International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 05, 2020
ISSN: 1475-7192

# Feature Selection for Gene Expression Data Analysis – A Review

Dr.R. Prema

***Abstract---*** *Gene selection in microarray data analysis is defined as the process of identifying a small number of informative and relevant genes that can find any sample from the dataset into the correct class. The feature selection process is categorized into three types: wrapper, embedded and filter techniques. Filter methods use statistical ranking for feature selection by ordering the features individually. They select the relevant features independent of any supervised learning algorithm. The wrapper techniques use a number of search methods to evaluate the possible subset of important features. From that it selects the subset of features that gives the best classification accuracy. In embedded methods, feature selection methods are incorporated in the training process. This paper reviews several feature selection methods used to find significant features from gene expression data for use in classification*.

***Keywords---*** *Feature Selection Methods, Microarray Gene Expression Data, Gene Selection, Classification*.

## I. INTRODUCTION

Feature selection is one of the significant machine learning tasks when faced with data having enormous size, large numbers of missing entries and high noise (Pena *et al*.2001). Leaving out relevant attributes or keeping irrelevant attributes may affect the performance of the machine learning algorithm. There are many feature selection approaches to assist in the classification of samples.

A gene expression dataset contains thousands of gene expression values, many of which may be redundant or irrelevant for Classification (Horng*et al*.2009). Therefore, statistical methods are required to find the most important genes before classification is carried out (Arauzo *et al*.2011). This paper delivers a review on three main feature selection approaches for microarray gene selection.

The rest of this paper is organized as following. Section 2 presents the details of the filter based feature selection techniques. Section3 and section 4 discusses a number of wrapper based feature selection approaches and embedded methods respectively. Section 5 concludes the paper and provides overall discussion.

## II. FILTER-BASED METHODS

A filter approach applies a statistical measure (SNR, T-Test etc) to assign a score to each feature without using a machine learning algorithm. Many filter-based feature selection algorithms have been developed based on information theory, statistical ranking, rough and fuzzy set theory, etc.

Yeung & Bumgarner (2003) developed two algorithms. The first method is the uncorrelated shrunken centroid algorithm, and the second one is based on error-weighted uncorrelated shrunken centroid algorithm. Both algorithms are included feature selection methods and classification procedures. Data with any number of classes can be applied

*Dr.R. Prema, Assistant Professor, Department of MCA, New Horizon College of Engineering, Bangalore. Email: premabit@gmail.com*

**DOI: 10.37200/IJPR/V24I5/PR2020695**
**Received: 16 Apr 2020 | Revised: 22 Apr 2020 | Accepted: 12 May 2020** 

to these algorithms. Both algorithms use the dependence measure between genes to minimize the number of selected genes. The error-weighted method takes variability estimates of repeated measurements into consideration to down-weight noisy genes. But uncorrelated algorithm is valid to microarray gene expression datasets with or without repeated measurements.

Correlation-based Feature Selection (CFS) is a simple filter method for feature ranking (Hall1999). It follows the principle that a good subset of features is one that includes features highly related with the class yet not related with each other. It uses a correlation- based heuristic evaluation function to rank the features. But it only finds the features which are more relevant but not redundant to any other features of the class. Wang *et al*. (2005) applied CFS on gene expression datasets. They recommended CFS for fast analysis of data.

Rough set theory is a mathematical tool for minimizing the redundancy in information systems. Li & Zhang (2006) proposed a gene selection method, Rough Maximum Interaction-Maximum Relevance (RMIMR). This algorithm selects the significant genes in terms of relevance and interaction. That is relevance of genes and the linear positive interaction between genes should be maximized. To resolve the RMIMR optimization problem, this method uses a simple heuristic algorithm. From the experimental results, it has been found that by using the proposed method classification accuracy can be improved significantly.

Meyer *et al*. (2008) presented a filter based technique for gene expression data called Double Input Symmetrical Relevance (DISR). It is based on a new information theoretic selection criterion. DISR relies on a measure of variable complementarity. They examined the application of the proposed feature selection method on microarray expression datasets. Also they stated that DISR is applicable to gene expression analysis for three main causes: Cost of computation is very low; using trivariate mutual information, highly multivariate mutual information is approximated; and potential two-by-two gene complementarity is also considered.

