A REVIEW ON FEATURE SELECTION METHODS AND HIGH DIMENSIONAL GENOMIC DATA

M. Sathya [1], Dr.S.Manju Priya [2]

ABSTRACT

For certain diseases, therapeutic intervention is limited due to varying symptoms, poor medical response and rapid disease development. Genomic studies have revolutionized the way of treating diseases by offering promising ways to detect disease risks, improve preventive methods and offering personalized treatments. The development of laboratory techniques to identify biomarkers generates a vast amount of genomic data. Genomic data comprised different microarray formats and these formats include information pertaining to molecular, biological and cellular levels. The genomic information composed of different micro units which are represented in microarrays. Each microarray is composed of different gene expression features. The choice of features on a microarray data greatly affects the quality of detecting the genes associated with a particular disease. Different methods continuously being investigated to handle high dimensional microarray data and feature selection methods are to distinguish genes from disease and non-disease conditions. This paper aims to review recent feature selection techniques that are applied on microarray dataset. Different dimensionality reduction methods are presented and the performance is analyzed with respect to computational complexity and accuracy in detecting diseases at gene functional level.

I. INTRODUCTION

The detection of thousands of gene expressions from DNA microarray offers many benefits to understand different diseases and their developments. DNA microarray technology aids the diagnosis of diseases especially cancers. The course of Cancer development is crucial to its treatment, and understanding DNA at molecular level is complex and challenging [4]. The capture of the abnormalities of genes with respect to their structure, molecule and function is aided through microarray and gene sequencing technology. The onset and development of a particular cancer can be predicted by monitoring genes expressed in both non-cancerous and cancerous cells. Microarray data are composed of genes and their expressions at different levels involving thousands of genes. Using microarray data, classification of cancerous cells and the type of cancer become possible. A microarray data represents a matrix in which rows represent gene expressions and columns represent disease conditions (Fig 1).

The Classification of a particular cancer and its type is challenging even with microarray data, as the gene expressions are present in thousands and matching a particular gene expression for a cancer type from thousands of genes become complex [14]. The analytical process of finding the relevant features is

[1]Research Scholar, Dept. of Computer Science, Karpagam Academy of Higher Education, Coimbatore.

[2]Associate Professor, Dept of CS, CA & IT, Karpagam Academy of Higher Education, Coimbatore.

assisted by data mining and machine learning techniques. It is challenging to choose a suitable method from a variety of machine learning algorithms as the classification ability and accuracy differs among different methods [20, 12, 11]. Selecting relevant features that can minimize the classification or prediction errors is studied endlessly and different feature selection techniques are being continuously introduced. Feature selection extends the ability to understand features present in the data and are broadly classified into filter, wrapper and embedded methods.

In a microarray dataset, all genes do not offer enough information and affect the classification accuracy as noises. Also, handling a large number of features require high computational power and time. However, the presence of irrelevant and noisy features may affect the disease classification accuracy. Selecting the genes that are relevant and contain information can be useful for an accurate classification of disease.

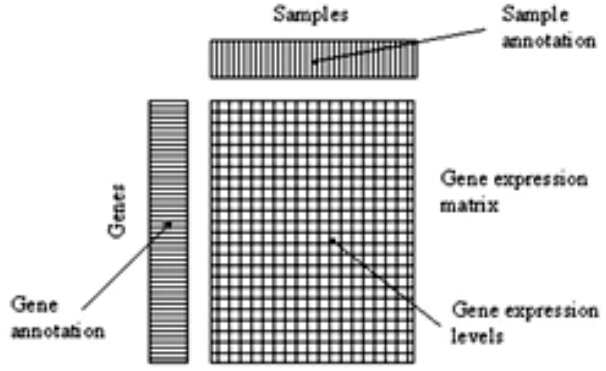

Figure 1 Microarray matrix with gene expressions in rows and Samples/diseases in columns

Feature selection extends the ability to understand features present in the data and are broadly classified into filter, wrapper and embedded methods. Filter methods uses intrinsic property to rank features such as gene information, similarity, distance and dependency functions to select genes. Wrapper methods select genes using a search method embedding a classifier inside to evaluate the feature importance to classification. The selection of gene subset is subject the classifier chosen. Embedded methods use feature selection as an extension of their models.

