International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

ACO Based Feature Selection : An Application for Medical Diagnosis

Nirmala M.
Department of Computer Science, Pondicherry University, Puducherry, Tamil Nadu, India

ABSTRACT

Medical Diagnosis is a system for the examination of a man's symptoms based on disorder. This issue has been examined and applied to several healthcare systems in medication. It has substantial fascination in the area of computer science because of huge cause for various ailments. The medical dataset contains immense number of immaterial and repetitive features. Not all the features are required to analyze whether the specific patient is having that specific sickness or not. Feature Selection (FS) is the system for finding the most essential features for a predictive model. This system is used to discover and eliminate not required, insignificant and repetitive features that do not contribute or diminish the accuracy of the predictive model. In this paper, we address the medical diagnosis by utilizing feature selection of Ant Colony Optimization (ACO). It is a nature inspired heuristic algorithm. By utilizing this algorithm, we will attempt the medical diagnosis using FS. Increment in features may lead to decrease in accuracy of the model. In order to attain the increase in accuracy of the model feature selection is used.

Keywords: Ant Colony Optimization, Feature Selection, Medical Diagnosis

I. INTRODUCTION

The liver plays the vital role in several body functions from protein production and blood clotting to cholesterol, glucose (sugar), and iron metabolism. There are more than hundred kinds of liver disease, which include hepatitis, alcoholic liver disease and fatty liver disease. Medical diagnosis is a difficult process in which numerous factors have involved. Machine learning approaches are very useful in medical diagnosis and treatment in which it reduced the time for diagnosing and it can help the physician in case of the treating the disease. Due to large number of unavailability of the physician, it is needed to use the machine learning process to diagnose the disease. A feature selection algorithm has also been proposed in order to identify the most stable, significant and discriminate features. Different techniques have been utilized for improving the accuracy of the classifier by discarding redundant and irrelevant features. Feature selection is also known as dimensionality reduction, which is used to reduce the number of features. By using this we can reduce the number of features and it can be improved the accuracy of the model. The attributes which are not contributed or lesser contributed toward the accuracy of the model can be avoided. So that the computation time can be reduced.

Feature Selection evaluates a subset of feature as a group of suitability. Feature Selection has divided into three methods. They are filter method, wrapper method and embedded method. Filter method evaluate each feature by scoring rank. Wrapper method uses the search algorithm to search through space of possible features and evaluate them by running the algorithm on the subset. Embedded techniques are embedded in and specific to a model.

Feature selection is used to select the best subset of features in the dataset. The notion of 'best' is related to attain the highest accuracy of the predictive model. The benefit of feature selection is it reduces overfitting. It means giving less opportunity to make decisions for the less redundant data based on noise. Feature selection improves the accuracy of the model and it reduces the training time. The algorithms train faster when using less data. It can greatly improve the accuracy of the classifier model. Therefore, it is important in finding the relevant features.

In this work, we will attempt the medical diagnosis for the disease liver disorder. Here we are using Ant Colony Optimization (ACO) for the feature selection. Feature Selection is the process of selecting the relevant features in which they can increase the efficiency of the model. Feature selection plays an important role in training classifiers. Assuming that the original feature space contains D features, the goal of the feature selection process is to find an optimal subset that contains only d ($d \leq D$) features. Feature selection can reduce the computational complexity and improve the classification accuracy. Different techniques are used to improve the accuracy of the classifier by removing irrelevant and redundant features. These techniques are used in medical imaging and medical diagnosis etc. The feature subset obtained by the feature selection algorithms are given as the input to the classifier. With the aim of improving the accuracy and the effectiveness of the medical disorder, the four classification algorithms like random forest, multi layer perceptron, adaboost and bagging are used.

In order to check the accuracy of the classification algorithms before feature selection and after feature selection the results are provided. The training set and test set are combined to form the single training dataset as tenfold cross validation is used to train the random forest, multi layer perceptron, bagging and adaboost.

