Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.463 – 468

RESEARCH ARTICLE

Selection of Most Responsible Genes for Cancer Disease from Large Attributed Dataset Using Hybrid Approach

Chandni Patel[1], Mahesh Panchal[2]
[1]Kalol Institute of Technology & Research Centre, India
[2]Deputy Director at Gujarat Technological University, India
[1] chandnip445@gmail.com; [2] mkhpanchal@gmail.com

Abstract*— High dimensionality has been a major problem for gene array-based cancer classification. Feature Selection (FS) is ordinarily used as a useful technique in order to reduce the dimension of the dataset. For that to get advantages from both methods of feature selection, Individual Feature Ranking (IFR) and Feature Subset Selection (FSS) are combined which is a hybrid approach. Information Gain from IFR and Ant Colony Optimization (ACO) from FSS are combined .To maintain the level of exploration and exploitation during search process in ACO, a Fraction method is used which is proposed in this paper.

***Keywords*— Feature Selection, Ant Colony Optimization, Level of exploration and exploitation in ACO**

I. INTRODUCTION

In many applications, like in medical science the size of a dataset is so large that learning might not work as well before removing unwanted features. Because in medical field a one sample of data has lots of features (i.e. million or billion) and from that sample to identify disease it is too critical. Same as for selecting most responsible or marker genes for Cancer diagnosis .Reducing the number of irrelevant/redundant genes from sample classifier accurately predicts the Cancer disease. This helps in getting a better insight into the underlying concept of a Cancer genes classification problem. Feature selection methods try to pick a subset of features that are relevant to the target concept. Feature selection is considerable importance in pattern classification, data analysis, information retrieval, machine learning and data mining application and achieved impressive classification results with classifier[2]. Feature selection can be seen as an optimization problem that involves searching the space of possible feature subsets to identify the optimal one. Many optimization techniques such as genetic algorithms (GAs) [1], tabu search (TS), simulated annealing (SA) and ant colony optimization algorithms (ACO) have been used for solving feature selection.

This paper introduces an ACO which is combining with IFR method and a Fraction method to maintain level of exploration and exploitation during search of optimal features.

The organization of the paper is as following; in section 2 Feature selection methods with their prone and cons have been explained. In section 3 ACO algorithms for feature selection is explained. The following section is Literature Review; in this section different algorithms have been explained with their limitation and after that proposed algorithm are introduced. Last section contains the conclusion about the study that is done so far and so forth future work.

II. FEATURE SELECTION METHODS FOR CLASSIFICATION

❖ **Feature Selection for Classification**: Feature selection can be defined as a process that chooses a minimum subset of M features from the original set of *N* features, so that the feature space is optimally reduced according to a certain evaluation criterion.Feature selection (FS) from the original set of features is highly desirable pre-process in order to remove any irrelevant or redundant features[1]. How feature selection is applied before classification is shown in Fig 1. It has desirable advantages when classification is performed on high dimensional data.

- Improving the classifier accuracy
- Providing faster and more cost-effective prediction
- Remove redundant and irrelevant features

❖ **Architecture[1]:** It is possible to derive a general architecture from most of the feature selection algorithms. There are four basic steps in a typical feature selection method[1] (see Fig 2)

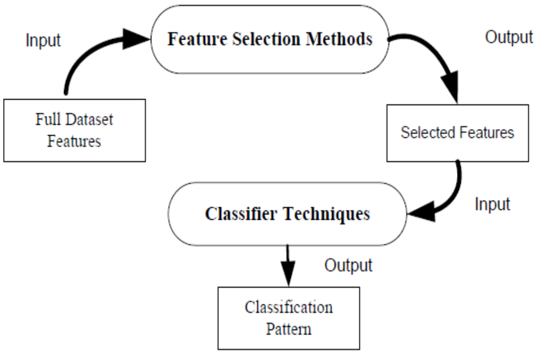

Fig 1 Feature selection for classification[1]

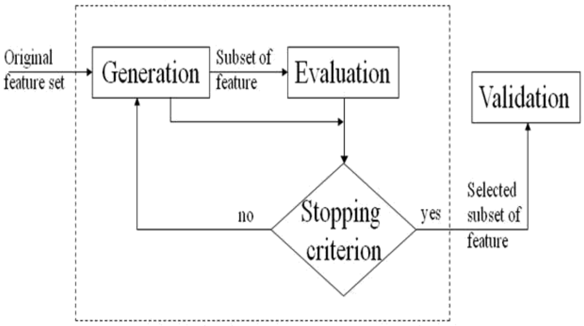

Fig 2 Architecture of Feature Selection[1]