This paper is organized into the following sections,-: section 2 discusses various feature selection studies carried out in recent times and section 3 concludes the paper with future direction.

II RELEATED WORKS

Geet al, [8] proposed a feature selection technique for identifying phenotype-based genes. The proposed a wrapper method is based on correlation and maximal information coefficient, which measures the MIC between genes. Using seventeen microarray dataset, the model performance is studied. The proposed method is capable of handling different types as the coefficients are discretized. Using KNN classifier as an evaluator, the features are reduced to subset through best first search method. The performance is compared with PAM, RRF and CFS and the results indicate that the proposed method outperforms PAM and RRF by 13%, while CFS has better accuracy.

Jain et al., [13] propose a two-step feature selection method utilizing CFS and improved BPSO. The two-step procedure selects a subset of genes that is relevant to cancer target. In the first step, using the correlation-based feature selection genes with higher predictive power are selected in a subset. The subset is optimized using IBPSO and wrapper NB algorithm, which later improves the classification accuracy. The CFS coupled IBPSO solves the problem of early convergence of

local optimum through weighting criteria. The performance of the model is evaluated using tenfold cross validation to avoid higher variances on eleven microarray datasets. The proposed model outperforms the existing methods with 100 % accuracy, with more than 60% reduction in gene size.

Azizet et al., [2] proposed a new feature extraction method by combining ICA and fuzzy backward feature elimination. The proposed methods target minimizing genes by only selecting the most relevant informative genes and eliminating non-informative genes. The proposed method selects independent components using backward fuzzy-feature elimination. The resulting subset is applied to improve SVM and NB classification performance.

Using five different cancer datasets, the classification performance of SVM and NB using the reduced gene subset is studied. The classification result shows the performance of SVM and NB highly improved because of gene reduction. SVM and NB show higher accuracy of 90% and 85% when compared to PCA (75%).

Sharbaf et al., [22] proposed a filtering approach using Fischer score and a wrapper method to select genes from microarray data. Using Fischer criterion, features are ranked and the highest-ranking features are filtered. The high ranking gene subset is then applied to cellular learning automata to learn the gene relationships and optimized through ant colony optimization to select the final subset of genes. The resulting subset is evaluated using tenfold cross validation on four different datasets on SVM, KNN and NB. After feature selection, the classification accuracy of SVM improved from 58% to 95%, KNN from 70% to 95% and NB from 94% to 97%. Based on AUC, genes with higher AUC are selected and when applied to classification, they show better classification rate.

Chinnaswamy et al., [3] proposed a hybrid model for selecting relevant features. The model uses correlation values and PSO for selecting best features. The feature-elimination in microarray dataset is carried out using CFS. The final feature subset is evaluated using an extreme learning machine, J48, RF, Rtree, Decision stump and Genetic programming on three datasets. The feature selection is studied on three different datasets using different algorithms stated above. ELM shows better accuracy (93%) than other methods. The proposed method significantly reduces gene numbers there by reducing computational costs and time.

Hu et al., [10] proposed a feature selection method for high dimensional microarray data. The proposed method uses a modified shuffle frog leap algorithm for dimension reduction. By using a balanced group arrangement based on memory weight and transfer function, it detects the features that have higher prediction rate. The modified SFLA is evaluated on nine different datasets using KNN classifier. The proposed method outperforms other methods such as IGA, IPSO and SFLA with 83% of accuracy and also significantly reduces the features in the high dimensional data.

He et al., [9] proposed a new process-oriented feature selection method (MINT) for high dimensional data, where features are higher than the samples. Dimensionality-reduction method is developed using maximum relevancy and minimum redundancy with respect to transduction of genes. The MRMR method is a measure that accounts for higher relevancy to the class and minimum dependence between genes. The

maximum relevancy (MR) calculates mutual information between target classes and features while maximum redundancy (mR) calculates mutual information between features and the performance of the classifier is based on these two mutual information values. The proposed method is tested on 4 different datasets and compared with MRMR. The performance results show that the proposed method outperforms MRMR in reducing the gene size.

