International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur, Tamilnadu,India.

ASSOCIATION RULE MINING THROUGH THE ANT COLONY OPTIMIZATION TECHNIQUES TO PREDICT CHRONIC OBSTRUCTIVE PULMONARY DISEASE

Dr. R. Hemalatha,
Associate Professor and Head,
PG & Research, Department of Computer Science,
Tiruppur Kumaran College for Women,
Tirupur,TamilNadu,India.

A.Babyshalini,
M.Phil Research Scholar,
PG & Research, Department of Computer Science,
Tiruppur Kumaran College for Women,
Tirupur, TamilNadu, India.

Abstract: Chronic Obstructive Pulmonary Disease (COPD) is the important cause of death in the world. Some most common chronic diseases are COPD, diabetes, cardiovascular disease and chronic respiratory disease. Early detection can save the life and survivability of the patients. In this paper propose model give a solution to predict chronic diseases. In this paper proposes a novel approach of applying the Ant Colony Optimization technique (ACO) for extracting the Association Rules (AR) from the database to detect COPD. This algorithm is broadly divided into three parts, in the first part, accept the data set of chronic symptoms which is a generalized way for creating the patterns for Chronic diseases Framework, and in the second part, find the relevant data from the patterns. It can choose the frequent symptoms only by using the support count value. The pheromone value which the support of the pattern of COPD symptoms. Subsequently by outcome analysis, prove the effectiveness of this algorithm. The focus of this paper is to provide specific information about chronic diseases for public. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. Preventive Health care knowledge is essential for clinical and administrative decision making.

Keywords: Chronic diseases, Association Rule (AR), Ant colony optimization (ACO), COPD.

I. INTRODUCTION

This paper is focused on the Chronic Obstructive Pulmonary Disease (COPD) for which Data mining method are used to analysis the collected data. COPD is one of major cause of humanity in developed and developing countries. COPD has a significant impact in the quality life of patient and the most common condition that requires hospitalization. COPD is a complex, chronic and progressive disease characterized by the chronic inflammation and irreversible air flow obstruction, which involves structural changes in the lung, figure 1 represents the Emphysema and Chronic Bronchitis in COPD. The principal symptoms are the difficulty in breathing, cough and expectoration and it is represented in figure 2. In the clinical presentation are different phenotypes, very heterogeneous, with prognostic and therapeutic clinical repercussions [1]. Though COPD is not a curable disease, smoking is the most effective measure for prevention and to stop the progression. COPD's clinical diagnosis must think about every patient with a respiratory difficulty, chronic cough or high production of secretions and a history of exposition to risk factors of the diseases [2].

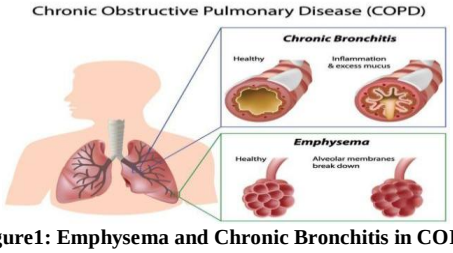


Figure1: Emphysema and Chronic Bronchitis in COPD

Different authors indicate as significant factors related to this condition: the masculine sex, age, consumption of tobacco (number of packages per year), cough, expectoration, difficulty in breathing and other respiratory symptoms [3-7].

Figure 2: Symptoms of COPD

This disease is characterized by some of the following symptoms: chronic cough, sputum production and dyspnoea. It can be confirmed in a clinical exam called spirometry, if they obtained values are 80% below of the Forced Expiratory Volume in 1 second (FEV1) and the ratio FEV1/FVC (Forced Vital Capacity) is lower than 0,7 [9]. This disease can be classified in four stages, according to the degree of severity in figure 3. The first stage, or Mild COPD, is characterized by a FEV1 value above or equal to 80% and the presence, or not, of chronic cough and sputum production. COPD is not usually detected at this first stage. The second stage, or Moderated COPD, is characterized by a value of the FEV1 between 50% and 79%, shortness of breath during exertion, chronic cough and sputum production. The third stage, or severe COPD, is characterized by a value of the FEV1 between 30% and 49%, greater shortness of breath, reduced exercise capacity and

fatigue. The fourth stage, or very severe COPD, is characterized by a value of FEV1 below 30% and the presence of chronic respiratory failure [8,9].