Navarro & Munoz (2009) described an Entropic Filtering Algorithm (EFA) for attribute selection, to generate a relevant subset of genes. This algorithm finds gene subsets that maximize the multivariate conditional entropy (normalized). The EFA was examined with many classification algorithms on a number of gene expression datasets. The results achieved were of good quality against previous algorithms' performance. They stated that EFA is a speedy feature selection algorithm, since the algorithm used a very few number of genes and achieved results in a half an hour computing time.

Sun *et al*.(2010) extended the ideas implemented in the Relief algorithm (Kira & Rendell 1992) and proposed a feature selection algorithm based on local learning. This algorithm decomposes an arbitrarily complex nonlinear problem into a set of locally linear ones. Then it learns feature relevance globally within the large-margin framework. It addresses many issues including problems with computational complexity, accuracy and algorithm implementation. The authors examined the application of the proposed algorithm on microarray datasets with a huge number of irrelevant genes. It was demonstrated that the proposed algorithm can be implemented efficiently, since it avoids any heuristic combinatorial search.

Many gene selection algorithms have been proposed based on empirical mutual information. These algorithms

**DOI: 10.37200/IJPR/V24I5/PR2020695**
**Received: 16 Apr 2020 | Revised: 22 Apr 2020 | Accepted: 12 May 2020** 

suffer from the sparseness issue, with gene expression datasets that contain a small number of samples and large number of genes or features. To overcome this problem, Zhu *et al*.(2010) proposed a model-based algorithm to calculate the entropy of class labels based on the model. The authors used multivariate Gaussian generative models for gene subset selection. They derived gene selection methods based on the D- 0ptimality and A-0ptimality criteria. The computation of all possible log-determinants is not efficient with the large number of genes. So, they proposed a number of algorithms to greatly minimize the cost. Their experiments on seven microarray datasets and comparison with five other feature selection methods showed the accuracy of the proposed algorithms.

Mishra &Sahu (2011) derived a model for gene selection using SNR ranking. They proposed two methods of gene selection. In the first method, the genes are clustered by k-means clustering. Then, SNR feature selection method is implemented to obtain the top- scored genes from each group or cluster, which is given to two machine learning algorithms for validation, such as SVM and kNN. In the second approach, the genes are ranked by implementing only SNR ranking, and the top- scored genes are given to the classifier and validated. It was concluded that the first method for gene selection is better in comparison to the second method, because after applying the SNR method and selecting the top-ranked genes from each cluster, a true pattern is given, which helps to improve the classification accuracy. The authors tested a Leukemia dataset with the proposed approach and used a 10-fold cross-validation method to validate the classifiers.

Bolon-Canedo*et al*.(2011) presented a statistical dependence measure for feature selection. The authors explored this method in a classification scenario with categorical class labels, where the previous study (Seth & Principe 2010) considered only continuous variables. They tested its performance over sixteen microarray datasets (binary and multiclass) and compared it against the minimum Redundancy Maximum Relevance feature selection method, in terms of five different classifiers. It was shown that the proposed method obtained performance better than or equal to that of mRMR over the binary datasets.

Song *et al*.(2012) introduced a feature selection algorithm, BAHSIC, which defines a class of backward elimination feature selection algorithms that make use of (i) kernels and (ii) the Hilbert-Schmidt independence criterion. It follows the principle that good features should be highly dependent on the class labels. This is a nonparametric dependence measure, which considers all modes of dependence between the features. The authors showed that the proposed method demonstrates good performance on microarray datasets compared with the more specialized state-of-the-art methods. Also, the authors stated that the BAHSIC algorithm is very competitive in terms of runtime performance.

Liu *et al*. (2013) proposed a method, called Robust Principal Component Analysis to find differentially expressed genes. They treated the differentially expressed genes as perturbation signals S and the non-differentially expressed genes as a low-rank matrix A. The matrix D of raw dataset is divided into two additive matrices A and S. Then, based on matrix S, the differentially expressed genes are found. Finally, using Gene Ontology (GO), the differentially expressed genes are identified. A larger number of comparisons and experiments on microarray data showed that the proposed algorithm is efficient.

