Abstract: Feature selection has become an elementary tool for processing high dimensional data. DNA microarray technology is used for the study of large number of genes simultaneously, which helps in determining the expression levels of the genes. Gene selection using high dimensional gene expression data is foremost and imperative for prediction and classification of disease. This gene expression data can be shown in the form of matrix and usually contains irrelevant, redundant and noisy data, so the study and analysis of data becomes very problematic. The prime purpose of feature selection approaches is to remove the curse of dimensionality, improve the performance and accuracy of classification and clustering algorithms by the elimination of these irrelevant features and reduction of noise. This paper explains the taxonomy of feature selection methods stating their respective pros and cons. It also presents a review on few feature selection approaches, mainly those that have been proposed over the past few years.

Keywords: Bioinformatics, DNA microarray, Feature selection, High dimensionality, Gene expression

1. Introduction

With the advancement in biomedical research for the classification of a disease, mainly cancer and for prognosis or diagnosis of disease at early stage, genes are taken into concern. As human genome contains tens of thousands of genes, studying and performing experiments on those genes at once is a cumbersome task. The study and analysis of these genes is done using microarray experiment. A microarray gene expression dataset can be represented in the tabular or matrix form in which rows represent genes and columns represent sample or time point and each table entry illustrates the expression level of a particular gene at a particular time or in a sample [1]. As there is large number of genes as compared to the small sample size, a feature selection method for dimensionality reduction is required. Feature selection is widely used in many fields including data mining, machine learning, pattern recognition and statistics. Its prime purpose is to choose most informative subset of features using some criteria.

Feature selection when used for microarray gene expression data is called gene selection. Apart from curse of dimensionality there are many other problems faced in gene selection like mislabeled data, redundant data, irrelevant and noisy data, and problem of cross-platform comparisons, erroneous and bias problem and difficulty in biological information retrieval [2]. Various gene selection methods and algorithms are proposed in literature which is able to reduce the dimensionality by removing irrelevant, redundant and noisy genes. The gene expression data which is drawn in the form matrix also contains some missing values which are caused by factors like insufficient resolution, image corruption or due to scratches on slide. Various methods have been used to fill those missing values. Commonly used Naïve methods are imputing the missing values with zeros or row averages or removing the gene expression profile containing missing value. These methods don’t provide complete estimation of missing value and looses useful information. So some other methods have been developed to impute the values like KNN impute method, Bayesian principal components analysis impute method, Support Vector Regression method (SVR) [3]. These methods give good estimation accuracy on the datasets.

Feature selection can be classified depending upon two different criteria. Firstly, on the basis of prior knowledge available, it is categorized as supervised, unsupervised and semi-supervised. Secondly, depending upon how they combine the selection algorithm with model building it can be subdivided into filter, wrapper, embedded, hybrid and ensemble methods. All these five methods possess their own advantages and disadvantages. Some utilize the advantages of other two methods to give better results. A large literature is available that provides...
the taxonomy of these methods [2], [4-7]. The latest development in feature selection methods is hybrid and ensemble methods. Hybrid method is less prone to overfitting than wrapper while ensemble is more flexible and robust.

The rest of this paper is organized as following. Section 2 presents the details of the background which includes bioinformatics, DNA microarrays, and taxonomy of feature selection methods. Section 3 discusses recent advancements in feature selection approaches and Section 4 concludes the paper and provides overall discussion.

2. Background

Bioinformatics: The Merriam-Webster dictionary defines bioinformatics as the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics [8]. Genes are the fundamental unit of heredity and the building blocks of DNA. Genome is the genetic material of an organism, comprising complete set of DNA, including all of its genes. The various applications of bioinformatics include comparative genomics, genome annotation, health and drug discovery, prediction of protein structure, sequence analysis etc.