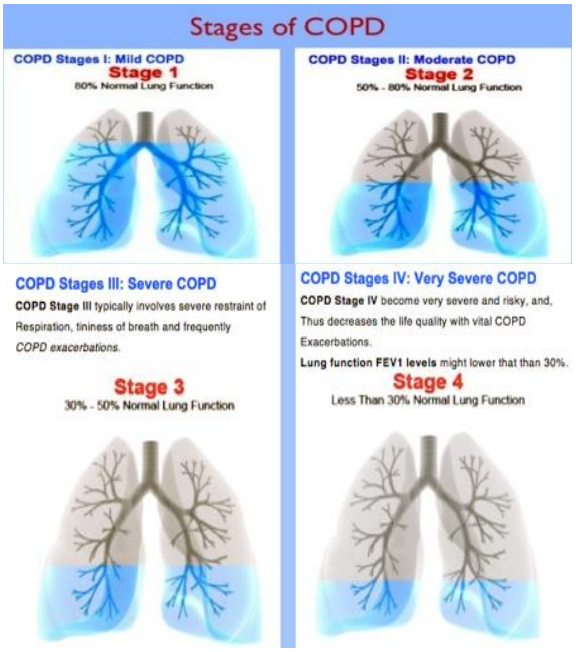


Figure 3: Stages of COPD

Data mining technique means the use of sophisticated data analysis tools to determine previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods in early detection of chronic disease. In classification learning, the learning strategy is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples. In association learning, any association among features is sought, not just ones that predict a particular class value. Association rule can be easily converted into classification rules. This decision tree is used to generate frequent patterns in the dataset. The data and item sets that occur frequently in the data base are known as frequent patterns. The frequent patterns that is most significantly related to specific chronic disease types and are helpful in predicting and its type is known as significant frequent pattern. Using this significant patterns generated by decision tree the data set is associate accordingly and risk scores are given.

Data mining techniques are implemented together to create a novel method to diagnose the existence of chronic disease for a particular patient. When beginning to work on a data mining problem, it is first necessary to bring all the data together into a set of instances. Integrating data from different sources usually presents many challenges. The data must be assembled, integrated, and cleaned up. Only then it can be used for processing through machine learning techniques. This developed system can be used by physicians and patients alike to easily know a person's chronic disease status and severity without screening them for testing chronic disease [10]. Also it is useful to record and save large volumes of sensitive information which can be used to gain knowledge about the disease and its treatment.

1.3 MINING ASSOCIATION RULES

Mining association rules from a large database of business data, such as transaction records, has been an important issue in the field of data mining. The problem of association rule mining can be divided into two sub-problems: (1) frequent itemset discovery and (2) association rules generation. It has also been shown that the overall performance of mining is seriously determined by the first sub-problem. Frequent itemset mining algorithms often generate a very large number of frequent itemsets and rules, which reduce both the efficiency and also the effectiveness of the mining algorithms since only the subset of the complete frequent itemsets and association rules is of interest to users. In addition, the users need an additional post-processing step to filter the large number of mined rules to determine the useful ones.

1.4 ANT COLONY OPTIMIZATION

The Ant Colony Optimization (ACO) algorithm is a meta heuristic that has a sequence 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.

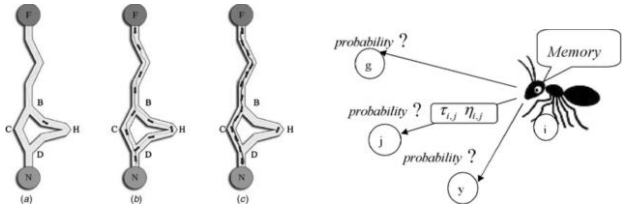


Figure 4: The behaviour of real ants.

They exploit user-specific constraints in the mining process to improve performance, or efficiency. Therefore, this study intends to use the ant colony system, which has recently been shown to be very promising in the areas of the travelling salesman problem and scheduling [11], for multiple dimensional constraints mining association rules. Furthermore, since data mining has rarely been applied to solve questions in medical science, this study uses data to find disease association rules. Here, an important issue is to find the potential disease and early prevention. The evaluation results show that the proposed method, using the ant colony system, really can provide more concise and accurate information than the conventional Association rule -based algorithm. The rest of this paper is standardized as follows. Section 2 summarizes some important literature review, and the workflow of proposed method in section 3, the proposed method is described in Section 4. Section 5 presents the evaluation results and discussion. Finally, concluding remarks are made in Section 6.