**DOI: 10.37200/IJPR/V24I5/PR2020695**
**Received: 16 Apr 2020 | Revised: 22 Apr 2020 | Accepted: 12 May 2020** 

Dai & Xu (2013) proposed a feature selection method based. The authors used fuzzy information gain ratio based on fuzzy rough set theory. Rough set theory is used widely as a feature selection approach. While microarray datasets contain continuous gene expression values, the crisp rough set theory cannot handle continuous values directly. The fuzzy rough set theory was introduced to proceed with continuous- valued gene selection. This approach was compared to many feature subset selection techniques on three cancer datasets. The results showed that the proposed method is efficient.

Hoque *et al*.(2014) introduced a greedy feature selection method, MIFS-ND. This method uses mutual information. This algorithm finds an optimal subset of features by minimizing the redundancy and by maximizing the relevance among features. To find the relevance and redundancy measures, mutual information is used (Battiti 1994). This algorithm combines both feature-feature mutual information and feature-class mutual information to select a subset of high ranked features. For all the features, feature-class mutual information also average feature-feature mutual information is calculated. From these results, a feature that has the maximum feature-class mutual information, but lowest feature-feature mutual information, is selected. For that purpose, this method used Non-dominated Sorting Genetic Algorithm-II (Deb *et al*.2000). It is an optimization algorithm. The performance of the selected features was analyzed using four classification algorithms on several datasets, including two microarray gene expression datasets. The performance of the proposed method was found to be significantly high in terms of both classification accuracy and execution time and for all the datasets.

Hoque *et al*.(2016) presented fuzzy mutual information-based feature selection method with non- dominated solution. This algorithm uses a fuzzy based mutual information measure. It selects features based on feature-class fuzzy mutual information and also based on feature-feature fuzzy mutual information. The proposed method was examined using gene expression and other machine-learning datasets. The classification accuracy was evaluated by using four different classifiers. A modified kNN algorithm was also developed to calculate the classification accuracy. The authors provided experimental analysis of the proposed method with several feature selection algorithms in terms of accuracy.

Significance Analysis of Microarrays (SAM) (Tusher *et al*.2001) is one of the filter-based method. Kang & Song (2017) applied small changes to SAM and proposed new filter-based gene selection techniques, since they observed from the results that the scores of the genes selected from SAM were lesser than estimated. To prove the performance of the proposed technique, a series of datasets are considered with different sample sizes and noise levels. The results proven that the proposed techniques found significant genes comparing with the original SAM method. By avoiding the overestimation of variance, target genes could be found in robust manner with the weighting schemes.

## III. WRAPPER-BASED METHODS

A wrapper approach uses machine learning techniques to calculate the classification accuracy produced by the use of the selected features in the classification. A few examples of learning techniques are simulated annealing, genetic algorithm, branch and bound method, particle swarm optimization, etc.

**DOI: 10.37200/IJPR/V24I5/PR2020695**
**Received: 16 Apr 2020 | Revised: 22 Apr 2020 | Accepted: 12 May 2020** 

Xiong *et al*.(2001) developed three feature wrappers for biomarker identification. The wrapper algorithms search through the space of feature subsets: support vector machines, logistic regression and linear discriminant analysis etc. The authors employed sequential forward search and sequential forward floating search methods to effectively carry out the computationally intensive search process. To evaluate the performance of the proposed gene selection, three gene expression datasets were considered. The experimental results demonstrated that, by identifying composite classifiers with several gene markers produces very high classification accuracy.

Maugis*et al*.(2009) presented feature selection for cluster analysis with Gaussian mixture models. A general model was proposed to specify the role of each feature (Raftery & Dean 2006).This model does not require any prior assumptions about the link between the discarded and selected features. A feature's role is achieved through an algorithm that embeds two backward stepwise feature selection algorithms for clustering and linear regression. The consistency of the resulting criterion is justified under regularity conditions.

Ai-Jun & Xin-Yuan (2009) proposed a Bayesian stochastic variable selection approach for gene selection. For simulating parameters from the posterior distribution, this method used simulation-based Markov chain Monte Carlo methods. It was also shown that the proposed algorithm is robust to the selection of initial values. This algorithm produces posterior probabilities of related genes for biological interpretation. To evaluate the performance of the proposed method, it was compared with other methods on Colon cancer and Leukemia datasets. From the results, it was stated that the proposed algorithm is efficient and used a small number of the most significant genes to perform classification.