1. **A generation procedure** to generate the next candidate subset of feature for evaluation
2. **An evaluation function** to evaluate the subset under examination. It measures the goodness of a subset produced by some generation procedure, and this value is compared with the previous best. If it is found to be better, then it replaces the previous best subset.
3. **A stopping criterion** to decide when to stop. Stopping criteria based on a generation procedure include: (i) whether a predefined number of features are selected, and (ii) whether a predefined number of iterations reached. Stopping criteria based on an evaluation function can be: (i) whether addition (or deletion) of any feature does not produce a better subset; and (ii) whether an optimal subset according to some evaluation function is obtained.
4. **A validation procedure** to check whether the subset is valid. It tries to test the validity of the selected subset by carrying out different tests, and comparing the results with previously established results.

❖ **Methods:** During generation and evaluation of features, Feature Selection has several methods or approaches with their examples[6].(See Table1)

Feature Generation	Individual Feature Ranking(IFR)	Most filters
	Feature Subset Selection(FSS)	GA,TS,SA,ACO
Feature Evaluation	Filter	mRmR
	Wrapper	Any Classifier(e.g. Naïve Bayes)

Table 1 Methods for Feature Selection

1. **Individual Feature Ranking[6]**: IFR measures each feature's relevance to the class and selects the top-ranked ones. Most filters like Information Gain(IG),Gain Ratio(GR),Chie Square[3] etc. The purpose of this technique is to discard irrelevant or redundant features from a given feature vector.

IFR is commonly used due to its simplicity and scalability but it has several disadvantages. Like The number of features retained is difficult to determine and lack of feature dependencies and classifier interaction[1].

2. **Feature Subset Selection[6]**: Attempts to find a set of features with good group performance. For finding a feature of subset it uses two approaches to search.

- Complete Search which is called as exhaustive search. Ideally, feature selection should exhaustively traverse all candidate subsets to find the optimal one. However, exhaustive search is known to be *NP*-hard and it becomes quickly computationally intractable.
- Heuristic Search which uses either Deterministic Search which uses greedy strategy to select according local change (such as such as SFS, SBS, SFFS, and SFBS) or Non Deterministic Search which attempts to find optimal solution in random fashion (such as GA, TS, ACO, SA and LVF).In recent years, Non Deterministic Search has been introduced to feature selection and have shown good performance.

FSS finds optimal Solution better than IFR but it become NP-Hard problem in terms of exhaustively traverse all candidate subsets to find the optimal one, less scalable and slower compare to IFR.

3. **Filter[6]**: Filters select good features based on data intrinsic measures, such as distance (e.g. fisher criterion, test statics, relief), consistency (e.g. inconsistency rate), correlation (e.g. Pearson coefficient correlation, information gain) and mRmR(maximum relevance and minimum redundancy). They show the relevance of a feature to the target class. These criteria are independent of any inductive learning algorithm.

4. **Wrapper[6]**: It utilizes a learning algorithm "wrapped" in the feature selection process to score feature subsets according to the prediction accuracy. Wrappers often select features with higher accuracy. Any classifier is used to find accuracy of features.

III. ANT COLONY OPTIMIZATIO FOR FEATURE SELECTION

ACO is mostly used for feature selection problem to get optimal feature subset from large search space. It is called as ACO based FS (Feature Selection) algorithm. In those algorithms, a large dataset is given as input. In which pheromone values are associate with each attribute (features) of dataset[4]. After that algorithm search for the optimal feature subset which contains features with higher pheromone value and finally that feature subset is given to any classification algorithm to get good predictive accuracy. General steps of ACO based Feature Selection algorithm is as follows[5].

1. Initialized pheromone value to each feature of data set.
2. Construct a solution(feature subset) by each ant.
3. Pass feature subset to classifier for evaluation by each ant and receive its accuracy.
4. Update best solution and best ant based on accuracy of classifier.
5. Check stopping criteria.

➢ ACO algorithms use two factors for guiding the search process[4]. This guideline is based mainly on the random and probabilistic behaviors of ants while selecting features during subset construction.

- Pheromone Value: It is numerical values as a simulation for the pheromone that real ants deposit on their way to and from their nest.
- Heuristic Information: It is the priori information which is used for the selection of candidate values. There are two types of heuristic information used by ACO algorithms Static and Dynamic.

➢ ACO algorithms use two types of search approach to find optimal solution from given search space[5].