II. LITERATURE SURVEY

The following are few related works, which deals with feature selection and classification. In recent years, various feature selection based classification techniques have been proposed. Saifnalband adityasundar et al[1] proposed the methodology could provide an effective non-invasive diagnostic tool for knee joint disorders, in which they have used two techniques. They were apriori algorithm and genetic algorithm to select the features. Least square support vector machines and random forest were proposed as classifiers to evaluate the performance of FS techniques. Ujjwalmaulik et. al [2] proposed scheme can achieve significant empirical success and was biologically relevant for cancer diagnosis and drug discovery. They have used fuzzy preference based rough set method for feature selection and semisupervised support vector machines for cancer classification. J. DhaliaSweetlin et al[3] proposed the method diagnostic accuracy of pulmonary bronchitis from CT images of the lung. They used ACO based feature selection for selecting the features and support vector machine for classification. Syed Muhammad Saqlain shah [4] proposed the technique that extracts reduced dimensional feature subset through probabilistic principal principal component. They used probabilistic principal component analysis and for the classification, they have used the support vector machine and radial basis function. Shamsul Huda et al[5]proposed a hybrid feature selection methodology with ensemble classification for the diagnosis of brain tumour. They proposed the globally optimized artificial neural network input gain measurement approximation (GANNIGMA) for the filter approach and for the wrapper approach artificial neural network input gain measurement approximation (ANNIGMA) with ensemble classification. Swati Shilaskar et al[6] implemented forward feature inclusion, back-elimination feature selection and forward feature selection with SVM classifier. Hybrid forward feature selection algorithm successfully reduces feature dimensions and improves accuracy of classifier. P. Jaganathan et al [7] proposed a filter-based feature subset selection based on fuzzy entropy measures and presented the different selection strategies for handling medical database classification. Javier Perez-Rodriguez et al [8] proposed the boosting feature subset selection algorithms. They have proposed the algorithms like adaboost feature selection algorithms, floatboost feature selection algorithm, multiboost feature selection algorithm, ReweightBoost feature selection algorithm. Xiao-Ying Liu et al [9] proposed the hybrid genetic algorithm for feature selection. They have used the hybrid genetic algorithm with wrapper- embedded approaches for feature selection. T. Vivekanandan et al [10] proposed a modified evolution algorithm for feature selection. This work is integrated with fuzzy analytic hierarchy process and feed forward neural network. The results showed that the modified differential evolution outperforms the traditional differential evolution in terms of execution time and minimization of critical attributes.

III. PROPOSED WORK

A. Feature Selection

The proposed work is selecting relevant features based on ACO. The ability of the real ants to find the shortest distance between the nest and food sources. Finding the shortest route is mainly due to the depositing the pheromone when they travel. Every ant follows the direction, which is rich in pheromone. The pheromone decays over time, resulting in less pheromone on less popular paths. The shortest route have higher pheromone level and the other routes have less pheromone level and it will diminished until all other ants follow the same route. Here all the routes are considered as the features and the routes, which are having high level of pheromones, are considered as the selected feature.

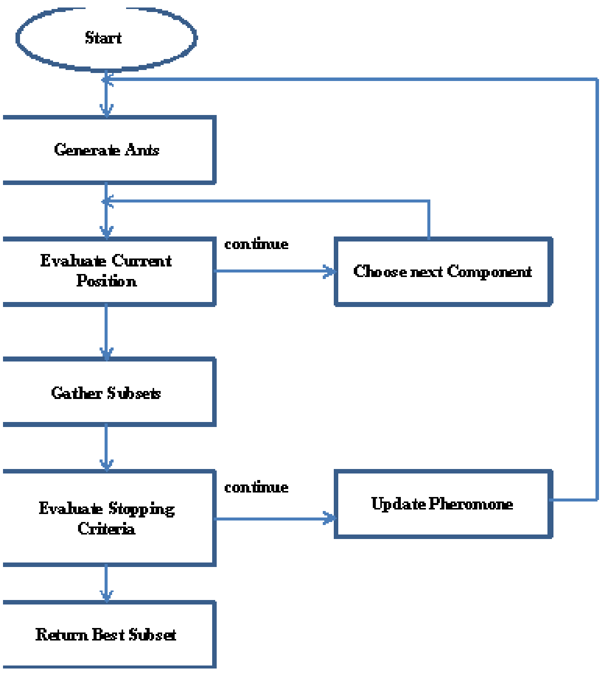

Figure 1. ACO based Feature selection

B. Algorithm for Ant Colony Optimization:

The feature selection process can be represented by:

- n features in the original set
- amount of pheromone levels τ
- Subset S
- m is the subset of feature (n>m)

Step 1) Initialise the number of ants k that must be equal to the number of features n

Step 2) Initialise the amount of pheromone levels τ

Step 3) Determination of subset is given by :

$$P_i^k(t) = \begin{cases} \frac{[\tau_i(t)]^{\alpha} * [\eta_i(t)]^{\beta}}{\sum_{u \in j^k} [\tau_u(t)]^{\alpha} * [\eta_u(t)]^{\beta}} & if\ i\ \in j^k \\ 0 & otherwise \end{cases} \quad (1)$$

Where, j^k is the set of feasible features that can be added to the partial solution. τ_i and η_i are the pheromone value and heuristic value associated with feature i. α and β are the parameter of their relative importance. Initially all features have equal value of τ and η.