Mollaee et al., [16] proposed an ensemble-based feature selection method for microarray classification. The ensemble method is composed of different ranking methods such as blogreg, Ttest and Fisher criteria and using the mean ranking from the three methods, genes are selected. The ensemble part of the proposed method can be treated as a pre-processing method and in the second part, features are selected using discriminant ICA optimized through PSO. The selected features are evaluated with SVM (Gaussian kernel) as the base classifier using six microarray datasets and compared with kernel PCA, Probability PCA, GPLVM, Isomap, FA, PCA and ICA. The classification accuracy for the proposed method (PSO-dICA) feature selection shows, PSO-dICA out performs other feature selection methods by avoiding local optimum which improves the differentiation of the target classes.

Sahu et al., [21] proposed a framework that uses clustering and feature ranking methods to remove irrelevant features in microarray data. Using unsupervised feature grouping using similarity is carried out through k-means clustering. Ranking techniques such as SNR, Ttest, and SAM are used to select high ranking features and using ensemble classifier the classification accuracy is compared. The ensemble classifier framework includes DT, NB, KNN and MLP algorithms and the performance is compared on four different datasets. The proposed ensemble model significantly improves the classification accuracy.

Das et al., [5] proposed a feature selection method using ensemble of parallel processing. The parallel processing involves bi-objective genetic algorithm using information theory and mutual information functions. The feature subset is generated using sampling replacement, where the non-informative features are removed. Using fourteen different datasets, the proposed Genetic algorithm-based feature selection is evaluated using NB, SVM, KNN, BOOST, MLP, SMO and DT. Bi-objective function filters the most relevant features and outperforms other feature selection methods for 9 datasets using non-dominated decision aggregation.

González et al., [19] proposed a framework that uses boosting based feature selection for classifying lung cancer subtypes. The framework is a typical CBR which targets for differentiating squamous cancer cells and adenocarcinoma using reduced gene features. The feature selection works at two levels, and in, the first level, most unrelated genes are removed and in the second distance-based property is used to retrieve cases with minimum feature genes. Base on the retrieved cases, the CBR then calculates the class probability using weighted KNN. The model is validated using different classifiers and finally, the proposed framework is able to produce high accuracy by retaining fewer cases with reduced gene subset avoiding training for new cases.

Wang et al., [24] introduced a wrapper-based framework for selecting highly informative gene

subset. Using Markov blanker method irrelevant feature genes are removed to produce gene subset with quality genes in microarray data. A ranking method guides the wrapper model to select relevant features. Symmetrical uncertainty is used to rank the feature genes based on the relationship to the class and a subset of genes is created. Later, the subset is evaluated using SFS, a greedy forward selection procedure that removes the redundant features in the subset through Markov blanket (SFS-MB). Incremental wrapper subset with MB is applied to select the features and remove redundant features. The performance is validated using ten different datasets using C4.5, KNN and NB as base classifiers with tenfold cross validation. The proposed framework with SFS-MB showed better performance with smaller gene features.

Nagpal et al., [17] have developed a new technique on selection of gene subsets through solving the problems of microarray dataset. The new technique adds feature score from random forest to mutual information for generating quality subset. Feature scores of the dataset are derived using correlated feature with higher distances from random forest. For the resulting subset with high rank features, mutual information is calculated and the ranks are added to generate qualitative mutual information. Different subsets are generated in the same way and the converging features are selected finally. IB1, C4.5 and PNB are used to evaluate the quality of the subset. The classification and computation time of the selected algorithms are improved greatly when applying the final feature subset with accuracy of 87%, 98% and 100%.

Gao et al., [7] proposed a new hybrid feature selection strategy that aligns both classification information and reduction redundancy. The proposed method uses minimum redundancy-Maximal new Classification (MR-MNCI) to address both class dependent and class independent of redundancy. The first step calculates MI between the class dependent and the second the new classification information and class independent information. Finally features are selected based on their importance. The proposed method is compared with other feature selection methods using twelve datasets. SVM and NB are used to test the classification performance. The accuracy of the proposed method achieved higher percentage than other methods on seven datasets.

Shukla et al., [23] proposed a new feature selection method for gene selection from microarray data. The proposed work is based on recursive PSO which alters gene space to fine particles by assigning SVM weights and integrated with ranking methods improves the classification. The PSO method is mainly used for feature reduction. The proposed work (RPSW) is compared with other feature selection methods such as F-score, Wilconxon's Rank and MI using five different microarray datasets. The RPSW achieved higher accuracy rate of 98% (SVM) in minimum steps while other method requires more recursive steps.