To study the behavior of various genes present in an organism, biologists need to study the activity of individual genes, which is a cumbersome task. With the advent of DNA microarray the expression levels of a large number of genes can be determined at once in a single experiment. The work of M. Madan Babu gives the overview of the basic concepts involved in the microarray experiments and also discusses various computational methods that can be used to derive fruitful results from such experiments [9]. The human genome consists of thousands of genes. At a particular moment, each of our cells has some combination of these genes which are turned on and other combinations are turned off. The work of biologists is to figure out which genes are turned on and which are off, that can be determined with the amount of mRNA produced. The central dogma of molecular biology explains the process of making of proteins from individual strands of DNA [10].

Feature selection methods: Feature selection is the method of determining the subset of original features. It is a dimensionality reduction technique having application in fields like data mining, pattern recognition, machine learning, statistics and gene selection. The objectives of feature selection are manifold, the major ones are: (a) To avoid overfitting and improve model performance i.e. prediction, (b) to provide faster and more cost effective models, and (c) to gain a deeper insight into the underlying processes that generated the data [11].

Feature selection is particularly used in cases where the gene expression data comprises of thousands of genes but with small sample size. There are three types of gene expression data available i.e. fully labeled, partially labeled or unlabeled which assists the generation of supervised, semi-supervised and unsupervised feature selection for class prediction [2]. Supervised feature selection is the basic method and uses the labeled data to train the feature selection model. It aims at selecting the feature subset based on some criteria for evaluating the significance and relevance of features. It requires labeling of data given by external knowledge which is a challenging task as it is expensive to generate labeled data and may be unreliable. Unsupervised feature selection is more exigent than semi-supervised and supervised feature selection as it is not accompanied by labeled classes. It is applied to circumvent redundant or missing value in microarray gene expression data [12]. Semi-supervised feature selection shows better results for gene expression microarray data as the data consists of large number of unlabeled data but limited labeled data. It combines advantages of both the approaches and hence, is able to find most promising and informative features. Various semi-supervised feature selection approaches were surveyed in [13], like density separation approach, transductive SVM, semi supervised based logistic discriminate procedure etc. These approaches help in increasing classification accuracy. Another semi-supervised cluster ensemble framework was introduced in [14] which apart from gene selection also helped in improving tumor clustering performance by selecting the optimal subset clustering solutions in the ensemble. The various advantages and disadvantages of supervised, semi supervised and unsupervised feature selection methods are summarized in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>Utilizes labeled data for training feature selection model</td>
<td>Expensive to generate labeled data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Worsens the risk of over-fitting the learning</td>
</tr>
</tbody>
</table>

Table 1. Advantages and disadvantages of feature selection methods
The stages in feature selection process are elaborated. (a) The process begins with the determination of starting point and the search direction. The search can be forward, backward elimination or bi-directional search, (b) then, a good search strategy is decided which is further categorized into exponential, sequential or randomized search, (c) then, an evaluation method for feature selection is decided from filter, wrapper, hybrid, embedded and ensemble, which results in quality feature subset extraction, (d) once the features are extracted, a stopping criteria is defined to avoid over-fitting and lower computational complexity, (e) if the process stops successfully, then validation of final results will take place using various error estimation or validation techniques and if some error occurs, then the process goes back to step 1 and will start from determination of search direction [2].

The methods of feature selection can also be classified on the basis of learning algorithm used as follows:

1. **Filter method** It is also known as open loop method, looks only at the intrinsic properties of the data like distance, correlation and consistency. Most of the filter methods explained in literature are univariate. They are computationally faster, easily scalable, efficient and are independent of the classifier. Major drawbacks of filter methods are that they avoid interactions with the classifier and, as most of the methods are univariate, they consider each feature separately hence ignoring dependencies among features. A number of reasons are provided to explain the domination of univariate method [4]. Filter method degrades classification performance, hence various multivariate filter methods were introduced like Correlation based feature selection (CFS), Markov blanket filter (MBF) etc. A comprehensive survey of filter feature selection methods is given in [6].