Any Partial Least Squares (PLS) - based feature selection is based on some kind of supervised learning approach. Ji *et al*.(2011) presented a PLS-based feature selection, which synthesizes genetic relatedness and is suitable for multiclass classification. The authors described three PLS-based indicators for gene selection by using the explanation difference of independent features on dependent features (class variable). The proposed method considered the combined effects of all the genes and the correlation among the genes. This method was tested in Acute Myeloid Leukemia and Acute Lymphoblastic Leukemia and Small Round Blue Cell Tumor datasets, integrated with Kernel Support Vector classifiers. A subset of significant genes with small numbers and high identification was achieved. The experimental results showed that the proposed PLS-based gene selection method is highly efficient for cancer classification.

Sharma *et al*. (2012) proposed a top-r feature selection algorithm for gene expression data analysis of sample classification. The two main procedures of the algorithm are Successive Feature Selection (SFS) and block reduction. A $d-$ dimensional feature vector is divided into $m$ equal blocks. All the blocks are allowed to process through the SFS method one at a time. In successive levels, features are dropped one at a time, and a subset of features is obtained. Then, the classification accuracy is calculated using a classifier, and the best subset of features is sent to the next level. This algorithm merges this smaller subset of genes to update the gene subset. The process is repeated until all subsets are merged into one top- $r$ informative gene subset. The experimental results indicated that the proposed method gave promising classification accuracy for all the gene expression datasets. The relevance of the selected genes in terms of their biological functions was also presented.

**DOI: 10.37200/IJPR/V24I5/PR2020695**
**Received: 16 Apr 2020 | Revised: 22 Apr 2020 | Accepted: 12 May 2020** 

Srivastava *et al*.(2014) evaluated the performance of filter versus wrapper gene selection techniques over three gene expression datasets, Ovarian Cancer, Lymphomas and Leukemia. A Relief F algorithm was used as a filter-based gene selection method, and a random feature subset selection algorithm was used as a wrapper-based gene selection method. A random subset selection algorithm generates a random subset of genes and assesses their quality independently. These genes are used to classify the samples with learning algorithms. Only the best gene subsets are kept based on classification accuracy. Then, the algorithm selects a pool of the most frequent significant genes, and uses discriminant analysis to search a subset of genes over a randomized subset of genes.

Hui*et al*.(2016) developed a wrapper-based feature selection algorithm for classification of cancer subgroups. They used the t-test method to reduce the number of genes in the dataset. Then a particle swarm optimization-based approach was employed to find the most significant genes. It is an example of a stochastic computation technique. It has developed from study of the interaction between communal groups of birds (Kennedy & Eberhart 1995). The proposed method was applied to the SRBCT microarray dataset to classify four subgroups: neuroblastoma, non-Hodgkin lymphoma, rhabdomyosarcoma and Ewing sarcoma. The proposed method identified a set of 14 differentially expressed genes that could classify four SRBCT classes with 100% accuracy.

## IV. EMBEDDED-BASED METHODS

The embedded approach combines a feature selection step and classifier construction. Some examples are decision trees, random forests, weighted naive Bayes, etc.

Random forest is an algorithm for classification that uses an ensemble of classification trees. Díaz-Uriarte& de Andrés (2006) used random forest for gene selection and classification. Each of the classification trees is built using a bootstrap sample of the data, and at each split the candidate set of features is a random subset of the features. To select genes, the random forests are iteratively fit; at each iteration, a new forest is built after discarding those features with the smallest feature importance and OOB error rate. This algorithm found very small numbers of non-redundant genes, while preserving classification accuracy.

Maldonado *et al*.(2011) introduced an embedded method called Kernel-Penalized SVM (KP- SVM). The main idea is to penalize the use of features in the dual formulation of SVM. SVM can eliminate features that have low relevance for the classifier by optimizing the kernel function. For each problem, the proposed method attempts to find the best suitable RBF- type kernel function, which consists of a minimal dimension obtained by combining the parameters of generalization, goodness of fit and feature selection. Also, the proposed method uses an explicit stopping condition that avoids the elimination of features that would negatively affect the classifier's performance. The authors performed experiments on four microarray datasets and compared the results with those of other known feature selection techniques. KP-SVM outperformed the alternative approaches and determined consistently fewer relevant genes.