Arunkumar et al., [1] proposed a new method for feature selection in high dimensional data using rough-set based algorithm. The algorithm uses a novel similarity measure on fuzzy rough set introduced in the study. Information gain is used to minimize the high dimension and the proposed fuzzy rough method is used to select relevant genes as well as remove irrelevant genes. The feature selection method is evaluated on three different microarray data using RF classifier. The selection of relevant genes is based on the similarity measure that estimates the relation of the

gene features to the target classes. The tenfold cross validation using RF produced 98% of accuracy over other existing methods.

Qi et al., [18] proposed a new unsupervised feature selection using matrix factorization. Apart from matrix regularization, the proposed method uses correlation between features and feature weight matrix as weights to regularization and detects redundancy. RMFFS outperforms NMS, GNMF and MFFS as a result of considering correlation between low redundant features. For unsupervised selection, K-means clustering is used to cluster the features and achieve a mean accuracy of 55% over other feature selection methods.

Liu et al., [18] proposed a new feature selection method for combined methods. In combined methods, development of newer methods is challenging due to difference in feature evaluation and rich feature interaction. To overcome these limitations, the proposed method is constructed using genetic algorithm, feature based and selection based, a hybrid model that fuses genetic algorithm and regularization. The evolutionary method optimizes the learning and search strategy and embeds tuning of solutions. The proposed HGAWE method outperforms other combined methods on five microarray data with a mean accuracy of 94%.

Dashtban et al.,[6] proposed a novel a biologically inspired method for gene selection from microarray dataset. Initially a subset of genes is filtered using Fisher score, secondly using binary Bat algorithm is applied to build a wrapper method which can potentially distinguish the relationship between genes and classes. The local search method is modified using random walk and named as multi objective binary BA (MOBBALS). The subset of genes selected through MOOBALS and its classification accuracy is evaluated using NB, DT, KNN and SVM classifiers. The classification methods achieved an accuracy of 99% using the subsets of genes generated by MOOBALS for three datasets.

Table 1. Performance of Various Feature Selection methods

References	Method	Feature Information	Data set	Accuracy(%)
(Ge, *et al* ., 2016)	Wrapper	MIC	17	(SVM, NB, DT, NN) 13% ↑
(Jain *et al* ., 2018)	Wrapper	CFS+iBPSO+ NB	11	NB 100%
(Aziz *et al*., 2016)	Filter	ICA+FBFE	5	SVM 90% NB 85%
(Sharbaf *et al.*, 2016)	Filter	Fischer score +ACO	4	SVM 95% KNN 95% NB 97%
(Chinnaswamy *et al.*, 2016)	Hybrid	PSO+CFS	3	J48 81% RF 93% Rtree 72% ELM 93%
(Hu *et al.*, 2016)	Wrapper	ISFLA	9	KNN 83%
(He *et al.*, 2016)	Transductive	MINT	4	NA
(Mollaee *et al.*, 2016)	Ensemble	dICA+PSO	6	SVM 94%
(Sahu *et al*., 2017)	Unsupervised	RPSW	5	Kmeans 99%
(Arunkumar *et al.*, 2018)	Filter	IG + RS	3	RF 98%
(Qi *et a* l., 2018)	Unsupervised	RMFFS	3	K-means 55%
(Liu *et al*., 2018)	Combined	GA+Regularization	5	HGAWE 94%
(Dashtban *et al* ., 2018)	Hybrid	GAFS	14	NB 91% SVM 94% KNN 86%
(Das *et al* ., 2017)	Ensemble	RS+MMI+GA	14	EFSGA 90%
(González *et al.*, 2017)	Framework	GBoost	1	SVM 94.5% NB 95.9% KNN 95.3%
(Wang *et al.*, 2017)	Wrapper	SFS-MB	10	1NN 79.9% NB 80.5% C4.5 80.5%
(Nagpal *et al.*, 2018)	Wrapper	QMI	4	NB 98.61% IB1 90.72%
(Gao *et al.*, 2018)	Hybrid	MRMNCI	12	SVM+NB 86%
(Shukla *et al.*, 2018)	Ensemble	RPSW	6	SVM 98% NB 94%

