Indian Journal of Science and Technology, Vol 8(35), DOI: 10.17485/ijst/2015/v8i35/86788, December 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Optimal Feature Subset Selection using Ant Colony Optimization

Sabeena S* and **B. Sarojini**

Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore - 641043, Tamil Nadu, India; sabeena.mphil@gmail.com, dr.b.sarojini@gmail.com

Abstract

Background/Objectives: Data mining is the process of extracting large volumes of raw data from hidden knowledge. The health care industry requires the use of data mining techniques as it generates huge and complex volumes of data. The applications of data mining techniques to medical data extract patterns which are useful for diagnosis, prognoses and treatment of diseases. This extraction of patterns allows doctors and hospitals to be more effective and more efficient. The huge volume of data is the barrier in the detection of patterns. Feature selection techniques mainly used in data preprocessing for data mining. **Methods/Statistical Analysis:** Classification task leads to reduction of the dimensionality of feature space, feature selection process is used for selecting large set of features. The Ant Colony Optimization based feature selection method is applied on cancer datasets. **Findings:** This research work proposes about feature selection mechanism based on Ant Colony Optimization. In an ACO algorithm, the activities of ants have significance for solving different combinatorial optimization problem which selects most relevant features. Through several iterations filter based method finds the optimal feature subset. Based on the similarity between features the feature relevance will be computed, that shows to the minimization of the redundancy. To validate the proposed feature selection method Support Vector Machine classification is applied. The accuracy of classification for whole feature set and the reduced feature subset are compared. The improved accuracy proves that the proposed feature selection approach has selected informative feature of the cancer datasets. **Applications/Improvements:** The possibilities of using PSO algorithm is applied for finding the best features in future. Other algorithms are also considered for further implementation.

Keywords: Ant Colony Optimization, Feature Selection, Support Vector Machine

1. Introduction

The most important and frequently used technique in data mining is feature selection which is used to data preprocessing. Functions like advances in data mining algorithm and developing mining performance which brings the immediate effect for feature selection[1]. Feature selection has been applied to a lot of areas namely face recognition, text categorization, finance and customer relationship management and cancer classification[2]. In the feature selection technique the data includes many unnecessary features. Redundant features provide more information rather than selected features, and irrelevant features provide no useful information in any situation[3]. Feature selection process is used for choosing a subset of features from a huge set of features to reduce the extensity of feature space for a strong classification task[4].

Ant Colony Optimization is a combinatorial optimization problem which selects most relevant features from the whole set of features[5]. By selecting a feature subset from the whole feature set improves the performance of classification algorithms. In this proposed work ant colony optimization is used as feature selection method to select features. Ant Colony Optimization is a distributed method in which a set of agents cooperate to find a good solution[6]. This is a filter based method which finds the optimal feature subset through several iterations. To reduce the redundancy among

* Author for correspondence

selected features multivariate approach is used[7,8]. SVM classification algorithm is applied to the resultant optimal feature subset obtained using ACO. The accuracy of classification for whole feature set and the reduced feature subset are compared. The increase in the performance of the classifier for the reduced feature subset validates the proposed feature selection approach[9].

This research work is organized as follows. The section 2 describes the methodology and the ACO algorithm. The experimental results and the comparison of different criteria are presented in section 3. Section 4 concludes this research work.

2. Methodology

The father of the Ant Colony Optimization was Dorigo and Gambardella who invented it in the early 1990's. Hard combinatorial optimization problems can be solved using Ant Colony Optimization (ACO). It is started by social habits of ants while searching for food. By solving hard combinatorial optimization problems cooperative work is carried out by each by each ant. It deposits a chemical substance, to communicate with other ants called pheromone. The intensity of the pheromone decreases due to the substance which evaporates over time[6]. To avoid trapping in a local minimum, to traverse new zone of the search area, and to decrease the possibility of selecting longer paths this process is being carried out[8]. While choosing two paths, ants prefer in higher possibility to select a path with more pheromone. The ants that select the shorter paths will get the food to the nest faster. Hence, shorter ways receive a higher number of pheromone than longer ways. Over time, all ants will be using the shorter ways to find the food. As a result, "evaporation of pheromone" and "probabilistic selection of paths", allow ants to find the shortest way, and they guide to flexibility for solving combinational optimization problems[9].

2.1 Feature Selection using ACO

Feature selection is the process of finding a subset of features, from the original feature set, optimally. Ant Colony Optimization method explores to find the optimal feature subset using some iterations. The main objective of the proposed method is to minimize the redundancy between them by selecting a subset of features. In this method lowest similarity features is selected by each ant to the previous selected features. Therefore, if a feature is selected by most of the ants, this indicates that the feature has the lowest similarity to the other features. The feature receives the greatest amount of pheromone, and the chances of its selection by the other ants will be increased in the next iterations. Finally, by using the similarity between features the top selected features will have high pheromone values. Thus, the proposed method selects the best features with minimum redundancy[6].