II.LITERATURE REVIEW

Different studies have been carried out by researchers which focus on diagnosis of chronic disease using the data available. Smoking is the biggest risk factor of chronic disease. The more years and larger number of cigarettes smoked the greater

the risk of developing chronic disease. The average age of someone diagnosed with chronic disease is 65 to 70 years old, but people who are younger can develop chronic disease.

Krishnaiah V, et al.,[13] proposed to a model for nearly detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient. Using generic chronic disease symptoms such as age, sex, Wheezing, Shortness of breath, Pain in shoulder, chest, arm, it can predict the likelihood of patients getting a chronic disease.

ParagDeoskar, et al.,[14] proposed to assorted data mining and ant colony optimization techniques for appropriate rule generation and classification, which pilot to exact COPD classification. In addition to, it provides basic framework for further improvement in medical diagnosis. This paper also surveys the aspects of ant colony optimization (ACO) technique. Ant colony optimization helps in increasing or decreasing the disease prediction value.

Sowmiya T, et al., [15] Chronic disease is one of the most dangerous COPD types in the world. These diseases can increased worldwide by uncontrolled cell growth in the tissues of the lung. Early detection of the COPD can save the life and survivability of the patients who affected by this diseases. Figure 5 represent the healthy vs affected COPD. In this paper we survey several aspects of data mining procedures which are used for chronic disease prediction for the patients.

Data mining concepts are useful in chronic disease classification. In this paper the aspects of ant colony optimization (ACO) technique in data mining is also be revised. Ant colony optimization helps in increasing or decreasing the disease prediction value of the diseases. This case study assorted data mining and ant colony optimization techniques for appropriate rule generation and classifications on diseases, which pilot to exact chronic disease classifications. In additionally to, it provides basic framework for further improvement in medical diagnosis on chronic disease.

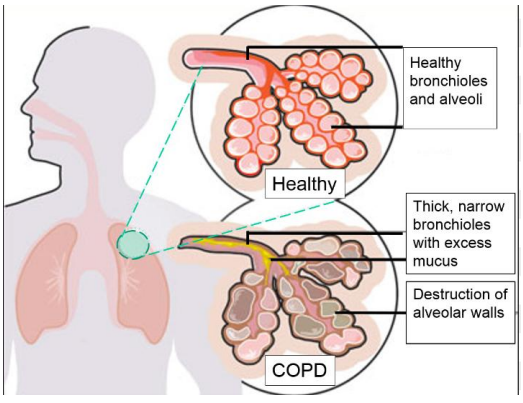


Figure 5: Healthy vs Affected COPD

III.EXISTING METHOD

COPD is the number one cause of cancer deaths in both gender worldwide. Smoking is the principal risk factor for development of lung cancer. The stage of COPD refers to the extent to which the cancer has spread in the body. Overall, 10-15% of COPDs occur in non-smokers. (Another 50% occur in former smokers). Two-thirds of the non-smokers who get COPD are women and 20% of COPDs in women occur in individuals who have never smoked. COPD is the most critical reason for death. Cancer research is generally clinical and/or biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. As the use of computers powered with automated tools, large volumes of medical data are being collected and made available to the medical research groups. Affect of respiratory system in COPD in figure 6.

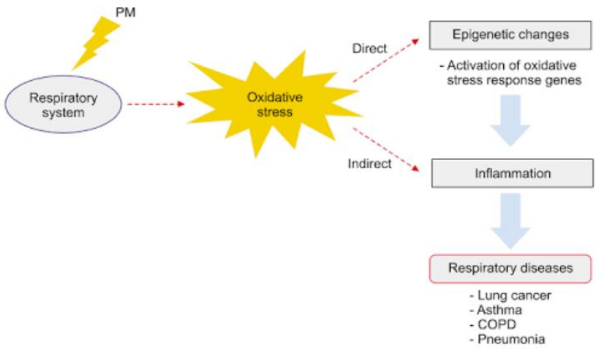


Figure 6: COPD affects in respiratory system

IV. WORK FLOW

The methodology used in this work is shown in the Figure 7. From the given input dataset the training dataset is selected and classified using the AR algorithm. Then to the classified results the optimization algorithm ACO is applied and the results are generated. The objective is to increase the level of performance of accuracy and to reduce the error rates using classification techniques integrating with optimization methods like ACO. In this paper a new algorithm, ARACO is proposed and this integrated approach helps in increasing the accuracy for better results. The performance is measured based on trained values of these methods in the given dataset.