2. **Wrapper method** It is also known as closed loop method, interacts with the classifier and depends upon the learning algorithm to estimate the value of given subset. It uses performance evaluation as feature evaluation criteria. It gives higher performance accuracy than filter as it considers dependencies among features. Among the disadvantages of wrapper method, one is that it is less general than filter method as it needs to be re-executed every time a new learning method is utilized. So it doesn’t provide any assurance that the result is optimal for different learning algorithms. Moreover, it is more prone to overfitting and requires extra computation time as most of the wrapper methods given in literature are multivariate. A comparison of filter and wrapper gene selection methods is done in [15]. The study proposes different

<table>
<thead>
<tr>
<th>Unsupervised</th>
<th>Semi-supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Unbiased, no experts or data analysts required for categorization of samples.</td>
<td>- Utilizes both labeled and unlabeled data</td>
</tr>
<tr>
<td>- High performance, no prior knowledge available.</td>
<td>- Often find most discriminative and informative features.</td>
</tr>
<tr>
<td>- Powerful way to uncover the unknown important understanding of classification of disease types.</td>
<td>- Labeled data is expensive.</td>
</tr>
<tr>
<td>- Neglects the feasible correlation between different features.</td>
<td>- Relies on some mathematical principles which are not universally valid for all data.</td>
</tr>
</tbody>
</table>
filter metrics for discrete and continuous microarray data. For discrete data, metrics used are Shannon Capacity, Euclidean distance, Kolmogorov dependence, Kullback-Leibler, while for continuous data P-metric and t-score is used.

3. Embedded method It interacts with the classifier and has better computational complexity than wrapper method, as it doesn’t allow the recurrent execution of classifier and study of each feature group. Unlike wrapper, it is less prone to overfitting but has major problem of computational complexity in high dimensional data and also it is classifier specific. Like wrapper it also considers dependency among the features. The recent development in feature selection methods is determined by hybrid and ensemble methods.

4. Hybrid method It is obtained by combining either two feature selection methods (wrapper and filter) or two methods belonging to same criteria. It combines the strengths of both the methods and results in higher performance than filter and better computational complexity than wrapper. It is less prone to overfitting than wrapper but like embedded and wrapper methods it is also classifier specific. In [16] a hybrid filter-wrapper method is followed which shows more accurate results with smaller feature subset when compared to individual filter and wrapper methods.

5. Ensemble method It is flexible and robust (as performance of feature selection don’t depends upon single subset) with high dimensional data. Its aim is to overcome the instability problem of other feature selection methods. Ensemble method builds a group of feature subsets and then constructs an aggregated result out of the group.

3. Overview of feature selection approaches

This section provides details of the work done by various researchers in the field of feature selection in previous few years. Feature selection’s main task is to decrease the number of genes and hence obtain a subset of promising genes, so that they can be used by different classifiers for classification of disease etc. The work by Banu et al. gives an evolutionary approach for informative gene selection for early disease diagnosis, especially cancer. It looks for the most interpretable features that have high impact on the disease and helps to recognize the target disease [17] The pre-eminent role of Support vector machine (SVM) for classification of disease is used by various researchers [11], [18-20]. Semi supervised SVM for cancer classification is depicted in [21] which uses fuzzy preference based rough sets for feature selection. Transductive SVM is combined with feature selection in [22] to improve prediction accuracy as compared to inductive SVM.

Heuristic approaches are more efficient and accurate for feature selection in high dimensional microarray dataset. A hybrid meta-heuristic algorithm called Genetic Bee Colony (GBC) algorithm which combines the advantages of two naturally inspired algorithms: Genetic Algorithm (GA) and Artificial Bee Colony (ABC) is explained by Alshamlan et al. A complete balance between exploitation and exploration is needed for meta-heuristic population based algorithms. ABC algorithm is simple, easy to implement and competitive but it suffers from problems like large computational time and cost, mainly for high dimensional data. For reducing this, preprocessing is done in which mRMR filter is used to filter noisy and redundant genes. Both ABC and GA have some limitations in their original algorithms which can be overcome by hybridization. ABC has good exploration process, hence it easily finds a new solution in optimization search space but suffers from poor exploitation causing long computational time for convergence. On the other hand, original GA has poor exploration process causing pre convergence problem but has good exploitation operations. GBC algorithm realizes the advantages of both ABC and GA and has balanced exploration and exploitation processes. Few modifications are done in the basic ABC and GA algorithm to enhance their abilities. The experiments on various binary and multi class datasets showed that GBC algorithm selects few genes with high classification accuracy, and also when compared with other algorithms like ABC, mRMR-ABC, mRMR-PSO, mRMR-GA. GBC gives better results [18].

Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) along with F-score is detailed by Fattah et al. to select the most informative genes. TOPSIS method aimed at selecting the best alternative which is nearest to the positive ideal solution and farthest from the negative ideal solution. A five step process is elaborated which includes, (i) matrix representation of gene expression data, (ii) computation of normalized decision matrix, (iii) computation of weighted normalized decision matrix, (iv) deduction of positive-ideal and negative-ideal solutions (v) computation of relative closeness to the ideal solution and ranking the order of performance. Gene ranking is done on the basis of TOPSIS method and top 250 genes are then selected and F distribution is then applied on it. Various classifiers are then used to
check the correctness of the proposed method and it is concluded that when reduced gene data set is 10 and SVM is used as classifier then it gives higher classification accuracy then other methods [19].

Ibrahim et al. uses Multi-level feature selection to select genes based on expression profiles. It combines both deep and active learning. Active learning is used here in unsupervised manner. The main components of the approach includes, (i) deep belief nets, which are graphical models that store interaction and behavior of the genes by generating high level representation of the genes, (ii) feature selection which is used to select the subset of genes and (iii) unsupervised active learning, to reduce the subset and hence increase classifier accuracy. It uses SVM and Random Forests classification to construct gene and sample classifier respectively. Deep Belief Nets (DBN) was first compared with random, statistical t-test, information gain, relief-f and chi-square. Then F-measure was used to compare DBN with above approaches. The result shows that DBN was able to achieve highest F-measure among all approaches [20].

In the work done by Nguyen et al. a novel approach for genes selection is proposed in which improvement to the conventional Analytic Hierarchy Process (AHP) is done. In the conventional AHP no quantitative criteria were taken and their results are derived by experts. This become issue when the problem being solved contains large number of criteria and is taken from different knowledge areas as the experts’ knowledge is restricted to the area of their expertise. Modified AHP is a composite approach where quantitative criteria is taken, which here uses five test statistics i.e. t-test, entropy, Receiver Operating Characteristics (ROC), Wilcoxon and Signal to Noise Ratio (SNR) [23]. Figure 1 represents the steps involved in modified AHP.

In hierarchy construction three levels tree structure is made in which five criteria are considered simultaneously. In priority calculation the range [1, 10] is considered for sorting the importance of a gene compared to other genes. Here the absolute judgment is replaced by rigorous judgments which give real numbers instead of rough numbers. Then eigenvectors (ε) and eigenvalues (λ) are calculated representing ranking scores of the genes. It is required that the matrix obtained should be consistent, so the elements must be transitive i.e. $x_{in} = x_{im} x_{mn}$. For consistency verification consistency index (CI) and consistency ratio (CR) is calculated. CR value 0 indicates that the judgments are purely consistent. As a result option performance matrix is obtained, which contain their respective eigenvectors. Ranking of genes is done to select the top ranked genes for the purpose of classification. Here IT2FLS is used for classification and its result is compared with that of multilayer perceptron (MLP), support vector machine (SVM) and fuzzyARTMAP (FARTMAP) by calculating average the leave one out cross validation (LOOCV) accuracy, area under the ROC curve (AUC), sensitivity and specificity of the mentioned classification methods for different gene subsets. Range of genes is set to perform classification techniques. Genes, if taken beyond the range will cause performance reduction. Small subsets of genes don’t provide enough information for classification whereas large number of genes would lead to high dimensionality problem. It is concluded that the performance of modified AHP is high as compared to other gene selection methods.