Step 4) Check the construction progresses of S to determine whether it is finished or not. If finished by all ants k then continue, otherwise, go to Step 3).

Step 5) Evaluate the selected subset

For j=1 to n

Estimate the Mean Square Error (MSE) of the classifying results obtaining the classifying features. Sort the subsets according to their MSE. Update the minimum MSE and store the corresponding subset of features

Step 6) Check whether the algorithm has executed certain number of iterations (t) or not

Step 7) Update the pheromone value and heuristic value of each feature i

Step 8) If the number of iterations is less than the maximum iterations or the desired MSE is not achieved go to step 3.

C. Classification:

The selected feature subsets are used to train the classifier in weka independently. Weka is software, which contains a collection of machine learning algorithms. The training is carried out using the 10 fold cross validation. The selected feature subsets are trained using the classifier in weka: Random Forest, Multi Layer Perceptron, Adaboost, and Bagging.

1) Random forest: Random forest algorithm used for both classification and regression task. The difference between the decision tree and the random forest is that the splitting the node and finding the root node randomly. The advantages of random forest are it can be used in categorical values; it can handle the missing values. It is the ensemble learning method for classification. It is operated by constructing a mass of decision trees at training time and outputting the class of individual trees.

2) Multi Layer Perceptron: It is class of feed forward artificial neural neutron. It has three or more layers of nodes. Each node is a neuron and it uses nonlinear activation function. Activation function is output of that node given the set of input. It uses back propagation for training and supervised learning method. It consists of two passes; they are forward pass and the backward pass. Forward pass it predicts the output according to the input but in backward pass is partial derivation of the cost function with different parameters are propagated back through the network. Multi layer perceptrons are fully connected, each node in one layer has the certain weight to every node in the following layer.

3) Adaboost: Adaptive boosting is also called the adaboost. It can be used with any kind of conjunction with other algorithms. The output of the other learning algorithm is combined into weighted sum that gives the output of the boosted classifier. It can be represented in the form:

$$F_T(x) = \sum_{t=1}^{T} f_t(x)$$

Where $f_{(t)}$ is the weak learn that takes x as the input and return its class. It is used to boost the performance of the machine learning algorithms. These are used to achieve the accuracy of the model.

4) Bagging: Bagging is also called the bootstrap aggregation is the ensemble algorithm. It is designed to improve the accuracy and stability of the learning algorithm. Bagging is to have various classifiers prepared on various under-sampled subsets and enable these classifiers to vote on a ultimate decision, standing out from simply utilizing one classifier. Boosting is to have a progression of classifiers to prepare on the dataset, yet step by step putting more accentuation on training examples that the past classifiers have fizzled in the expectation of that the following classifier will focus around these harder examples.

IV. EXPERIMENTAL RESULTS

This section presents the description of the dataset for the disease liver disorder, implementation and the results obtained. The dataset is taken from UCI Machine Learning Repository. This dataset contains 10 features that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. This dataset contains data of 416 liver patient and 167 non-liver patients. It is collected from the northeast part of Andhra Pradesh, India. The class label is used to separate the group into liver patient or non-liver patient. It contains 441 male patients and 142 female patients.

A. Feature Selection using dataset

Feature selection for liver disorder dataset is implemented using ACO. Out of the ten features, four features were selected. In order to develop the most significant and accurate model, the relevant features must be selected.

Table 1. Dataset Description

Sl.No	Attribute	Description
1.	Age	Age of the patient
2.	Gender	Gender of the patient
3.	TB	Total Bilirubin
4.	DB	Direct Bilirubin
5.	Alkphos	Alkaline Phosphotase
6.	Sgpt	Alamine Aminotransferase
7.	Sgot	Aspartate Aminotransferase
8.	TP	Total Protiens
9.	ALB	Albumin
10.	A/G	Ratio Albumin and Globulin Ratio

B. Accuracy attained before and after FS:

The classification is carried out using weka toolkit, which has the numerous collection of in built machine learning algorithms. Here we have compared the accuracy of the classification algorithm using feature selection ACO in terms of before feature selection and after feature selection. After training the dataset to the classification algorithms we have found that accuracy of the classification algorithm has been increased in compared with before feature selection. We have provided the comparison table for classification algorithm, before and after doing feature selection using ACO. We trained dataset in some of the classification algorithm in Weka.

Table 2. Accuracy attained before and after FS

Classification Algorithm	Before FS	After FS
Random Forest	71.157	72.1934
Multi Layer Perceptron	68.04	69.948
Adaboost	70.9845	71.5026
Bagging	70.8117	71.3299

In the above table, it has shown that feature selection based on ACO has increased the accuracy of the classification algorithm. In all above algorithm the accuracy is increased in compared to classification algorithm before feature selection. In comparison with all the four classification algorithms, random forest classification algorithms achieved highest accuracy.

