# PSO Search-based Feature-selection Method for High Dimensional Data

**M. Sathya, S. Manju Priya**

***Abstract--- The growth of gene expression data from various techniques continues to expand. Addressing problems associated with high dimensional data and selecting relevant features have become more essential. Selection of relevant genes allows researchers to computationally explore gene expression to find functional genes, disease-causing genes and drug interactions to target specific genes. In the data mining community several feature-selection methods and techniques are continuously being studied and introduced. Selecting the best feature ranking method is still challenging. Feature-selection methods combine search methods and feature evaluation to find relevant features. The choice of search method has a significant relationship with feature ranking scores. In this paper, a new feature-selection method using PSO search strategy to derive high ranking feature subset is introduced. The extracted feature subset is experimentally studied on classification of Colon tumor using Colon dataset. The findings of the study show that PSO based search strategy shows better results than other methods. The study concludes that the proposed method can be used for high dimensional and classification problems on microarray dataset.***

***Keywords--- Microarray dataset, Feature-selection, Classification, Search Strategy, Feature Ranking, Particle Swarm Optimization, High Dimensional Problem.***

## I. INTRODUCTION

Gene selection is the process of selecting relevant genes that are expressional at different conditions. A microarray dataset is composed of genes in columns and samples in rows. The samples correspond to different conditions at molecular and cellular functional units. Microarray experiments are used to measure the expression levels of genes across different samples. Using the expressional genes, the characteristics of the expressions are predicted. The characteristics expression can be drug interaction, diseases, tumor or other biological properties. The prediction of the functional level is complex to address and analyze, since for a single functional outcome almost thousands of genes are involved. These genes may or may not contribute to the functional classes defined. In gene expression analysis, a large number of genes are generated from small number of samples. The generation of a large number of genes causes computational load and complexity to analyze. Gene-selection targets selection of appropriate genes that are relevant to the functional units. Feature-selection is the primary target for such high dimensional problems.

In recent years, several techniques have been introduced to address high dimensional problems. In microarray data, removing genes that are irrelevant or redundant reduces the high dimensions of the dataset. To reduce the feature genes, techniques such as information gain, relief, correlation-based feature-selection, chi-square-based feature-selection and many other methods are studied. The main challenge lies in selecting the best method to reduce the feature numbers. In a microarray gene expressional data, the feature-selection method has to address two main problems, namely reducing the number of features and keeping features unaffected from data processing. A feature-selection method can be opted for microarray data, only if it satisfies the two conditions. Also, the benefit of feature-selection on microarray data is that, while training a classifier, smaller samples of features avoid over-fitting and improves the classification performance.

Feature-selection method combines search methods and feature ranking. Ranking of features using different search methods has different rank scores for a feature and the choice of the search method is important for high dimensional problems, since the search method is primarily targeted to filter the features from thousands of genes. Different search methods have different strategies to explore the features and may cause computation complexity in the feature space. Exhaustive search, random search, breadth first search, heuristic search and non-exhaustive search, methods are extensively used to select features. Each search problem has its own merits and demerits in high dimensional problems.

Feature-selection methods are categorized into wrapper method, filter method and embedded methods. Wrapper uses search methods and classifiers to evaluate the feature relevancy, filter methods are independent of the classifiers and use statistical metrics such as distance, correlation, mutual information to score feature importance, while embedded method estimates feature ranks as a part of model building, where the method ranks the features and selects the features and builds models on the selected features. In this paper a new feature-selection algorithm is proposed using PSO (particle swarm optimization) search method for high dimensional data. This paper is organized into 5 sections. Section 1 discusses introduction to microarray data and gene selection, section 2 related works, section 3 methodology, section 4 experimental results and section 5 concludes the paper.

## II. RELATED WORKS

Ge, *et al.*(2016)proposed a wrapper based feature-selection to identify genes. The method used MIC and correlation between genes to measure the feature score. The selected features were classified using different classifiers to test the performance.