Work done by Kar et al. aimed to reduce the dimensionality and is suitable for identifying types of cancer. A stochastic optimization approach called Particle Swarm Optimization (PSO) is used coupled with SVM and K-nearest neighborhood (KNN) as classifiers. It uses threshold values to each candidate solution vector to select or discard the genes. Performance comparison of KNN and SVM shows that KNN classifier is superior to SVM. PSO is preferable for multi variable optimization problems. It is an evolutionary approach which embeds swarm intelligence to engineering applications. Each particle here represents a candidate solution. Personal best position of the particles is decided and the global bestparticle is found. The representation of entire data sample ${S[m, n]}$, where $m$ denotes the number of genes and $n$ is the number of samples is done and then k-fold cross validation technique is employed to get $k$ equal subsets of samples from the entire dataset. Out of these $k$ subsets, single subset is taken for testing as validation data and $(k-1)$ subsets are taken as training data and the step is repeated $k$ times, until each of the $k$ subset is used once as validation data. Fitness value is obtained by combining both CV accuracy and number of genes.
This work is applied on three microarray datasets i.e. SRBCT, ALL_AML and MLL. This also includes detailed comparison of various approaches of feature selection on each of the given datasets depending on number of genes selected, cross validation accuracy, training and test accuracy. It is concluded that PSO adaptive KNN based approach is better in terms of number of genes selected and outperforms classification accuracy [24].

K-Means is one of the most popularly used clustering algorithms. It makes $k$ subsets of the provided dataset in such a way that all the objects in the given subset are closest to the same centroid [25]. Various approaches focus on improving the initial centroids of the $k$-means algorithm [26-30] hence increasing the accuracy of the algorithm. This method is commonly used, since it is easy to understand and has linear time complexity [31]. But they don’t consider clusters and feature selection in full feature space. A wrapper based feature selection approach that uses population based search technique PSO to generate subsets and $k$-means algorithm for the evaluation of subsets is explained. It deals with the problem of informative gene selection in high dimensional data. It gives the detailed explanation of basic PSO algorithm. Each particle maintains its current position ($x_j$), current velocity ($v_j$) and personal best position ($y_j$). The new velocity of each particle is calculated using equation (1) [32]

$$v_j(n) = w v_j(o) + c_1 r_1 (y_j - x_j(o)) + c_2 r_2 (y_g - x_j(o))$$

where $n$ is new, $o$ represents old, $w$ is inertia weight, $r_1$ and $r_2$ are random numbers in the range [0, 1], $c_1$ and $c_2$ are acceleration constants, $v_j(o)$ is the particle’s current best solution, $y_g(o)$ is best solution of particle found so far and $y_g$ is the global best position. Next the new position of each particle is calculated by equation (2)

$$x_j(n) = x_j + v_j(n)$$

The position that minimizes the fitness criteria is taken and each particle will consider its and that of other particles knowledge. The final set obtained will determine the optimal solution.

Deepthi et al. proposed a method in which PSO is mapped into clustering gene expression data where initially a subset of genes is taken randomly representing the whole swarm. K-means clustering is used to calculate fitness value, then to find its initial centroids $k$-means++ algorithm is used. Once the initial centroids are chosen, then the process to assign points to the closest ones repeats and new centroid calculation began. The main idea behind this process is the minimization of squared error criteria which is given in equation (3) [33]

$$\text{sum of squared error} = \sum_{j=1}^{k} \sum_{x \in C_j} d^2(c_j, x)$$

where $c_j$ is the centroid of cluster $C_j$, $x$ is an object that belongs to cluster $C_j$ and Euclidean distance is the used distance measure. Three datasets Leukemia, MLL, DLBCL datasets were taken and four methods Random Projection, Ranker followed by (PCA), Correlation Feature Selection (CFS) subset using Sequential Forward Search (SFS) and $k$-means algorithm were applied. The performance analysis is done by considering two criteria i.e. within cluster sum of squared error (WC error) equation (4) and cluster purity equation (5).

$$\text{WC error} = \sum_{j=1}^{k} \sum_{x \in C_j} (x - m_j)^2$$

WC error determines the cohesion of elements in a cluster. Less value of WC error indicates that the clusters are more compact.

$$\text{Purity}(X \cap C) = \frac{1}{M} \sum_{k=1}^{K} \max |x_k \cap C_j|$$

The value of Purity determines the accuracy of the clusters. It ranges from 0 to 1. As the value approaches 1, it means accurate clustering. According to the above two criteria the proposed method of integrating PSO and $k$-means performs better than other methods and helps to achieve an accuracy of 70-80% for all the datasets taken.