The feature relevance leads to the minimization of the redundancy which will be calculated based on the similarity within the features. The steps to perform ACO feature selection is described below. In this technique, before the feature selection method starts, the search space must be expressed as a suitable form for ACO. Therefore, the search space is expressed as a fully connected undirected weighted graph, $G = < F , E >$ where $F = \{F_1, F_2, \ldots, F_n\}$ indicates a set of all features in that each feature denotes a node in the graph, $E = \{(F_i, F_j): F_i, F_j \in F\}$ indicates the graph boundary. The connection of the boundary $(F_i, F_j) \in E$ will be set to the correlation value between F_i *and* F_j. Figure 1 shows the illustration of the feature selection problem.

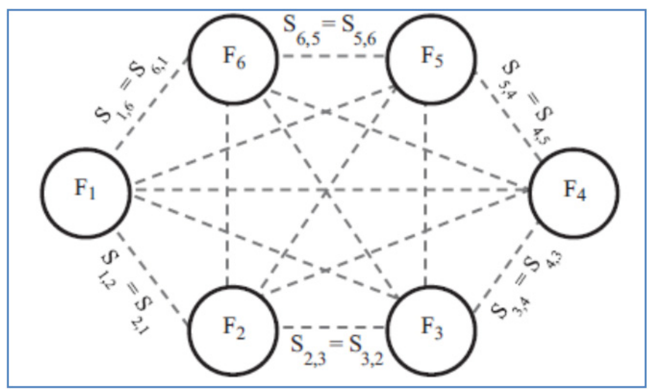

Figure 1. The representation of the feature selection problem.

The structure of algorithm is as follows:

Step 1: The purpose of a similarity related with the boundary between features i and j. The similarity between any two features is identified by calculating the perfect value of the cosine similarity among them. The cosine similarity between features A and B is calculated using the following equation.

$$sim(A,B) = \left| \frac{\sum_i^p = 1(a_i b_i)}{\left(\sqrt{\sum_i^p = 1^{a_i^2}}\right)\left(\sqrt{\sum_i^p = 1^{b_i^2}}\right)} \right| \quad (1)$$

Here A and B shows two features with p-dimensional vector ($A=\{a_1$, a_2 ,..., $a_p\}$ $B=\{b_1$, b_2 ,..,bp}).The similarity value of two features represented as 0 and 1 where 1 indicates similar features and 0 represents non-similar features. If the calculated similarity value between two features is greater than zero the features are similar.

By using ACO method in the feature selection problem, "heuristic data" and "desirability" must be defined by ACO algorithm. In the proposed method, the heuristic data is described as the contrary of the similarity within the features. A desirability measure τ_i $\forall i = 1\ldots n$, ecalled "pheromone", which is related to the features and will be refreshed by ants regularly.

Step 2: The proposed method is composed of many iterations. The group of pheromone allocated to each node before iterations start. In every iteration, *NAnt* ants are fixed randomly on the various nodes. Continuously according to a probabilistic "state transition rule" each ant traverses the nodes in a repetitive way until an iteration stopping rule is satisfied. Stopping rule is described as the number times the nodes that should be selected by each ant.

Step 3: The state transition rule aims to select features using highest pheromone values and smallest similarities to previously selected features. Feature counter array stores the features that is been selected by any ant.

Step 4: Then, at the end of the iteration, by applying a "global updating rule" the number of pheromone for each and every node is updated. Based on its feature counter value the number of pheromone for every node is computed. The ants tend to provide more pheromone to nodes with higher feature counter values. As well as, a portion of the pheromone melts on all nodes.

Step 5: The task is repeated till a given amount of iterations are achieved. Next, the features are grouped of their pheromone values in lower order. Then, the top selected features with highest pheromone values are selected as the final feature subset.

The "state transition rule" is based on a sequence of the heuristic data and the vertex pheromone values as shown below:

When the ant k is placed on the feature i, the upcoming feature j can be choosed by a greedy method or in a probabilistic method. In the greedy method, the upcoming feature is selected based on the following equation:

$$j = \underset{u \in j_i^k}{\arg} \max\left\{[\tau_u][\eta(F_i, F_u)]^{\beta}\right\}, if\ q \le q_0 \tag{2}$$

where j_i^k is the unvisited feature set, τ_u is the pheromone assigned to the feature u, $\eta\ (F_i\ ,F_u) = \frac{1}{sim}(F_i, F_u)$ is the inverse of the similarity between features i and u, β is a parameter that is used in the substance of pheromone versus similarity (β > 0), $q_0 \in [0,1]$ is a constant parameter, and q is a random value in the meantime [0,1].