**Manuscript received February 01, 2019**

**M. Sathya,** Research Scholar, Dept. of Computer Science, Karpagam Academy of Higher Education, Coimbatore.

**Dr.S. Manju Priya,** Associate Professor, Dept. of CS, CA & IT, Karpagam Academy of Higher Education, Coimbatore.

This method outperformed PAM (Partition around Medodies), RF (Random forest) and CFS (correlation based feature-selection) methods with an improved accuracy.

Wang,*et al.*(2018) proposed a new feature-selection method that combined CFS and PSO to filter features. This method's efficiency was measured using six classifiers. The model improved classification accuracy of J48, RF, KNN (k-nearest neighbors), SVM (support vector machine) and MLP.

Jain,*et al.*(2018), using CFS and improved BPSO, proposed a new method for feature-selection. This method worked in two steps, in the initial step, genes were filtered using their correlation and in the second step, the selected genes were optimized using IBPO. When using the selected features the model showed100% performance accuracy.

Aziz,*et al.*(2016) proposed a feature-extraction technique using IC and fuzzy backward elimination. Using PCA, the performance of this method and that of the feature-selection method are compared. The genes that are extracted using this proposed method, improved the classification accuracy of SVM to 90% and NB to 85%.

Sahu,*et al.*(2017) proposed an unsupervised method to reduce the features in microarray data. The similarity between clusters was ranked through SNR, SAM and t-test, and the genes were filtered. The selected features were applied to DT (Decision tree), KNN, and MLP for performance. The study on four different microarray data showed that the ensemble model's performance accuracy improved greatly.

Wang,*et al.*(2017) proposed a wrapper-based feature-selection method using Markov blanker and SU to rank the features. Using SFS, the irrelevant features were removed. The quality of the subset was tested using C4.5, KNN and NB (Navie Bayes) across 10 different data sets. The model showed a higher performance with better accuracy.

Nagpal, *et al.*(2018) proposed a new method for selecting relevant genes using random forest and MI. In the initial step, the genes were filtered using random forest and the subset of features was derived using MI. Classifiers such as IB1, C4.5 and PNB were used to evaluate the quality of the gene subset. PNB achieved better classification results.

Gao,*et al.*(2018) proposed a new hybrid feature-selection method using minimum redundancy-Maximal new Classification (MR-MNCI). This method estimated the class dependence of features, and the features were selected. The resulting subset improved SVM and NB classifiers performance.

Shukla,*et al.*(2018) proposed a new method to filter genes using recursive PSO, and assigning SVM weights, the features are ranked. The highest-ranking features were filtered to a subset and compared against five datasets. This method achieved better accuracy of 98% than other methods.

Arunkumar,*et al.*(2018) introduced a new method based on rough set to select the best features. The rough set method used similarity score to select the genes. The selected genes were classified using RF model tenfold cross validation, and the performance of the model was compared on three different datasets. RF model achieved an accuracy of 98%.

Qi,*et al.*(2018) using unsupervised matrix factorization proposed a new feature-selection method. This method estimated the features' weights using correlation to remove irrelevant features. This method outperformed NSM, FNMF and MFFS methods with a mean accuracy of 55%.

Dashtban, *et al.*(2018) proposed a novel method for gene selection using Fisher score. Binary Bat algorithm was used to build a wrapper method using random walk search (MOBBALS). The subset of genes filtered using the proposed method was applied to NB, DT, KNN and SVM classifiers to check model performance. This model performance of 99% was obtained using this method.

## III. METHODOLOGY

The objective of the study is to propose a feature-selection model with the best search strategy. Since feature-selection method involves search method and feature ranking, different search methods use different search functionalities to generate a feature subset. A subset derived from two different search methods differs subsequently with respect to the feature selected. Feature ranking methods include Correlation, Chi-square, information gain, Relief-f, symmetrical uncertainty etc., while search method includes exhaustive search, best first, simulated annealing, genetic algorithm, greedy forward search, PSO, etc., which are extensively used in data mining and feature-selection. Search strategy work on the assumption that there is no correlation between features, but among features there exist a correlation. Utilizing the correlation between features, the search strategy can be improved to select features with more discriminating power between classes. The search methods used in the study are designed to utilize the correlation among features to derive a subset.