Dheenathayalan et al. have used Genetic Algorithm (GA) for the selection of remarkable genes from microarray datasets. They used GA as a preprocessing technique and implemented it using R language and then the significant genes are used for disease diagnosis or prognosis. For further processing data mining techniques are used. Genetic algorithms are the evolutionary algorithms which rely on the concept of natural selection and population genetics. Various parameters are taken for the identification of informative and significant genes, like (i) chromosomes which are evaluated using the fitness function, (ii) population, it is fixed size group of chromosomes, (iii) selection operator, which is used to give preference to the better chromosomes, (iv) genetic operators like crossover and mutation operator. The genes identified using
genetic algorithm is biologically relevant [34]. Table 2 gives the details of various feature selection approaches used in literature. The work discussed above focuses on feature selection using gene expression data. The accuracy and effectiveness of the feature selection approach is based on parameters like selected number of genes, classification performance, and computing time. All these papers use one or more of the above parameters to find the effectiveness of the feature selection approach described.

Table 2. Literature details of feature selection approaches

<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Feature Selection approach</th>
<th>Dataset(s)</th>
<th>Classifier/Clustering Algorithm</th>
<th>Validatn process</th>
<th>Author, Year</th>
<th>Feature Selection approach</th>
<th>Dataset(s)</th>
<th>Classifier/Clustering Algorithm</th>
<th>Validatn process</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Fattah et al., 2013) [19]</td>
<td>TOPSIS and F-score</td>
<td>Colon</td>
<td>KNN, J48, SVM, NB</td>
<td>10-fold cross validation</td>
<td>(Subhajit Kar et al. 2015) [24]</td>
<td>PSO(Particle Swarm Optimization)</td>
<td>SRBCT, ALL_AML, MLL</td>
<td>Adaptive KNN</td>
<td>K-fold cross validation</td>
</tr>
<tr>
<td>(Rania Ibrahim et al., 2014) [20]</td>
<td>(MLFS) Multi level feature selection approach</td>
<td>Breast, HCC, Lung, Prostate, Colon, Ovarian</td>
<td>SVM and Random forests (RFs)</td>
<td>10-fold cross validation</td>
<td>(K.Dheenathayalan and J.Ram Singh 2014) [34]</td>
<td>Genetic Algorithm</td>
<td>Lymphoma, MLL, K-means clustering</td>
<td>WC error, Cluster purity</td>
<td></td>
</tr>
</tbody>
</table>

4. Conclusion

This paper provides review of feature selection techniques using gene expression data. It explains the DNA microarray, the problems involved in the process of feature selection including high dimensional data, irrelevant, redundant and noisy data. It provides the details of feature selection approaches using supervised, unsupervised and semi-supervised methods. In general, it is observed that a lot of work is being done in supervised feature selection methods mainly using filter evaluation framework. Comparatively little work is done in unsupervised and semi-supervised feature selection methods, but they also provide promising results with unlabelled or partially labeled data. Gene selection in which genes are taken as features is an important task as the subset of informative and
promising genes obtained after applying different feature selection methods are used by different classifiers to achieve best classification accuracy, disease prediction etc. This paper discusses the classification of feature selection methods stating its advantages and disadvantages. Filter approach is used by many supervised and unsupervised feature selection methods due to its generalization property and independence of classifier. Many researchers put huge and productive efforts in filter, wrapper and embedded methods but very few researches worked on ensemble and hybrid methods. Ensemble method due to its robustness and flexibility over high dimensional data, and hybrid method which combines the advantages of both filter and wrapper proves to give propitious results. Thus the advancement of hybrid and ensemble methods for feature selection can be considered auspicious.

References

[13] Innocent, Neethu, and Mathew Kurian. "SURVEY ON SEMI SUPERVISED CLASSIFICATION METHODS AND FEATURE SELECTION.
[24] Kar, Subhajit, Kaushik Das Sharma, and Madhubanti


