113333312200221122333333322212121222
302311221111112212222121111223133311131333333
302311121112111221112121122222133322233112233
303310331112112222222111112222333321222222222
3023201221212111212212122132112222112 2?211122
302220212111022312222222333221222231222212111
302210222112022212222222121333133223121322322
303220222112022212222221122222211121131333233
302110221111011121202111112212133311333231223
302321223112122212202221122222122322212221222
3023102331111122212202222121222211111221222122
```

Figure 3: Dataset

Then we find the frequency of the dataset, which is shown in table 1.

Table 1: Frequency

Type	Frequency
1	480
0	107
3	278
?	5
2	954

Then we analyse that for 107 cases there is no detection and for 5 cases the result is ambiguous means the possibility and non-possibility of symptoms are equal.

For applying ant colony optimization we set the types as the ant and the support value as the pheromone.
Support of 1:480/587=81.77%
Support of 2: 954/1061=89.91 %
Support of 3:278/385=72.20%

The above phenomenon is shown in table 2. Then updating is performed by the pheromone trails and evaporation value of the pheromone. It is updated by the below formula:

∑P1+P2+P3…….+Pn-evaporation value

P1,P2,P3 are the pheromone value. In our case the evaporation value is 0.2 which is increased by 0.2. Means in first trial we subtract by -0.2 then by 0.4, then by 0.6 and finally by 0.8. The range is from 0 to 1. If the evaporation value is exceeds from the range then in this case the evaporation value is fixed by 0.8. The Updation is only performed when the values of pheromone trail is less than the pheromone value which is shown in table 3. Finally by applying the ant colony optimization technique we achieve better detection accuracy for each type of cancer symptom. In our case the individual symptom accuracy we detected is 82 %, 90 % and 81% for 1, 2,3 type of cancer symptom. This phenomenon is shown in figure 4. Then we find the average accuracy which is 84 %.

We compare our result by Yao Liu and Yuk Ying Chung research [12]. In [12] the result shows that for breast cancer data, their proposed DPSO is on par with Naive Bayes with the highest accuracy. And as for the lung cancer data, they achieve 68.33% by DPSO which performs well by getting the highest accuracy of which is around 11.7% and 20.2% better than Naive Bayes and SVM respectively, and 3.8% more accurate than PSO. According to [16] the accuracy observed by them is following:

1) DPSO (new) 68.33
2) PSO (new) 64.44
3) PART 48.14
4) SMO 48.14
5) Naive Bayes 56.67
6) KNN 45
7) Classification Tree 46.67

We consider all the above results for comparison. The comparison graph is shown in Figure 5. In our case when comparing with those values we achieve better detection rates as comparison to the traditional technique as shown in figure 5.

Table 2: Initial Pheromone

Initial Pheromone	
Ants	Pheromone
1	0.82
2	0.9
3	0.72

Table 3: Pheromone Trails

Initial Pheromone	
Ants	Pheromone
1	0.82
2	0.9
3	0.8133

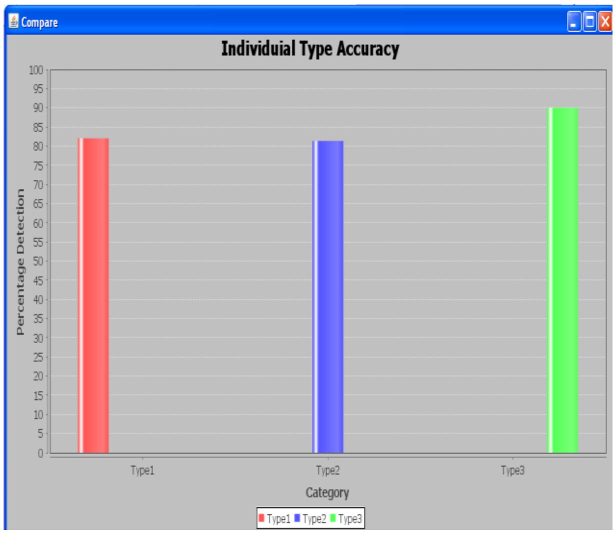


Figure 4:Individual Type Accuracy

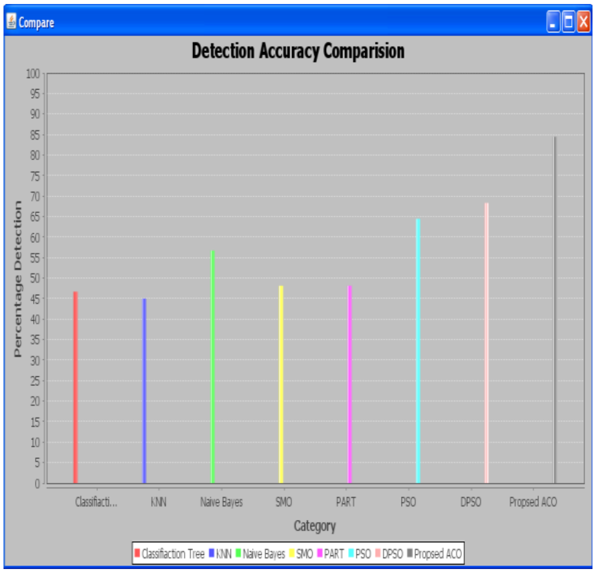


Figure 5: Detection Accuracy

6. CONCLUSION AND FUTURE SCOPE

The use of data mining techniques in Lung cancer classification increases the chance of making a correct and early detection, which could prove to be vital in combating the disease. In this paper we present an effective approach which is based on association and optimization for lung cancer prediction. Our result shows the effectiveness of our approach. In future it can be extended for other cancers like breast cancer, heart diseases and blood cancer.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

REFERENCES

[1] D. J. Hand and S. D. Jacka, Discrimination and Classification. New York: Wiley, 1981.