Xiang *et al*. (2012) proposed a feature selection and multiclass classification framework based on Discriminative Least Squares Regression (DLSR). This algorithm enlarges the distance between different classes with the conceptual framework of Least Squares Regression. The algorithm works as follows. Firstly, the proposed algorithm

**DOI: 10.37200/IJPR/V24I5/PR2020695**
**Received: 16 Apr 2020 | Revised: 22 Apr 2020 | Accepted: 12 May 2020** 

develops a training model with compact form for DLSR. This model translates the one- versus-rest training rule for multiclass classification problems. Then, based on this compact model, a learning framework with sparse representation both on the regularization term and on the LSR term is developed for variable selection. The homogeneous coordinate representation for the LSR achieved an efficient and effective solution to variable selection. The authors examined the proposed method using several datasets, including eight microarray datasets. Additionally, the authors presented theoretical analyses of the derived model for classification.

Canul-Reich *et al.*(2012) introduced an embedded gene selector, the iterative feature perturbation method. This method starts with the entire dataset with all genes and uses a backward elimination algorithm. At each iteration, the algorithm removes the least important genes and reduces the gene set. The importance of each gene relies on the impact on classification accuracy that each gene has when disturbed by noise values. The gene is considered relevant, if adding noise leads to a big change in the prediction results. Likewise, non-relevant genes will give less or no impact on the prediction accuracy. Then, the non-relevant genes are removed. The author applied the proposed method on four cancer datasets.

Table 1 gives complete reviews of feature selection methods for microarray gene expression data.

Table 1: Review of Feature Selection Methods for Microarray Gene Expression Data

| S.No. | Reference | Feature selection type | Topic | Methodology | Dataset |
|---|---|---|---|---|---|
| 1 | Yeung& Bumgarner (2003) | Filter | Multiclass classification of microarray data with repeated measurements: application to cancer | Uncorrelated Shrunken Centroid algorithm, and the Error-Weighted Uncorrelated Shrunken Centroid algorithm | Multiple tumours and Breast cancer |
| 2 | Wang *et al.*(2005) | Filter | Gene selection from microarray data for cancer classification - machine learning approach | Correlation-based feature selection | Leukemia and DLBCL |
| 3 | Li & Zhang (2006) | Filter | Gene Selection Using Rough Set Theory | Rough Maximum Interaction-Maximum Relevance | Leukemia and Colon datasets with two and multi-class labels |
| 4 | Meyer *et al.*(2008) | Filter | Information-theoretic feature selection in microarray data using variable complementarity | Double input symmetrical relevance, information-theoretic selection | Leukemia, Brain tumor, Prostate, DLBCL |
| 5 | Navarro & Munoz (2009) | Filter | Gene subset selection in microarray data using entropic filtering for cancer classification | Maximizes the normalized multivariate conditional entropy with respect to the class labels | Lung, Prostate and Breast cancer |
| 6 | Sun *et al.*(2010) | Filter | Local Learning Based Feature Selection for High Dimensional Data Analysis | Relief algorithm-based selection | DLBCL, Prostate and Breast cancer |
| 7 | Zhu *et al.*(2010) | Filter | Feature selection for gene expression using model-based entropy | Empirical mutual information based selection | ALL, AML, Lymphoma, SRBCT datasets |
| 8 | Mishra &Sahu (2011) | Filter | Feature Selection for Cancer Classification: A Signal-to-noise Ratio Approach | Data is clustered by *k*-means clustering and then SNR ranking to clusters | Leukemia dataset |