**Particle swarm optimization:** PSO is a Meta heuristic search method developed using bird's behavior in search of food. The idea behind PSO is the use of mutual information sharing among birds. When birds randomly search for food, they do not know the place where it is available. Here, place is referred to the solution (global optima) and birds are aware of their distance (local optima) from the food source (fitness). By adjusting their distance (velocity) between the source and their location, birds reach the food source. Using these criteria, the search strategy works to find the optimal solution. The new velocity and new position are given by formula 1 and 2

$$vi(k+1) = vi(t) + 1i(pi - xi(t)) + 2i(G - xi(t)) \text{-----} \mathbf{1}$$

$$xi(t+1) = xi(t) + vi(t+1) \text{---} \mathbf{2}$$

where, i is the $i^{th}$ particle in the vector, t the time index, vi the velocity of $i^{th}$ particle, xi the position of $i^{th}$ particle, pi the best position found by $i^{th}$ particle (local solution) and G the best (global) position found.

**Evolutionary Search:** Evolutionary algorithm is a simple process of population-based method derived from natural evolution. The natural process of evolution follows reproduction, mutation, combination and selection, and inspired from the natural process, the method optimizes the population using fitness function.

The computational process involves initialization, selection, combination and termination. Initialization involves deriving a population from available solutions. The selection step involves selection of fitness function from the characteristics of the members in the population and derive solutions. Based on the fitness, the members are chosen for the next generation through combinations. The combination produces newer generations that are engaged in replacing the members in the generation. The process is repeated until fitness criteria are met according to the threshold. Evolutionary algorithms are used for both minimization and maximization and are suitable for multi-objective problems. In feature-selection, each subset is a combination of generations that compete for the solution.

**Greedy Search:** Greedy search tries finding solution through finding local solutions. The solutions are selected with respect to local benefit at each step of the process. The algorithm finds solution for every step and summarizing the steps, the global solution is arrived at. The optimization of solution is achieved through the local optimum. When approaching a solution for maximization problems, the local solution at each step is used to construct a maximization solution. The feature subsets are derived by finding each feature that maximizes the performance of the model, and similarly for subset, a new combination of features is used to build the subset, comparing it with the previous one; if the new subset performs well, then the older subsets are dropped. The main drawback with this type of search is that it does not perform on all the features present in the dataset.

**Maximum Relevancy Minimum Redundancy:** mRMR is a selection process that aims to find relevant and redundant features and creates a subset. The method creates a subset using a strategy where features are correlated to the class (relevant) and there is poor correlation between features (redundant). The relevancy score is computed using f-score, and the redundancy is calculated using correlation coefficient. Depending upon the type of feature that is continuous or discrete, MI and correlation are applied. For discrete variables, mutual information score is used to compute the redundancy level of the feature. Features that are highly correlated with class are termed as relevant features and features that have high inter-correlation with features are termed redundant. The relevancy and redundancy formula are given by:

$$\text{Relevancy} = \max V_F, V_F = \frac{1}{|S|}\sum_{i \in S} F(i,h) \quad (1)$$

$$\text{Redundancy} = \min W_c, W_c = \frac{1}{|S|^2}\sum_{i,j} |cor(i,j)| \quad (2)$$

## IV. EXPERIMENTAL SETUP

To find out the best search strategy different search methods such as PSO, evolutionary search, mRMR and greedy search are used to guide the feature-selection process of correlation based feature-selection. To test the performance of the search strategy in feature-selection, Weka tool is selected for conducting experiments, and the tool provides a wide range of feature-selection methods and search strategies. To test the model performance, SVM and NB are used as the base classifiers.

*Data Set*

The experimental study uses Colon microarray dataset to study the model performance. The data set is obtained from Princeton university gene expression project. The dataset contains 2000 features of expression of genes and 62 samples. Out of the total samples 22 are normal and 40 are tumor samples. Initially the dataset is portioned into testing and training set in 70:30 ratio. The model is evaluated using the testing set. After partition, the testing set contains 19 samples (30%) and the training set 43 samples (70%).