- Exploitation: Searching in one direction which is experienced by previous solution.
- Exploration: Preventing ants from converging common path. That common path attracts ants to follow in same direction as pervious ants have followed. So that different ants get different solutions which increase the probability of getting better solution.

In ACO, a set of new rule has been designed to get proper balance between exploration and exploitation. It is necessary to get efficient result from any ACO algorithm.

IV. LITERATURE REVIEW

A. Genetic Algorithm[6]

A GA is an optimal search method. GA has also been introduced to feature selection [34]. Given a full set of N genes, each subset is represented as a string of length N as [$g1g2 . . . gN$], where each element takes a Boolean value (0 or 1) to indicate whether a gene is selected or not. A GA seeks for the optimal solution by iteratively executing genetic operators to realize evolution. Based on the principle of "survival of the fittest," strings with higher fitness are more likely to be selected and assigned a number of copies into the mating pool. Next, crossovers and mutations are performed on parent string with their probability. Each string in the new population is evaluated based on the fitness.

- Limitation: It has very random behaviour so it may mislead the search process.

B. Tabu Search[6]

TS algorithm is a meta heuristic method that guides the search for the optimal solution making use of flexible memory, which exploits the search history. TS is based on the assumption that solutions with higher objective value have a higher probability of either leading to a near-optimal solution, or to a good solution in a fewer number of steps. In each iteration, a TS moves to the best admissible neighboring solution, either with the greatest improvement or the least deterioration. A tabu list records the reverse of the most recent T moves to avoid cycling. Tabu but satisfies the aspiration criterion, then it is picked and made the new solution. A Tabu list prevent search from returning also help guide the search to achieve the optimal solution more quickly.

- Limitation: No result is available for high dimensional data.

C. Ant Colony System

Ant colony system is considered one of the most successful ACO algorithms. In which Pheromone value that associate with feature subset and Heuristic information that represents the desirability of feature. Features that belong to good solutions will contain larger pheromone values. Each ant passes its feature subset to the classifier SVM and receives its accuracy. Update the best solution and the best ant based on the accuracy of the classifier.

- Limitation: Exploration and exploitation level are not maintained during search process.

V. THE PROPOSED ALGORITHM

In this proposed algorithm, there are some basic methods are used with my proposed Fraction method.

- Hybrid IFR and FSS to get advantages from both Feature generation method.
- A Fraction method to maintain level of exploration and exploitation during search.
- mRmR (Maximum Relevancy and Minimum Redundancy) is used for feature subset evaluation.
- ➢ **User defined value:**

  Pheromone_value P, No. of Iteration **i**, No. of Ants **k**
  Subset_size **s** ,Best_subset **BS[]** ,Current_subset **CS[]**

- ➢ **Steps of Proposed Algorithm**

```
1. Load database having N attributes
2. Calculate Information Gain/Gain Ratio value of each attribute
3. Select $M(M<N)$ top most attributes which has higher value than given threshold value $t$
4. Initialize the parameters of Ant Colony Optimization Algorithm
5. For each Iteration i
        For each Ant k
        If (k==1) then
            Create Current_subset (CS) by Random method AND
```

```
                    Evaluate CS by MRMR method AND
                    Update pheromone_value P, Find Avg_Of_P, BS=CS
                     k=k+1;
                Else
                    Create Current_subset (CS) by Fraction Method AND
                    Evaluate CS by MRMR method AND
                    If (CS is better than BS)
                            Update pheromone_value P, Find Avg_Of_P, BS=CS
                       Check Stopping Criteria;
                    Else
                        No Updating AND
                        Check stopping Criteria
     Update pheromone_value P by BEST ANT at each iteration
6.   Train the classifier using the resulting optimal Best_subset (BS).
7.   Measure the accuracy of classifier
```

➢ **Stopping Criteria:**

- A predefined number of iterations are reached
- An optimal feature subset according to the evaluation is obtained.
- No progression is made in result.

➢ **Fraction Method**: It selects features from dataset according to user defined fraction value.

Example: In ACO, a pheromone value is associated to each feature in dataset and finds the average of that pheromone value.