III. CONCLUSION

Feature selection is becoming more important as it reduces the dimensionality problems especially in genomic data. In the classification problem, the accuracy of the classifier is often affected by redundant features, and evidently from this survey, it has become a more fundamental way to select relevant features through efficient feature selection techniques particularly in microarray dataset. Apart from traditional methods such as wrapper, filter and embedded, hybrid models, combined models and ensemble feature selection have huge potential to uncover feature importance (Table 1). Since the growth of data volume and data complexity in drug discovery and cancer related studies, the need for feature selection for microarray data is exponentially growing. More sophisticated methods are in demand to meet the current biological discoveries and research requirements. At the same time, it is clear that the ensemble methods are promising to select relevant features and also improve the classification performance. As future work new ensemble-based feature selection method would be appropriate for selecting relevant features for high dimensional microarray data.

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) Chinnaswamy, A& Srinivasan, R (2016). Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data. In Innovations in Bio-Inspired Computing and Applications (pp. 229-239). Springer, Cham

4) Dancey, J.E, Bedard, P.L, Onetto, N& Hudson, T.J (2012). The genetic basis for cancer treatment decisions. Cell, 148(3), 409-420.

5) Das,A. K., Das, S., & Ghosh, A. (2017). Ensemble feature selection using bi-objective genetic algorithm. Knowledge-Based Systems, 123, 116-127.

6) 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.

7) Gao, W, Hu, L, Zhang, P& Wang, F (2018). Feature selection by integrating two groups of feature evaluation criteria. Expert Systems with Applications.

8) Ge ,R, Zhou, M,Luo, Y, Men,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.

9) He,D, Rish., I, Haws.D& Parida.L (2016). Mint: Mutual information based transductive feature selection for genetic trait prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 13(3), 578-583

10) Hu.B, Dai. Y, Su. Y, Moore. P, Zhang. X, Mao. C & Xu, L (2016). Feature selection for optimized high-dimensional biomedical data using the improved shuffled frog leaping algorithm. IEEE/ACM transactions on computational biology and bioinformatics.

11) Huang, C. L, Liao, H. C & Chen, M (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Systems with Applications, 34(1), 578-587.

12) Huang, Y. L & Chen, D. R (2005). Support vector machines in sonographer: application to decision making in the diagnosis of breast cancer. Clinical imaging, 29(3), 179-184.

13) 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.

14) Kuo, W. P, Kim, E. Y, Trimarchi, J, Jenssen, T. K, Vinterbo, S. A& Ohno-Machado, L (2004). A primer on gene expression and microarrays for machine learning researchers. Journal of Biomedical Informatics, 37(4), 293-303.

15) Liu, X.Y, Liang, Y, Wang, S, Yang, Z.Y, & Ye, H. S (2018). A Hybrid Genetic Algorithm with Wrapper-Embedded Approaches for Feature Selection. IEEE Access, 6, 22863-22874.

16) Mollaee, M& Moattar, M. H (2016). A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. Biocybernetics and Biomedical Engineering, 36(3), 521-529.

17) Nagpal, A& Singh, V (2018). A Feature Selection Algorithm Based on Qualitative Mutual Information for Cancer Microarray Data. Procedia Computer Science, 132, 244-252.

18) Qi, M, Wang, T, Liu, F, Zhang, B, Wang, J, & Yi, Y (2018). Unsupervised feature selection by regularized matrix factorization. NeuroComputing, 273, 593-610.

19) Ramos-González, J, López-Sánchez, D, Castellanos-Garzón, J.A., de Paz, J. F & Corchado, J. M (2017). A CBR framework with gradient boosting based feature selection for lung cancer subtype classification. Computers in biology and medicine, 86, 98-106.

20) Ryu, Y.U, Chandrasekaran, R, & Jacob, V.S(2007). Breast cancer prediction using the isotonic separation technique. European Journal of Operational Research, 181(2), 842-854.

21) 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.

22) Sharbaf, F.V, Mosafer, S, & Moattar, M.H. (2016). A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics, 107(6), 231-238.

23) Shukla, A. K, Singh, P& Vardhan, M (2018). A hybrid gene selection method for microarray recognition. Biocybernetics and Biomedical Engineering

24) 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.

25) Zhou, Z. H, Jiang, Y, Yang, Y. B & Chen, S. F (2002). Lung cancer cell identification based on artificial neural network ensembles. Artificial Intelligence in Medicine, 24(1), 25-36.

26) Zou, H & Hastie, T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.