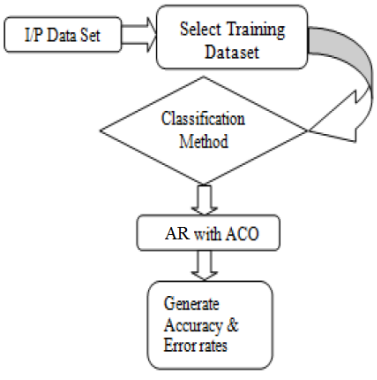


Figure 6: Workflow diagram

V. 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 chronic disease dataset from UCI repository (http://archive.ics.uci.edu/ml/datasets.html)[12]. The ants are the symptoms of the lung COPD. 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 [16][17].

Pheromones are chemicals ants place on their trails; they attract other ants. Further ants will travel on a shorter trail to a food resource and deposit more pheromones than on lengthy trails. The pheromones slowly evaporate over time. 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 many numbers 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: AR-ACO

- Generate association rules
- For each frequent itemset *Y*,
- For each proper nonempty subset *P* of *Y*,
- Let *Q* = Y - *P*
- P -> Q is an association rules if
- Confidence(P -> Q) ≥ minconf,
- support (P -> Q) = support (PUQ) = support(Y)
- Create all symptoms as the ant for the feasible partial solution.
- An ant k has a memory Mk that it can use to store information on the path it followed so far.
- The stored information can be used to build feasible solutions, evaluate solutions and retrace the path backward.
- An ant k can be assigned a start state and more than one termination conditions which depends on the iteration.
- Initialization is done by the support value.
- Ants start from a start state and move to suitable neighbour states, building the solution in an incremental way. The procedure stops when at least one termination condition ek for ant k is fulfilled.
- An ant k located in node i can move to node j chosen in a feasible neighbourhood Nki through probabilistic decision rules. This can be formulated as follows in 16.
- An ant k can shift to any node j in its suitable neighbourhood with S is a set of all states.
- A probabilistic rule is a function of the following.
- The values stored in a node local data structure $P_i = [p_{ij}]$ called ant routing table attained from pheromone trails and heuristic values,
- The ant's own memory from previous iteration, and The problem constraints.
- When shifting from node i to neighbour node j, the ant can update the pheromone trails τ_{ij} on the edge (i, j).
- 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 in figure 8. 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[18].

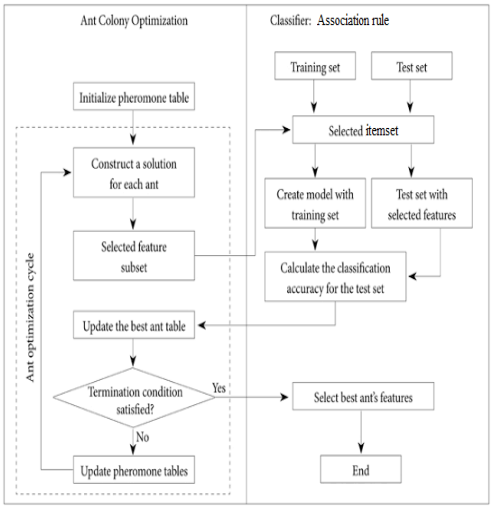


Figure 8: Algorithm flow

Real data for COPD has been collected and preprocessed using the data mining approach and the data are converted into CSV format for further processing. ACO algorithms have been implemented using MATLAB R2010a. Parameters for ant is set as 100, 1000 and execution time, accuracy were estimated for the two algorithms for 5 iteration[19]. Accuracy and execution time for two algorithms applied to different data set.

VI.PROPOSED FRAMEWORK

A. *Data Set*

Real COPD dataset contains 125 instances, 13 attributes. It is represented in the table 1.

S.no	Attribute name	Type
1	age	Real
2	sex	Binary
3	chest pain type (4 values)	Nominal
4	resting blood pressure	Real
5	serum cholestoral in mg/dl	Real
6	fasting blood sugar > 120 mg/dl	Binary
7	resting electrocardiographic results (values 0,1,2)	Nominal
8	maximum heart rate achieved	Real
9	exercise induced angina	Binary
10	oldpeak = ST depression induced by exercise relative to rest	Real
11	the slope of the peak exercise ST segment	Ordered
12	number of major vessels (0-3) colored by flourosopy	Real
13	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect	Nominal

Table 1:Dataset of COPD

The dataset is obtained from UCI repository. The database of patients is collected and the most accurate and major attributes and feature which help in predicting COPD such as: age(in years),sex(m/f),chest pain type(typical angina, atypical angina, non angina , asymptomatic),resting blood pressure(in mmHg),cholesterol(mg/dl), high fasting blood sugar(1/0),resting electrocardiographic results(1/0), maximum heart rate achieved, resting heart rate , exercise induced angina(1/0), St depression induced by exercise relative to rest,number of color vessels, obesity, thal(3=normal, 6= fixed defect, 7=reversible defect) are considered.