**DOI: 10.37200/IJPR/V24I5/PR2020695**
**Received: 16 Apr 2020 | Revised: 22 Apr 2020 | Accepted: 12 May 2020** 

| S.No. | Reference | Feature selection type | Topic | Methodology | Dataset |
|---|---|---|---|---|---|
| 9 | Bol´on-Canedo*et al.*(2011) | Filter | Statistical dependence measure for feature selection in microarray datasets | A measure of monotone dependencies used to assess the relevance and redundancy | Leukemia, DLBCL, Colon, Prostate, Lung, Breast cancer datasets |
| 10 | Song *et al.*(2012) | Filter | Feature Selection via Dependence Maximization | Hilbert-Schmidt Independence Criterion based selection | Hepatitis, Breast cancer and other benchmark datasets |
| 11 | Liu *et al.*(2013) | Filter | Robust PCA-based method for discovering differentially expressed genes | Genes of differential expression are identified according to the perturbation signals | Colon cancer |
| 12 | Dai & Xu (2013) | Filter | Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification | Uses a fuzzy equivalence relation, instead of crisp equivalence relation | SRBCT, Hepatocellular carcinoma and Colon cancer |
| 13 | Hoque*et al.*(2014) | Filter | MIFS-ND: A Mutual Information-based Feature Selection Method | Combines both feature-feature mutual information and feature-class mutual information | Lymphoma and Colon cancer |
| 14 | Hoque*et al.*(2016) | Filter | A fuzzy mutual information-based feature selection method for classification | Uses fuzzy mutual information measure with Non-Dominated solution | Colon, Leukemia and Breast cancer |
| 15 | Kang & Song (2017) | Filter | Robust gene selection methods using weighting schemes for microarray data analysis | Uses a different variance structure, leading to robustness on identifying significant genes in the presence of outliers | Fusarium, Leukemia |
| 16 | Xiong*et al.*(2001) | Wrapper | Biomarker Identification by Feature Wrappers | Sequential forward search and sequential forward floating search algorithms | Affymetrixoligonucleotide array, Breast cancer |

| S.No. | Reference | Feature selection type | Topic | Methodology | Dataset |
|---|---|---|---|---|---|
| 17 | Maugis*et al.*(2009) | Wrapper | Variable Selection for Clustering with Gaussian MixtureModels | Embeds two backward stepwise feature selection algorithms | Simulated datasets |
| 18 | Ai-Jun & Xin-Yuan (2009) | Wrapper | Bayesian variable selection for disease classification using geneexpression data | Simulation-based Markov chain Monte Carlo methods | Colon and Leukemia datasets |
| 19 | Ji *et al.*(2011) | Wrapper | PLS-Based Gene Selection and Identification of Tumor-Specific Genes | Algorithm considers combined effects of all the genes and is integrated with a linear kernel support vector classifier | SRBCT and ALL/AML datasets |
| 20 | Sharma *et al.*(2012) | Wrapper | A Top-r Feature Selection Algorithm for Microarray Gene Expression Data | Successive feature selection and block reduction | SRBCT, Prostate and MLL datasets |
| 21 | Srivastava *et al.*(2014) | Wrapper | Filter vs. Wrapper approach for optimum gene selection of high dimensional gene expression dataset: An analysis with cancer datasets | ReliefF algorithm with random feature subset selection algorithm | Ovarian Cancer, Lymphomas & Leukemia datasets |
| 22 | Hui*et al.*(2016) | Wrapper | A Particle Swarm optimization Based Gene Identification Technique for Classification of Cancer Subgroups | T-test with particle swarm optimization | SRBCT dataset |
| 23 | Maldonado *et al.*(2011) | Embedded | Simultaneous feature selection and classification using kernel-penalized support vector machines | Penalize the use of features in the dual formulation of SVM | CMA and LMA datasets |

**DOI: 10.37200/IJPR/V24I5/PR2020695**
**Received: 16 Apr 2020 | Revised: 22 Apr 2020 | Accepted: 12 May 2020** 

| S.No. | Reference | Feature selection type | Topic | Methodology | Dataset |
|---|---|---|---|---|---|
| 24 | Xiang *et al.*(2012) | Embedded | Discriminative Least Squares Regression for Multiclass Classification and Feature Selection | The algorithm enlarges the distance between different classes under LSR conceptual framework. | CAR, Glioma, Lung, MLL, SRBCT, CLL-SUB-111, GLA-BRA-180, and TOX-171 |
| 25 | Canul-Reich *et al.*(2012) | Embedded | Iterative feature perturbation method as a gene selector for microarray data | Used the backward elimination approach, an embedded featureselector. This determined which features areto be eliminated in the next step | Colon, Leukemia and Lung cancer |

## V. CONCLUSION

The advent of microarray gene expression data has posed a challenging task in gene selection, because of the large number of input features and small sample size. This chapter summarizes the various gene selection methods, which are classified based on the type of supervised learning algorithm used, namely filter, wrapper and embedded. From the literature it is found that the wrapper and embedded-based algorithms are computationally inefficient compared to the filter methods. The filter methods are preferred for high- dimensional gene selection in that they do not use a supervised learning algorithm, possess better generality, and require less computational complexity.