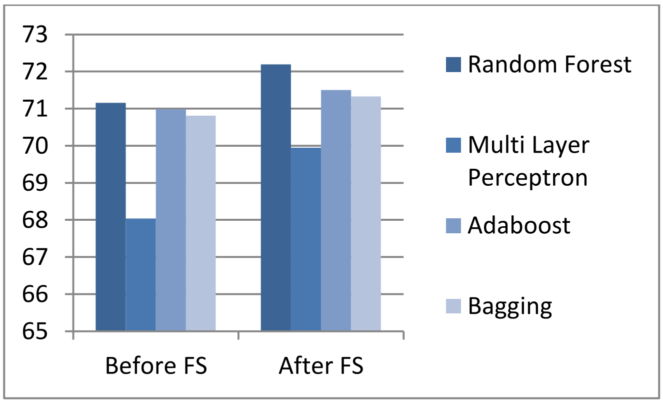

Figure 2. Comparing the accuracy (%) attained before and after feature selection

Figure 2 compares the results between four Random forest, Multilayer Perceptron, Adaboost and Bagging. The results verified that the performance of four classifiers before and after feature selection has increased in terms of accuracy. In the all four classification algorithms, random forest has highest accuracy in compare with other three algorithms.

C. Performance Evaluation

The effectiveness of the classifier performance can be evaluated using different parameters. They are specificity, precision, sensitivity/recall and accuracy. Specificity is the capacity of the test to correctly recognize those without the disease (true negative rate). Precision is the ratio of correctly assigned samples to the total number of samples classified. Recall is the ratio of correctly assigned samples to the total number of samples actually. Accuracy is defined as the percentage of samples classified correctly.

Specificity = TN/(TN+FP)
Precision = TP/(TP+FP)
Sensitivity/Recall = TP/(TP+FN)
Accuracy = no. of samples correctly classified/ total no. of samples taken.

Where TP is actual true positives correctly recognized by the system, FP is actual negatives but identified as positives, TN is actual true negatives and correctly identified as negatives, FN is actual positives but identified as negatives

V. CONCLUSION

In this paper ACO, based feature selection technique has been proposed. This technique could provide a accurate diagnostic system for medical diagnosis like liver disorders. In order to discard irrelevant and redundant features, feature selection algorithms are used. Feature selection algorithms are used to select relevant, stable and most significant feature. The performance of this ACO based feature selection algorithm is evaluated using Random Forest, Multi Layer Perceptron, Adaboost and Bagging classifiers. The results verified that ACO based feature selection algorithm has increased accuracy. We also compared the results with before feature selection and after feature selection. This study has confirmed that this ACO based feature selection has increased the accuracy in above said classification algorithms

VI. REFERENCES

[1]. Nalband S, Sundar A, Prince AA, Agarwal A. Feature selection and classification methodology for the detection of knee-joint disorders. Computer methods and programs in biomedicine. 2016 Apr 1;127:94-104.

[2]. Maulik U, Chakraborty D. Fuzzy preference based feature selection and semisupervised SVM for cancer classification. IEEE transactions on nanobioscience. 2014 Jun; 13(2):152-60.

[3]. Sweetlin JD, Nehemiah HK, Kannan A. Feature selection using ant colony optimization with tandem-run recruitment to diagnose bronchitis from CT scan images. Computer methods and programs in biomedicine. 2017 Jul 1;145:115-25.

[4]. Shah SM, Batool S, Khan I, Ashraf MU, Abbas SH, Hussain SA. Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis. Physica A: Statistical Mechanics and its Applications. 2017 Sep 15;482:796-807

[5]. Huda, Shamsul, et al. "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis." IEEE Access 4 (2016): 9145-9154.

[6]. Shilaskar, Swati, and Ashok Ghatol. "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases." Expert Systems with Applications 40.10 (2013): 4146-4153.

[7]. Jaganathan, P., and R. Kuppuchamy. "A threshold fuzzy entropy based feature selection for medical database classification." Computers in Biology and Medicine 43.12 (2013): 2222-2229.

[8]. Perez Rodriguez J, de Haro-Garcia A, del Castillo JA, Garcia-Pedrajas N. A general framework for boosting feature subset selection algorithms. Information Fusion. 2018 Mar 23.

[9]. Xiao-Ying Liu, Yong Liang, Sai Wang, Zi-Yi Yang, and Han-Shuo Ye. A Hybrid Genetic Algorithm with Wrapper-embedded approaches for feature selection, 2018.

[10]. Vivekanandan T, Iyengar NC. Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. Computers in biology and medicine. 2017 Nov 1;90:125-36.