[2] S. Goss, S. Aron, J. L. Deneubourg, and J. M. Pasteels. Self-organized Shorcuts in the Argentine Ant. Naturwissenschaften, 76:579–581, 1989.

[3] M. Dorigo and M. Maniezzo and A. Colorni. The Ant Systems: An Autocatalytic Optimizing Process. Revised 91-016, Dept. of Electronica, Milan Polytechnic, 1991.

[4] Agrawal R and Srikant R, "Fast Algorithms for Mining Association Rules", Proc of the International Conference on Very Large Databases. Santiago,USA, 1994:487-499.

[5] M. Dorigo, Gianni Di Caro, and Luca M. Gambardella. Ant Algorithms for Discrete Optimization. Technical Report Tech. Rep. IRIDIA/98-10, IRIDIA, Universite Libre de Bruxelles, Brussels, Belgium, 1998.

[6] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in Advances in Knowledge Discovery & Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth,and R. Uthurusamy, Eds. Cambridge, MA: MIT Press, 1996, pp. 1–34.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification. New York: Wiley, 2000.

[8] Junzo Watada, Keisuke Aoki, Masahiro Kawano, Muhammad Suzuri Hitam, Dual Scaling Approach to Data M Journal of Advanced Computational Intelligence Intelligent Informatics (JACIII), Vol. 10, No. 4, pp. 441-447, 2006.12.

[9] Ping-Hung Tang, Ming-Hseng Tseng," Medical Data Mining Using BGA and RGA For Weighting of Features In Fuzzy KNN Classification", Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009.

[10] Shyi-Ching Liang, Yen-Chun Lee and Pei-Chiang Lee , "The Application of Ant Colony Optimization to the Classification Rule Problem", 2011 IEEE International Conference on Granular Computing.

[11] Arezoo Modiri and Kamran Kiasaleh," Permittivity Estimation for Breast Cancer Detection Using Particle Swarm Optimization Algorithm", 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, August 30 - September 3, 2011

[12] Yao Liu and Yuk Ying Chung, "Mining Cancer data with Discrete Particle Swarm Optimization and Rule Pruning", IEEE 2011.

[13] Amog Rajenderan," Data Preparation for Web Mining A survey", International Journal of Advanced Computer Research (IJACR),Volume-2 Number-4 Issue-6 December-2012.

[14] Pragati Shrivastava,Hitesh Gupta, "A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (IJACR) ,Volume-2 Number-3 Issue-5 September-2012.

[15] M. Dorigo and G. Di Caro. New Ideas in Optimisation. McGraw Hill, London, UK, 1999.

[16] Hnin Wint Khaing," Data Mining based Fragmentation and Prediction of Medical Data", IEEE 2011.

[17] S.H. Zainud-Deen, Walaa M. Hassen, E. M. Ali, K.H. Awadalla and H.A. Sharshar, "Breast Cancer Detection Using a Hybrid Finite Difference Frequency Domain and Particle Swarm Optimization Techniques", 25th National Radio Science Conference (NRSC 2008).

[18] M. H. Mehta," Hybrid Genetic Algorithm with PSO Effect for Combinatorial Optimisation Problems", International Journal of Advanced Computer Research (IJACR), Volume-2 Number-4 Issue-6 December-2012.

[19] Priyanka Dhasal, Shiv Shakti Shrivastava, Hitesh Gupta, Parmalik Kumar, "An Optimized Feature Selection for Image Classification Based on SVMACO", International Journal of Advanced Computer Research (IJACR) ,Volume-2 Number-3 Issue-5 September-2012.

[20] Anshuman Singh Sadh, Nitin Shukla," Association Rules Optimization: A Survey", International Journal of Advanced Computer Research (IJACR), Volume-3 Number-1 Issue-9 March-2013.

[21] Anshuman Singh Sadh, Nitin Shukla," Apriori and Ant Colony Optimization of Association Rules", International Journal of Advanced Computer Research (IJACR) Volume-3 Number-2 Issue-10 June-2013.