In the probabilistic method, the upcoming feature j will be selected based on the probability $P_k\ (i,j)$ which is described as follows:

$$P_k(i,j) = \frac{[\tau_j][\eta(F_i, F_j)]^{\beta}}{\sum_{u \in j_i^k}[\tau_u][\eta(F_i, F_j)]^{\beta}}, if\ j \in J_i^k\ if\ q > q_0 \tag{3}$$

State transition rule based on the parameters q *and* q_0 which is a deals between Exploitation and Exploration. If $q \le q_0$, then ants choose the perfect feature in the greedy way, or else, each feature has a possibility of being selected according to its probability value that is calculated using exploration. The purpose of the probabilistic method is to avoid being captured inside a local optimum. The connection of both the probabilistic and the greedy methods are called "pseudo-random-proportional rule"[6,7].

The "global updating rule" is used to entire nodes at the end of the ant's traverse using the given equation:

$$\tau_i(t+1) = (1-q)\tau_i(t) + \frac{FC[i]}{\sum_j^n = 1FC[j]} \tag{4}$$

where n is the number of unique features, and are the total number of pheromone values of feature i at times t and t+1, thus, ρ is a pheromone evaporation parameter, and FC[i] is the counter according to feature i.

3. Experimental Results

Feature selection is the effective method is to select most suitable features from the whole feature set to raise the work of classification or clustering algorithms. Classification techniques are effective tool in order to classify the cancer data and measuring the accuracy. The effective classification tool is Support Vector Machine (SVM)[10]. In the proposed method, cancer datasets are processed more effectively using feature selection and classification. The feature selection method is based on Ant Colony Optimization and the classification method is Support Vector Machine (SVM). The ACO algorithm is implemented using Java Net Beans IDE 8.0.1. The SVM classification is done using Weka 3.6.

3.1 Parameter Settings

The proposed method applies several threshold values. It involves a different number of adjustable parameters. The beginning amount of pheromone for each feature is set to 0.2 ($\tau_i = 0.2$) pheromone evaporation coefficient is set to $\rho = 0.2$, the exploration control parameter is set ($q_0 = 0.7$), parameter β is set to 1 (β=1), and finally the number of ants for each dataset is set to the number of its original features (NAnt = #features).

3.2 Results of ACO based Feature Selection

Ant Colony Optimization based feature selection technique was used to select the features in breast cancer and liver cancer datasets. The results of feature Selection using Ant Colony Optimization in Breast and Liver Cancer datasets are shown in Figure 2. The experimental results show that the feature reduction of WDBC Breast cancer dataset is 55.55% and for BUPA Liver Disorder is 33.33%. Ant Colony Optimization based feature was applied to the breast and liver cancer datasets to find the optimal features from the original feature set. Breast cancer dataset attribute names denoted by {A0,A1,A2,A3,A4,A5,A6,A7,A8} and selected features are {A1,A2,A3,A8}. Liver cancer (BUPA Liver Disorder) dataset attributes names denoted by {A0,A1,A2,A3,A4, A5} and selected features are {A0,A1,A2,A5}.

Then, the whole features and selected features are applied in SVM classification technique. The percentage of feature reduction for WDBC dataset is 55.55% and for BUPA Liver Disorder is 33.33%. Support Vector Machine is used for classification. The classification is applied for the whole features and reduced feature subset. The accuracy of the SVM classification for whole feature set for WDBC dataset is 94.42% and for the reduced feature subset is 96.56%. There is a 2.12% of improvement in accuracy. The accuracy of the SVM classification for whole feature set for BUPA Liver Disorder dataset is 91.27% and for the reduced feature subset is 92.44%. There is a 1.17% of improvement in accuracy. The Table 1 represents SVM classification accuracy and Table 2 represents the feature reduction accuracy.

Table 1. SVM classification accuracy

Cancer Datasets	Whole set of features	SVM Accuracy for whole set of features	Selected Features	SVM Accuracy for selected features
Wisconsin Breast Cancer Dataset	9	94.42%	4	96.56%
BUPA Liver Disorder	6	91.27%	4	92.44%

Table 2. Percentage of feature reduction

Cancer Datasets	SVM Accuracy for whole set of features	SVM Accuracy for selected features	Percentage of feature reduction
Wisconsin Breast Cancer Dataset	94.42%	96.56%	55.55%
BUPA Liver Disorder	91.27%	92.44%	33.33%