Let Avg_Pheromone_Value =0.4, Subset_size=70, Total No of features=400, Fraction_value=0.6(60%) then Fraction method selects 70 features from 400 as fallows

Part 1:60% of 70 =42 features whose pheromone value is greater than 0.4(Avg_Pheromone_Value)
Part 2:40% of 70 =28 features whose pheromone value is less than or equal to 0.4(Avg_Pheromone_Value)

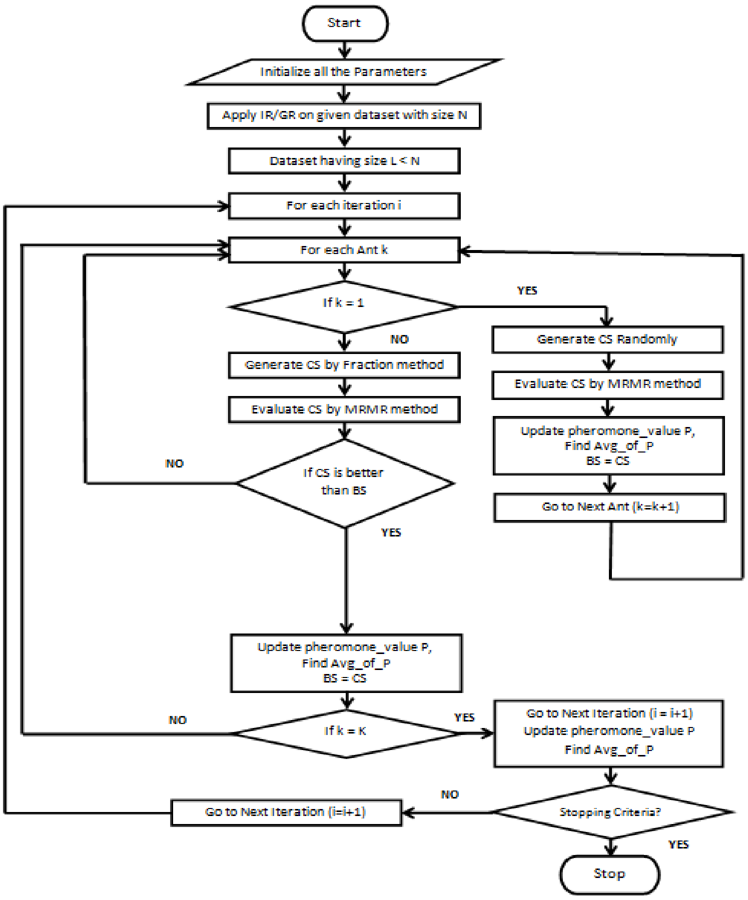
Start
Initialize all the Parameters
Apply IR/GR on given dataset with size N
Dataset having size L < N
For each iteration i
For each Ant k
If k = 1
YES
NO
Generate CS Randomly
Evaluate CS by MRMR method
Update pheromone_value P, Find Avg_of_P BS = CS
Go to Next Ant (k=k+1)
Generate CS by Fraction method
Evaluate CS by MRMR method
NO
If CS is better than BS
YES
Update pheromone_value P, Find Avg_of_P BS = CS
NO
If k = K
YES
Go to Next Iteration (i = i+1) Update pheromone_value P Find Avg_of_P
NO
Go to Next Iteration (i=i+1)
Stopping Criteria?
YES
Stop

Fig 3. Flow Chart of Proposed Algorithm

VI. CONCLUSION AND FUTURE WORK

In this paper, an implementation of proposed algorithm is done .In that two objective one is dealing with high dimensional data and proper guidelines to search process are achieved. After getting optimal feature subset which is given to classifier it will generate efficient performance with classifier in terms of accuracy. In future work this proposed algorithm is tested with different parameters values and dataset.

REFERENCES

[1] Norshafarina Binti Omar, Fatimatufaridah Binti Jusoh 2, Mohd Shahizan Bin Othman, Roliana Binti Ibrahim, "*Review of Feature Selection for Solving Classification Problems*" Journal Of Information Systems Research And Innovation, Issn: 2289.

[2] Huan Liu, Hiroshi Motoda , Rudy Setiono, Zheng Zhao "*Feature Selection: An Ever Evolving Frontier in Data Mining*"

[3] Jasmina Novaković, Perica Strbac, Dusan BULATOVIĆ "*Toward Optimal Feature Selection Using Ranking Methods*" Faculty *of Computer Science, Megatrend University, Serbia*.

[4] "Feature selection for classification using an ant colony system" e-Science 2010:Sixth IEEE international conference on e-Science. Brisbane, Australia .Dec. 2010.

[5] Monirul Kabir, Md Shahjahan and Kazuyuki Murase "*Ant Colony Optimization Toward Feature Selection*".

[6] Jiexun Li, Hua Su, Hsinchun Chen, "*Optimal Search-Based Gene Subset Selection for Gene Array Cancer Classification*" Fellow, IEEE, and Bernard W. Futscher.