There have been frequent ways to predict the risk of COPD; Figure 9 shows the basic flow of prediction that is followed.

B. *Analysis of Data:*

It is one of the most vital steps as the data in the database contain most of the surplus and noisy data. So by analysis of data, data cleaning/preprocessing, data integration can be performed to fill up misplaced values, eliminate the redundant data because handling missing value and unnecessary data would lead to erroneous output.

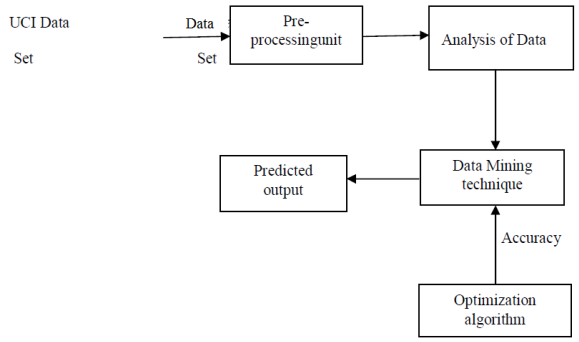


Figure 9: The framework of COPD

C. *Optimization of Algorithm:*

Various algorithms can be applied here to explore the best attribute which will evaluate the fitness value which is assigned to each attribute (individual). Algorithms which can be used can be Ant Colony Optimization.

D. *Prediction Engine:*

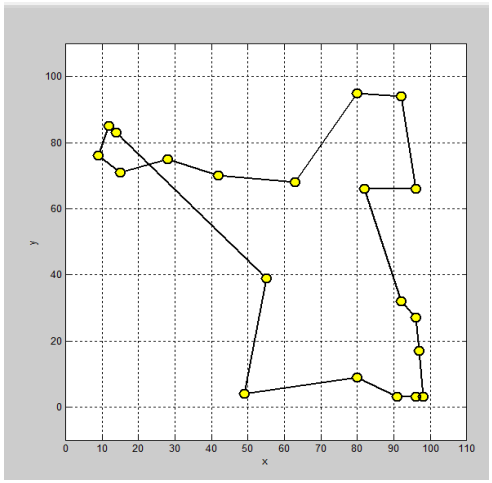
Predicts whether the person has a COPD or will suffer in future.

VII.MODEL EVALUATION RESULTS AND DISCUSSION

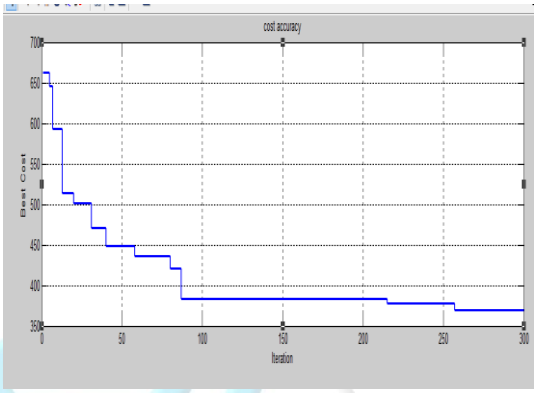
This section demonstrates how the proposed algorithm works by using the UCI Database to find COPD using association rules with ACO. Graph 1, represent the iteration of ACO. Then, the results from an expert questionnaire are employed to demonstrate the reliability of the mining rules. They are also compared with the Apriori method.

Graph 2 and 3, represents the cost accuracy of Ant Colony Optimization. In graph 2, cost accuracy are shown in X and Y axis. Graph 3, plots the cost accuracy in three directions i.e., X, Y and Z axis.

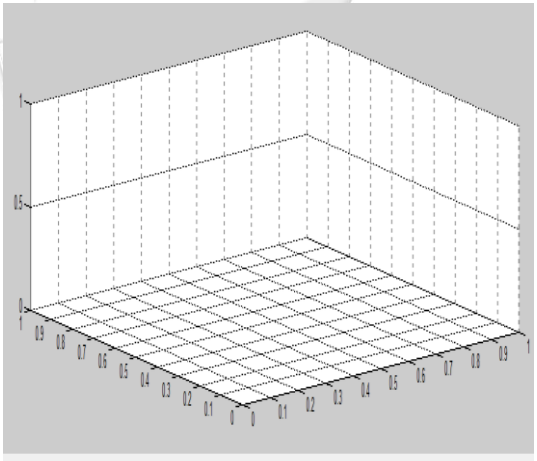
The variations in ACO algorithms are better in detection of COPD than the normal ACO algorithm. Accuracy is used to estimate the process. Accuracy measures the degree of closeness of measurement of a quantity to the actual value of the quantity.



