An Optimized Feature Selection Method for High Dimensional Data

J. Priyadharshini[1*], C. Kanimozhi[2]
[1]*PG Scholar, Department of Computer Science and Engineering, Anna University, BIT-Campus, Tiruchirappalli, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Anna University, BIT-Campus, Tiruchirappalli, India*
*Corresponding author: priyadharshini9524@gmail.com

***Abstract*: High dimensional datasets consists of large number of both relevant and irrelevant features, hence the computational and prediction time to process the dataset increases. Feature selection (FS) extracts the most relevant features which are known as subsets for prediction and the computational time can be reduced. The dataset is taken from National Centre of Biotechnology Information (NCBI), which is a widely used benchmark dataset for feature selection from Microarray Gene Expression. In gene expression data analysis, the problems of cancer classification and gene selections are closely related. Selecting informative genes is essential for classification performance. However, high dimensional dataset causes a high computational cost and over fitting during classification. Thus it is necessary to reduce the dimension of data by feature selection. In this paper mean based Genetic Algorithm (GA) is proposed to select the optimal subsets from the raw dataset based on the mean value of the features and the accuracy of the subset is evaluated using a classifier of Support vector machine (SVM), which reduces the complexity of the model in terms of computational cost and size. The proposed method is compared with the ant colony optimization (ACO) algorithm and the result shown that the proposed method has a better accuracy rate.**

***Keywords*: ACO, Accuracy, Feature selection, GA, Subset.**

1. Introduction

Data Analytics is the science of analyzing raw data to conclude the facts present in information. Many of the techniques and the process of data analytics have been automated into a mechanical process and algorithm that works over raw data for human consumption. Microarray gene expression technology has an enormous effect on cancer research. It is a powerful technique when it comes to diagnosing and identifying the disease genes for human cancers. Moreover, it has been vastly used to identify cancer-related genes using feature selection methods. A microarray expression dataset can be represented in a tabular form, in which each row represents a particular gene, each column to a sample and every entry of the matrix is a measured expression level of a particular gene in a sample. Dimensionality Reduction (DR) is a very important issue in the processing of high dimensional data. High-dimensional datasets contain unrelated and repeated features.

Feature selection is the most significant pre-processing step in mining large-dimensional data. Time complexity is very high for choosing the subset of features and for further analysis or to design the regression if the number of features and targets in the dataset is large. [9]

Feature selection (FS) is the process of significantly reducing the dimensionality of the feature space, while maintaining an accurate representation of the original data. Its main advantages are improved classification performance, reduced learning speeds, facilitating data interpretation, and improved generalization capability of the predictions.

This research work is organized as follows. The section 2 describes the related work. Section3 describes the Methodology. The proposed system and experimental results are presented in section 4 and Section 5. Section 6 concludes this research work.

2. Related Work

Rohana et al., [1] proposed a framework based on a genetic algorithm (GA) for feature subset selection and compared with various existing feature selection methods. It includes the ability to overcome multiple feature selection criteria and find small subsets of features that perform well for a particular inductive learning algorithm of interest to build the classifier.

Peng et al., [2] in general many of the swarm intelligent algorithms that simulate the social behavior of living beings are used as feature selection algorithms. The proposed method uses the one of the swarm intelligent algorithm for feature selection based on ant colony optimization. This algorithm is combined with the classifiers for selecting the more appropriate and useful features.

Raid Alzubi et al., [3] discussed an accurate hybrid feature selection method for detecting the most informative SNPs and selecting an optimal SNP subset. It is based on the fusion of a filter and a wrapper method, the Conditional Mutual Information Maximization method (CMIM), and the Support Vector Machine respectively (SVM).

Shuai An, et al., [4] discussed a new feature weighting approach which is based on the local nearest neighbors for gene selection, called local-nearest-neighbors-based feature weighting (LNNFW). LNNFW shrinks the distances between the target neighbors and magnifies the distances between the local differently labelled instances. This method can be applied naturally to the multi-class problems and does not require extra modification.

Jingyu Hou et al., [5] proposed a gene expression programming (GEP) based model to predict lung cancer from microarray data. It implements gene selection methods to extract the significant lung cancer-related genes and described about different GEP-based prediction models.

Mehrdad J et al., [6] proposes feature selection based on the Hilbert-Schmidt independence criterion (HSIC) and singular value decomposition (SVD). This algorithm is computationally fast and scalable to large datasets and only one row of the data has to be examined at a time. It does not require the whole microarray dataset to be stored in memory, and thus can easily be scaled to large datasets.