*Evaluation Metrics*

Different search methods such as PSO, greedy search, evolutionary search and mRMR are used to select the features. The quality of the subset is evaluated using NB and SVM classification algorithms. The classification metrics are computed using true positives (TPs), false positives (FPs), false negatives (FNs) and true negatives (TNs) from the confusion matrix. Generally, there are four different metrics such as accuracy, sensitivity, precision and F1. Accuracy is the proportion of predictions that are true. Sensitivity is the proportion of positive instances classified as positive, whereas specificity is the proportion of negative instances classified as negative. To test the performance of the classifier, accuracy is used as the evaluation metric.

## V. RESULTS AND DISCUSSION

The performance of the search methods for generating subsets using PSO, greedy search, evolutionary search and mRMR is compared using NB and SVM classifiers. According to table 1, the subset generated by greedy search contains 10 features, PSO generated subset with 381 features, 44 features selected from evolutionary search and 24 from mRMR. Since Colon dataset contains 2000 features, the selection of features is computationally expensive, and the process of selecting features from microarray dataset consumes more time. The experiment is carried out using subsets derived from feature-selection. In gene expressional analysis, the feature numbers and also their expressional relevancy to the functional aspects of genes are to be considered, as traditional feature-selection only targets relevancy to classes. Out of 2000 features, 381 features filtered from PSO search can be helpful to rule out the relevancy of the genes with the functional aspects, while other methods filter much reduced features pertaining to class relevancy, and it becomes difficult to interpret the relevancy of the features to the functional aspects. Classifiers NB and SVM are used on the feature subset for classification of tumor and normal samples on testing set. According to Table 2, NB using PSO search achieves the highest accuracy of 62.63%, while greedy search and evolutionary search achieve an accuracy of 54.63% and 52.63%, and ranker search using mRMR achieves only 47.36%. SVM using PSO search achieves the highest accuracy of 81.94%, while on greedy search, and evolutionary search achieves an accuracy of 79.94% and

75.94, and using ranker mRMR search, SVM achieves78.94%. NB and SVM show the highest accuracy performance using PSO search with 381 features.

**Table 1: Feature Subset Generated**

| | CFS+ Greedy search | CFS+ mRMR | CFS+ Evolutionary Search | **CFS+ PSO Search** |
|---|---|---|---|---|
| Colon | 10 | 24 | 44 | **381** |

From the classification performance, high dimensional problem can be effectively managed through feature-selection, and feature-selection removes redundant features and improves classification performance. The difference in search methods and features in the subset largely affect the classification results. The study shows that performance of the models improves when using PSO search. On the other hand, when comparing the models, SVM shows better results than NB. The lowest accuracy of 47.36% of NB and 75.94% of SVM show that SVM classification is better than NB (Fig 1).

**Table 2: Accuracy of the proposed model**

| | CFS+ Greedy search | CFS+ mRMR | CFS+ Evolutionary Search | **CFS+ PSO Search (proposed)** |
|---|---|---|---|---|
| NB | 54.63% | 47.36% | 52.63% | **62.63%** |
| SVM | 79.94% | 78.94% | 75.94% | **81.94%** |