## REFERENCES


[1] Ai-Jun, Y & Xin-Yuan, S 2009, 'Bayesian variable selection for disease classification using gene expression data', *Bioinformatics,* vol. 26, no. 2, pp. 215-222.

[2] Arauzo, A, Aznarte, JL&Benítez, JM 2011, 'Empirical study of feature selection methods based on individual feature evaluation for classification problems', *Expert Systems with Applications,* vol. 38, no. 7, pp. 8170-8177.

[3] Battiti, R 1994, 'Using mutual information for selecting features in supervised neural net learning', *IEEE transactions on neural networks,* vol. 5, no. 4, pp. 537-550.

[4] Bol´on-Canedo, V, Seth, S, Sanchez-Maro˜no, N & Alonso-Betanzos, A 2011, 'Statistical dependence measure for feature selection in microarray datasets', *Proceedings of nineteenth European Symposium on Artificial Neural Networks,* pp. 27-29.

[5] 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,* vol. 13, pp. 211-221.

[6] Deb, K, Agrawal, S, Pratap, A & Meyarivan, T 2000, 'A fast elitist non-dominated sorting genetic algorithm for multi- objective optimization: Nsga-ii', *Proceedings of International conference on parallel problem solving from nature,* pp. 849-858

[7] Díaz-Uriarte, R & de Andrés, SA 2006, 'Gene selection and classification of microarray data using random forest', *BMC Bioinformatics,* vol. 7, no. 3, pp. 1-13.

[8] Hall MA 1999, Correlation-based Feature Selection for Machine Learning. Ph.D. thesis, University of Waikato.

[9] Hoque, N, Ahmeda, HA, Bhattacharyyaa, DK & Kalitab, JK 2016, A Fuzzy Mutual Information-based feature selection method for classification', *Fuzzy Information and Engineering,* vol. 8, no.3, pp. 355-384.

[10] Hoquea, N, Bhattacharyyaa, DK &Kalitab, JK 2014, 'MIFS-ND:

[11] A Mutual information-based feature selection method', *Expert Systems with Applications,* vol. 41, no. 14, pp. 6371-6385.

[12] Horng, JT, Wu, LC, Liu, BJ, Kuo, JL, Kuo, WH& Zhang, JJ 2009, 'An expert system to classify microarray gene expression data using gene selection by decision tree', *Expert Systems with Applications,* vol. 36, no. 5, pp. 9072-9081.

[13] Hui, KH, Ooi, CS, Lim, MH, Leong, MS & Al- Obaidi, SM 2016, 'An improved wrapper-based feature selection method for machinery fault diagnosis', *PLoS ONE,* vol. 12, no. 1, pp. 1-10.