Cheng Liu et al., [7]., In this study, discussed a sparse logistic regression model with structured penalized regularization for feature selection in gene expression data, which can identify the unknown correlation structure within the data. Structured penalized logistic regression model can facilitate the selection of highly correlated genes, it demonstrates that the structured penalized regularization is capable of selecting relevant features and identifying important structure within the data and it achieves better prediction performance

Jian Tang et al., [8] describes a new MI-based feature selection approach for microarray data. This method depends on two strategy relevance boosting to displays the relevance for class labelling for the selected features and the other is feature interaction enhancing, which has simple aggregation based evaluation.

Jovani Taveira de Souza et al., [9] introduced two efficient methods for reduction methods. Attribute selection and Principle Component Analysis (AS and PCA) were applied to compare their performances in the selected databases. It demonstrates the consistency based subset evaluation and minimum redundancy and maximum relevance (CSE-mRMR) which represents excellent classification performance and presented better results.

Nada Almugren et al., [10] reviewed and compared the hybrid approaches that accomplished the bio-inspired evolutionary methods as the wrapper method. It stated that the classification is a challenging task due to the high dimensionality found in a small sample size of gene expression data. It concludes that the genetic algorithm GA is the most applied wrapper method and it achieves the highest accuracy with relatively small numbers of selected genes.

3. Methodology

A. *Genetic Algorithm*

Genetic algorithm is a part of Bio-inspired computing. Genetic algorithm is a search algorithm based on the principles of natural selection and genetics. In GAs, a population of chromosomes indicates candidate solutions for the problem. Each chromosome is represented with fixed-length bits. The primary population of chromosomes is created by distributing 1 s and 0 s arbitrarily.

B. *Role of Genetic Operators*

1) *Crossover*

Swapping parts of the solution with another in chromosomes or solution**.** The important role is to give combining the solutions and convergence in a subspace.

2) *Mutation*

Changing parts of one solution randomly to increases the diversity of the population. It provides a mechanism to escape from the local optimum.

3) *Selection of the fittest*

The use of the solutions with high fitness to pass to next a generation, which is often carried out in terms of some form of selection of the best solutions.

C. *Ant Colony Optimization*

Ant Colony Optimization is a combinatorial optimization problem which selects most relevant features from the whole set of features. 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. This is a filter based method which finds the optimal feature subset through several iterations.

4. Proposed System

A mean based Genetic algorithm, an optimization algorithm is applied on the dataset to select the optimal features of subset and the accuracy of the optimal subset is evaluated using the Support Vector Machine.

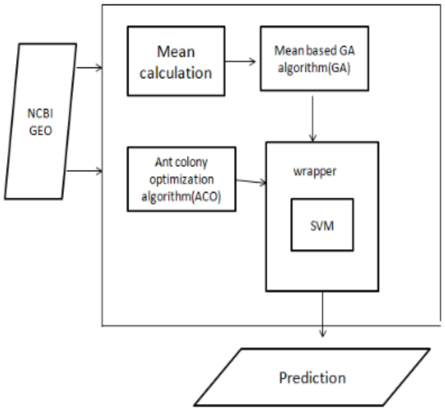

NCBI GEO
Mean calculation
Mean based GA algorithm(GA)
Ant colony optimization algorithm(ACO)
wrapper
SVM
Prediction

Fig. 1. Architecture diagram

A. *Data Collection*

Despite the different interesting applications of Microarray gene expression, the large number of datasets is available. The dataset is collected from the National Centre for Biotechnology Information (NCBI).

GDS3096: Tumour epithelium and surrounding stromal cells were isolated using laser capture micro dissection of human breast cancer to examine differences in gene 16 expression based on tissue types from inflammatory (IB) and non-inflammatory breast cancer (non-IBC).

B. Mean calculation

The mean value is calculated from the raw dataset by computing the column values of each feature and the calculated values are combined and stored as a separate feature subset. The subset acts like a separate data frame and it is given to the Genetic algorithm. The calculated mean values act as a criteria for selection of optimal features.

C. Feature Selection

The purpose of feature selection is to determine the minimum number of feature subsets that are essential and appropriate for the classifier to classify the normal and cancer cell genes. The feature subset p is always less than the original feature set m.

D. Accuracy estimation

The accuracy of the reduced subset is estimated using the classifier of Support Vector Machine. The main advantages of using SVM is that it is effective, while it also works well when using high dimensional data it works well when the number of features is greater than the number of samples.

5. Experimental Results

Feature selection is the effective method to select most suitable features from the whole feature set to raise the work of classification. Classification techniques are effective tool in order to classify the features and measuring the accuracy. The effective classification tool is Support Vector Machine (SVM). In the proposed method, datasets are processed more effectively using feature selection method. The feature selection method is based on mean based Genetic algorithm and Ant Colony Optimization. The classification method is Support Vector Machine (SVM). The proposed method was implemented using R 3.6 tool.