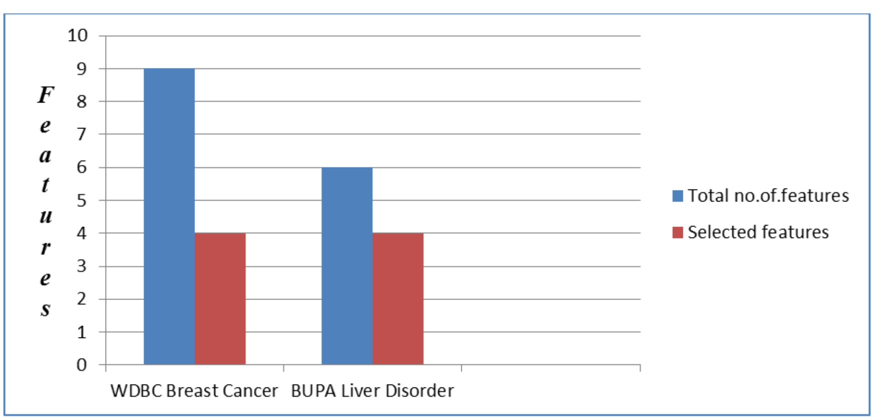

Figure 2. Ant Colony Optimization based Feature Selection.

The graphical representation of performance of SVM classifiers on WDBC and BUPA Liver Disorder Datasets are shown in Figure 3.

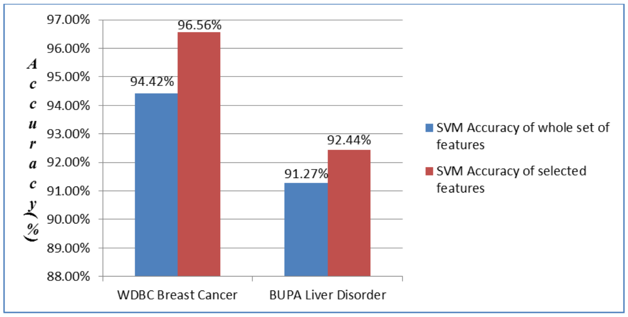

Figure 3. Performance of SVM classifier.

4. Conclusion

Cancer is considered as a dangerous disease caused mainly by environmental factors. There are different types of cancers, all types of cancer cells continue to grow, divide and re divide instead of dying and form new abnormal cells. An early prediction is needed in order to give detail treatment to the cancer patients and to help reduce the mortality and morbidity rate. The techniques used for this proposed methods are, feature selection method using ant colony optimization and support vector machine classification on medical datasets. The reduced feature subset and improved accuracy shows that diagnosis can be made with limited number of features. The selected input datasets are Breast cancer, and Liver cancer dataset. The proposed feature selection technique is ant colony optimization based feature selection, which is used to select a subset of features with minimum redundancy between them. Finally, the method selects best features with minimum redundancy. SVM classification is applied to the resultant feature subset and performance of the classifier is observed. Major goal of classification was to predict the target class accurately for each case in the data. The SVM classifier was applied to both selected features and whole set of features. The result obtained shows that the selected feature subset acquired the higher accuracy. The experiment results show that the implementation of feature selection process can generally improve the accuracy of the classifier.

5. References

1. Ouardighi A, Aboutajdine D. A powerful feature selection approach based on mutual information. International Journal of Computer Science and Network Security. 2008; 8(4):116–21.
2. Chen B, Chen L, Chen Y. Efficient ant colony optimization for image feature selection. Signal Process. 2013; 93(6):1566–76.
3. Meiri R, Zahavi J. Using simulated annealing to optimize the feature selection problem in marketing applications. European Journal of Operational Research. 2006; 171(3):842–58.
4. Biesiada J, Duch W. Feature selection for high-dimensional data: A pearson redundancy based filter. Computer Recognition Systems 2 Springer. 2007; 45:242–9.
5. Kanan HR, Faez K. An improved feature selection method using Ant Colony Optimization (ACO) classified on face recognition system. Applied Mathematics and Computation. 2008; 205: 716–25.
6. Tabakhi S, Moradi P, Akhlaghin F. An unsupervised feature selection algorithm based on ant colony optimization. Elsevier Journal. 2014; 32:112–23.
7. Dorigo M, Maniezzo V, Colorni A. Ant system: Optimization by a colony of cooperating agents. IEEE Transaction Systems. 2008; 26(1):29–41.
8. Aghdam, Aghaee G, Basiri. Text feature selection using ant colony optimization. Expert Systems with Applications. 2009; 36(3):6843–53.
9. Dorigo M, Gambardella M. Ant colony system - A cooperative learning approach to the traveling salesman problem. IEEE Transaction Evaluation. 2005; 1(1):53–6.
10. Unler A, Murat A, Chinnam RB. mr2PSO - A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. Information Science. 2011; 181(20):4625–41.