**DOI: 10.37200/IJPR/V24I5/PR2020695**
**Received: 16 Apr 2020 | Revised: 22 Apr 2020 | Accepted: 12 May 2020**

[14] Ji, G, Yang, Z & You, W 2011, 'PLS-Based Gene Selection and Identification of Tumor-Specific Genes', *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 41, no. 6, pp. 830-841.
[15] Kang, S & Song, J 2017, 'Robust gene selection methods using weighting schemes for microarray data analysis', *BMC Bioinformatics,* vol. 18, no. 1, pp. 1-15.
[16] Kennedy, J &Eberhart, RC 1995, 'Particle swarm optimization', *Proceedings of IEEE International conference on neural networks,* pp. 1942-1948.
[17] Kira, K & Rendell LA 1992, 'A practical approach to feature selection', *Proceedings of ninth International conference on machine learning,* pp. 249-256.
[18] Li, W & Zhang, W 2006, 'Gene selection using rough set theory', *Proceedings of the first International conference on rough sets and knowledge technology,* pp. 778-785.
[19] Liu, JX, Wang, YT, Zheng, C, Sha, W, Mi J &Xu Y 2013, 'Robust PCA based method for discovering differentially expressed genes', *Proceedings of International conference on Intelligent,* pp. 25-29.
[20] Maldonado, S, Weber, R & Basak, J 2011, 'Simultaneous feature selection and classification using kernel-penalized support vector machines', *Information Sciences,* vol. 181, no. 1, pp. 115-128.
[21] Maugis, C, Celeux, G & Martin-Magniette, M 2009, 'Variable Selection for Clustering with Gaussian Mixture Models', *Biometrics,* vol. 65,no. 3, pp. 701-709.
[22] Meyer, PE, Schretter, C &Bontempi, G 2008, 'Information-theoretic feature selection in microarray data using variable complementarity', *IEEE Journal of Selected Topics in Signal Processing,* vol. 2, no. 3, pp. 261-274.
[23] Mishra, D &Sahu, B 2011, 'Feature selection for cancer classification: A Signal-to-Noise Ratio approach', *International Journal of Scientific & Engineering Research,* vol. 2, no. 4, pp. 1-7.
[24] Navarro, FG & Munoz, LB 2009, 'Gene subset selection in microarray data using entropic filtering for cancer classification', *Expert Systems,* vol. 26, no. 1, pp. 113-124.
[25] Pena, JM, Lozano, JA, Larranaga, P &Inza, I 2001, 'Dimensionality reduction in unsupervised learning of conditional Gaussian networks', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 6, pp. 590-603.
[26] Raftery, AE & Dean, N 2006, 'Variable Selection for Model-Based Clustering', *Journal of the American Statistical Association,* vol. 101, no. 473, pp.168-178.
[27] Seth, S & Principe, JC 2010, 'Variable selection: A statistical dependence perspective', *Proceedings of ninth international conference on machine learning and applications,* pp. 931-936.
[28] Sharma, A, Imoto, S &Miyano, S 2012, 'A Top-r Feature Selection Algorithm for Microarray Gene Expression Data', *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 9, no. 3, pp. 754-764.
[29] Song, L, Smola, A, Gretton, A, Bedo, J & Borgwardt, K 2012, 'Feature selection via dependence maximization', *Journal of Machine Learning Research,* vol. 13, pp. 1393-1434.
[30] Srivastava, B, Srivastava, R & Jangid, M 2014, 'Filter vs. wrapper approach for optimum gene selection of high dimensional gene expression dataset: an analysis with cancer datasets', *Proceedings of the International Conference on High Performance Computing,* pp. 1-6.
[31] Sun, Y, Todorovic, S &Goodison, S 2010, 'Local learning based feature selection for high dimensional data analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 9, pp. 1610 - 1626.
[32] Tusher, VG, Tibshirani, R & Chu, G 2001, 'Significance analysis of microarrays applied to the ionizing radiation response', Proceedings of the National Academy of Sciences, pp. 5116-21.
[33] Wang, Y, Tetkoa, IV, Hallb, MA, Frankb, E, Faciusa, A, Mayera, KF&Mewesa, HW 2005, 'Gene selection from microarray data for cancer classification- a machine learning approach', *Computational Biology and Chemistry,* vol. 29, pp. 37-46.
[34] Least squares regression for multiclass classification and feature selection', *IEEE Transactions Neural Net works Learning Systems,* vol. 23, no. 11, pp. 1738-1754.
[35] Feature Wrappers', Genome Research, vol. 11, pp. 1878-1887.
[36] microarray data with repeated measurements: application to cancer', *Genome Biology,* vol. 4, no. 12, pp. 83.1-83.19.
[37] Zhu, S, Wang, D, Yu, K, Li, T & Gong, Y 2010, 'Feature Selection for Gene expression using model-based entropy', *IEEE transactions on computational biology and bioinformatics,* vol. 7, no. 1, pp. 1610-1626.

**DOI: 10.37200/IJPR/V24I5/PR2020695**
**Received: 16 Apr 2020 | Revised: 22 Apr 2020 | Accepted: 12 May 2020** 