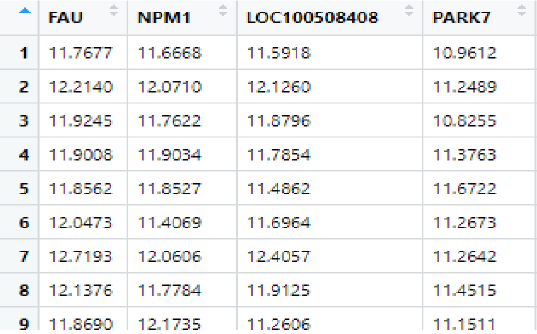
A. Parameter Settings

The proposed method applies several threshold values. It involves a different number of adjustable parameters. In mean based Genetic algorithm mutation and crossover probability is set to *pm*=0.001, *pc*=0.75.

Fit function: It takes two input arguments, a vector of indices into rows of the population matrix, and a context list within which any other items required by the function can be resolved. In ACO, pheromone evaporation coefficient is set to ρ = 0.2, the pheromone deposited by each ant is set (q = 0.7), iteration=5 and finally the number of ants population = 10.

B. Results

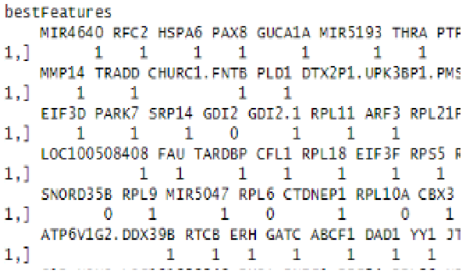
Mean based genetic algorithm and Ant Colony Optimization based feature selection techniques was used to select the features from GDS3096 dataset.



	FAU	NPM1	LOC100508408	PARK7
1	11.7677	11.6668	11.5918	10.9612
2	12.2140	12.0710	12.1260	11.2489
3	11.9245	11.7622	11.8796	10.8255
4	11.9008	11.9034	11.7854	11.3763
5	11.8562	11.8527	11.4862	11.6722
6	12.0473	11.4069	11.6964	11.2673
7	12.7193	12.0606	12.4057	11.2642
8	12.1376	11.7784	11.9125	11.4515
9	11.8690	12.1735	11.2606	11.1511

Fig. 2. Optimal subset using mGA

From figure 2 Top 17 positioned features are selected as the optimal subset in mean based genetic algorithm.



```
bestFeatures
     MIR4640 RFC2 HSPA6 PAX8 GUCA1A MIR5193 THRA PTF
1,]        1    1     1    1      1       1    1
     MMP14 TRADD CHURC1.FNTB PLD1 DTX2P1.UPK3BP1.PMS
1,]      1     1           1    1
     EIF3D PARK7 SRP14 GDI2 GDI2.1 RPL11 ARF3 RPL21F
1,]      1     1     1    0      1     1    1
     LOC100508408 FAU TARDBP CFL1 RPL18 EIF3F RPS5 F
1,]             1   1      1    1     1     1    1
     SNORD35B RPL9 MIR5047 RPL6 CTDNEP1 RPL10A CBX3
1,]         0    1       1    0       1      0    1
     ATP6V1G2.DDX39B RTCB ERH GATC ABCF1 DAD1 YY1 J1
1,]                1    1   1    1     1    1   1
```

Fig. 3. Optimal subset using ACO

In ACO feature selection method 94 features were selected. From figure 3, 1 represents the selected feature and 0 represents the unselected features.

The results of feature Selection using mGA and ACO and evaluation of accuracy rate using SVM were tabulated

Table 1
Results Representation

Dataset	GDS3096
Total features	100
Optimal features(ACO)	94
Optimal features(mGA)	17
Accuracy rate (ACO-SVM)	87%
Accuracy rate (Mg-SVM)	89%

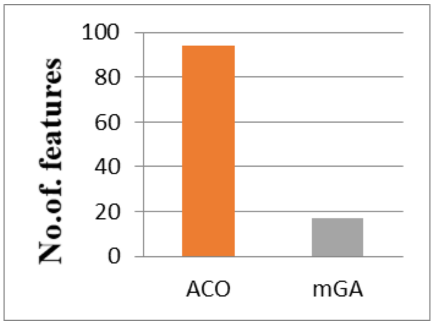


Fig. 4. Features selection

From figure 4, using ACO algorithm 94 features were selected and 17 features were selected by mGA method.