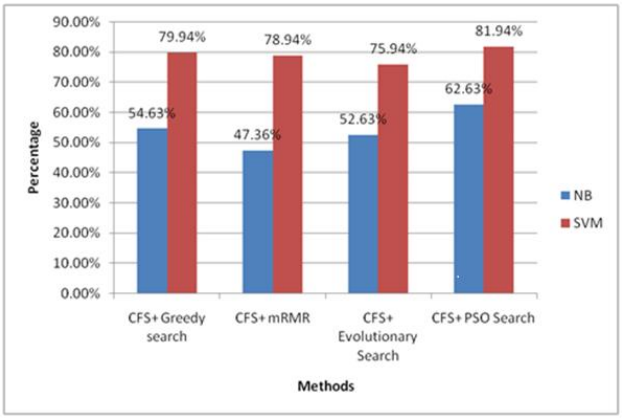


**Fig. 1: Accuracy of SVM and NB using PSO search**

## VI. CONCLUSION

Feature-selection is extensively required for high dimensional microarray datasets as microarray datasets are high dimensional. Gene expression analysis involves large samples and experiment conditions to study gene expressions. For efficient feature-selection, search method plays an important role as search process guides the feature-selection process. The study proposed four different search strategies such as PSO, greedy search, evolutionary search and mRMR to select relevant features and remove redundant features. The experimental study on Colon microarray dataset shows that the proposed PSO based search strategy shows better results on NB and SVM. The SVM model with PSO based on CFS, achieves the highest accuracy of 81.94% compared to other models. PSO based search method greatly reduces the feature numbers and thus PSO can be effectively used on microarray dataset for high dimensional problems. The study can be extended using a greater number of gene expressional microarray datasets to compare the performance across various dimensions. The performance of the proposed method focuses on high-featured numbers involving multi classes and provides an opportunity for utilizing the method for gene selection in expressional studies.

## REFERENCES


1. Arunkumar C, & Ramakrishnan, S. (2018). Attribute selection using fuzzy roughset based customized similarity measure for lung cancer microarray gene expression data. Future Computing and Informatics Journal.
2. Aziz R, Verma, C. K., & Srivastava, N. (2016). A fuzzy based feature-selection from independent component subspace for machine learning classification of microarray data. Genomics data, 8, 4-15.
3. Dashtban M., Balafar M., & Suravajhala, P. (2018). Gene selection for tumor classification using a novel bio-inspired multi-objective approach. Genomics, 110(1), 10-17.
4. Gao W., Hu, L., Zhang, P., & Wang, F. (2018). Feature-selection by integrating two groups of feature evaluation criteria. Expert Systems with Applications.
5. Ge R., Zhou, M., Luo Y., Meng, Q., Mai, G., Ma, D.,& Zhou, F. (2016). McTwo: a two-step feature-selection algorithm based on maximal information coefficient. BMC bioinformatics, 17(1), 142.
6. Jain I., Jain, V. K., & Jain R. (2018). Correlation feature-selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification. Applied Soft Computing, 62, 203-215.
7. Nagpal A., & Singh V. (2018). A Feature-selection Algorithm Based on Qualitative Mutual Information for Cancer Microarray Data. Procedia Computer Science, 132, 244-252.
8. Qi M., Wang T., Liu F., Zhang, B., Wang, J., & Yi, Y. (2018). Unsupervised feature-selectionby regularized matrix factorization. Neurocomputing, 273, 593-610.
9. Sahu B., Dehuri S., & Jagadev A. K. (2017). Feature-selection model based on clustering and ranking in pipeline for microarray data. Informatics in Medicine Unlocked, 9, 107-122.
10. Shukla A. K., Singh P., & Vardhan, M. (2018). A hybrid gene selection method for microarray recognition. Biocybernetics and Biomedical Engineering
11. Wang, A., An, N., Yang, J., Chen, G., Li, L., & Alterovitz, G. (2017). Wrapper-based gene selection with Markov blanket. Computers in biology and medicine, 81, 11-23.
12. Wang, H., Ke, R., Li, J., An, Y., Wang, K., & Yu, L. (2018). A correlation-based binary particle swarm optimization method for feature-selection in human activity recognition. International Journal of Distributed Sensor Networks, 14(4),
13. Sun, L., Zhang, X., Xu, J., Wang, W., & Liu, R. (2018). A Gene selection approach based on the fisher linear discriminant and the neighborhood rough set. Bioengineered, 9(1), 144-151.
14. Tabakhi, S., Najafi, A., Ranjbar, R., & Moradi, P. (2015). Gene selection for microarray data classification using a novel ant colony optimization. Neurocomputing, 168, 1024-1036.
15. Dai, J., & Xu, Q. (2013). Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. Applied Soft Computing, 13(1), 211-221.
16. Dash R., & Misra B. (2017). Gene selection and classification of microarray data: a Pareto DE approach. Intelligent Decision Technologies, 11(1), 93-107.