Graph 1: Ant Colony Optimization



Graph 2: Cost accuracy of ACO



Graph 3: Cost accuracy of ACO in 3D

VIII.CONCLUSION

The use of data mining techniques in chronic disease classification increases the chance of making a correct and early detection, which could prove to be vital in combating the disease. In this work, an effective approach which is based on association and optimization for chronic disease prediction. Our result shows the effectiveness of our approach. In future, the proposed work may be combined with Neural and Genetic Algorithms and can also be applied for the unsupervised

learning. The future enhancement can be applicable for image based analysis for MRI and CT scan images.

IX.REFERENCE

[1]. Miravitlles M, Calle M, Soler-Cataluña JJ (2012) Clinical phenotypes of COPD:identification, definition and implications for guidelines. Arch Bronconeumol 48: 86-98.

[2]. Qaseem A, Wilt TJ, Weinberger SE, Hanania NA, Criner G, et al. (2011) Diagnosis and management of stable chronic obstructive pulmonary disease: a clinical practice guideline update from the American College of Physicians,American College of Chest Physicians, American Thoracic Society, and European Respiratory Society. Ann Intern Med 155: 179-191.

[3]. Price DB, Tinkelman DG, Halbert RJ, Nordyke RJ, Isonaka S, et al. (2006) Symptom-based questionnaire for identifying COPD in smokers. Respiration 73: 285-295.

[4]. Freeman D, Nordyke RJ, Isonaka S, Nonikov DV, Maroni JM, et al. (2005) Questions for COPD diagnostic screening in a primary care setting. Respir Med 99: 1311-1318.

[5]. Dirven JA, Tange HJ, Muris JW, van Haaren KM, Vink G, et al. (2013) Early detection of COPD in general practice: patient or practice managed? A randomised controlled trial of two strategies in different socioeconomic environments. Prim Care Respir J 22: 331-337.

[6]. Dirven JA, Tange HJ, Muris JW, van Haaren KM, Vink G, et al. (2013) Early detection of COPD in general practice: implementation, workload and socioeconomic status. A mixed methods observational study. Prim Care Respir J 22: 338-343.

[7]. López Varela MV, Montes de Oca M, Halbert R, Muiño A, Tálamo C, et al. (2013) Comorbidities and health status in individuals with and without COPD in five Latin American cities: the PLATINO study. Arch Bronconeumol 49: 468-474.

[8]. Rabe KF, Hurd S, Anzueto A, Barnes PJ, Buist SA, Calverley P, Fukuchi Y, Jenkins C, Rodriguez-Roisin R, van Weel C, Zielinski J, Global Initiative for Chronic Obstructive Lung D. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Gold executive summary. *American journal of respiratory and critical care medicine*. 2007;176:532-555

[9]. GOLDCOPD. From the global strategy for the diagnosis, management and prevention of copd, global initiative for chronic obstructive lung disease (gold) 2011. . 2011

[10]. Thangaraju P, Karthikeyan T, Barkavi G, Mining COPD Data for Smokers and Non-Smokers by Using Data Mining Techniques, *International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014.*

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

[12]. 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.

[13]. Krishnaiah V, Narsimha G, Subhash Chandra N , Diagnosis of COPD Prediction System Using Data Mining Classification Techniques, *International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45.*

[14]. Parag Deoskar, Dr. Divakar Singh, Dr. Anju Singh, Mining COPD Data And Other Diseases Data using Data Mining Techniques, *A Survey, Volume 4, Issue 2, March – April (2013).*

[15]. Sowmiya T, Gopi M, Thomas Robinson, Optimization of COPD using Modern Data Mining Techniques, *International Journal of Engineering Research Volume No.3, Issue No.5.*

[16]. 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.

[17]. 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.

[18]. Karaboga, N., A. Kalinli, and D. Karaboga, \Designing digital IIR filters using ant colony optimisation algorithm," Engineering Applications of Artificial Intelligence, Vol. 17, No. 3, 301{309, 2004.