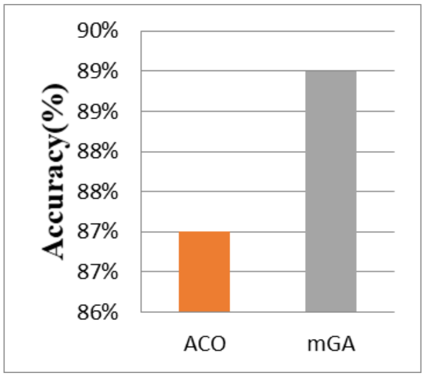


Fig. 5. SVM Accuracy Estimation

The accuracy of the selected optimal subset was evaluated using the Support vector machine. ACO obtained 87% and mGA obtained 89% for their optimal subset.

6. Conclusion

Feature selection is used to find the optimal features from high dimensional data. This project proposed a mean based genetic algorithm to select the optimal feature from a dataset. The dataset was taken from the NCBI repository. The accuracy of the selected featured was measured using the Support Vector Machine and compared with the ACO feature selection method. The results have shown that the mean based Genetic Algorithm had higher accuracy rate with low number no of features while comparing to the ACO feature selection method. Simultaneously the proposed method reduced the complexity of large dataset in terms computational time and size of the dataset. In future, the feature selection work will be extended by considering other bio inspired algorithm for feature selection.

References

[1] Pragadeesh, C., Jeyaraj, R., Siranjeevi, K., Abishek, R. and Jeyakumar, G., 2019. Hybrid feature selection using micro genetic algorithm on microarray gene expression data. *Journal of Intelligent & Fuzzy Systems*, 36(3), pp. 2241-2246.

[2] Peng, H., Ying, C., Tan, S., Hu, B. and Sun, Z., 2018. An improved feature selection algorithm based on ant colony optimization. *IEEE Access*, *6*, pp. 69203-69209

[3] Alzubi, Raid, Naeem Ramzan, Hadeel Alzoubi, and Abbes Amira. "A hybrid feature selection method for complex diseases SNPs." *IEEE Access* 6 (2017): 1292-1301

[4] An, Shuai, Jn Wang, and Jinmao Wei. "Local-nearest-neighbors-based feature weighting for gene selection." *IEEE/ACM transactions on computational biology and bioinformatics* 15, no. 5 (2017): 1538-1548.

[5] Azzawi, Hasseeb, Jingyu Hou, Yong Xiang, and Russul Alanni. "Lung cancer prediction from microarray data by gene expression programming." *IET systems biology* 10, no. 5 (2016): 168-178.

[6] Gangeh, Mehrdad J., Hadi Zarkoob, and Ali Ghodsi. "Fast and scalable feature selection for gene expression data using hilbert-schmidt independence criterion." *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 14, no. 1 (2017): 167-181.

[7] Liu, Cheng, and Hau San Wong. "Structured Penalized Logistic Regression for Gene Selection in Gene Expression Data Analysis." *IEEE/ACM transactions on computational biology and bioinformatics* 16, no. 1 (2017): 312-321.

[8] Tang, Jian, and Shuigeng Zhou. "A new approach for feature selection from microarray data based on mutual information." *IEEE/ACM transactions on computational biology and bioinformatics* 13, no. 6 (2016): 1004-1015.

[9] De Souza, J.T., De Francisco, A.C. and De Macedo, D.C., 2019. Dimensionality Reduction in Gene Expression Data Sets. *IEEE Access*, *7*, pp. 61136-61144.

[10] Almugren, N. and Alshamlan, H., 2019. A Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification. *IEEE Access*, *7*, pp.78533-78548

[11] Jain, D. and Singh, V., 2018. An efficient hybrid feature selection model for dimensionality reduction. *Procedia Computer Science*, *132*, pp. 333-341.

[12] Lamba, M., Munjal, G. and Gigras, Y., 2018. Feature Selection of Micro-array expression data (FSM)-A Review. *Procedia computer science*, *132*, pp. 1619-1625.

[13] Wang, H., Jing, X. and Niu, B., 2017. A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data. *Knowledge-Based Systems*, 126, pp.8-19

[14] Fayyazifar, N. and Samadiani, N., 2017, October. Parkinson's disease detection using ensemble techniques and genetic algorithm. In *2017 Artificial Intelligence and Signal*Processing Conference (AISP) (pp. 162-165). IEEE.

[15] Bonilla-Huerta, E., Hernandez-Montiel, A., Morales-Caporal, R. and Arjona-López, M., 2016. Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data. *IEEE/ACM Transactions on* Computational Biology and Bioinformatics (TCBB), 13(1), pp. 12-26.