International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume V, Issue X, October - 2018

Feature Selection Methods: A Review on Gene Selection

Priya K
Department of Information Technology
Kannur University
kodothpriya@gmail.com

Dr. N K Narayanan
Principal
College of Engineering, Vadakara
nknarayanan@gmail.com

Abstract- . *Feature selection based on dimensionality reduction is an important issue in a wide range of scientific disciplines. Dimensionality reduction technique which is commonly used in the field of machine learning, pattern recognition, data mining etc. This technique help to select relevant feature from original set of features. Many approaches for dimensionality reduction have been proposed. This paper present a review on the current and relevant feature selection researches in gene expression microarray analysis.*

1. INTRODUCTION

Feature selection is used in the domains where the datasets comprise of thousands of features but with relatively small sample size (e.g., gene expression data). Feature selection that is applied to gene expression data is also known as gene selection [1]. Gene selection is necessary as the data usually contains many irrelevant, redundant and noisy expressions, and also is effective for early tumor detection and cancer discovery as it leads to a more reliable cancer diagnosis or prognosis and a better clinical treatment [1].

The methods for gene selection are broadly divided into three categories: filter, wrapper and embedded methods [2]. A filter method relies on general characteristics of the training data to select genes without involving any classifier for evaluation [2]. Most filter methods consider each feature separately with ignoring feature dependencies, which may lead to worse classification performance when compared with other types of feature selection methods [2]. In addition to considering feature dependencies, the wrapper methods take into account the interaction between feature subset search and model selection. However, the wrapper methods have a higher risk of over fitting than filter ones and are very computationally intensive [3]. Embedded methods have the advantage that they include the interaction with the classification model, while being far less computationally intensive than wrapper methods [4]. The wrapper and embedded methods consider the features' relevance by evaluating their utility for achieving accurate predication or exploiting data variance and distribution, and the selected genes are usually poorly explicable

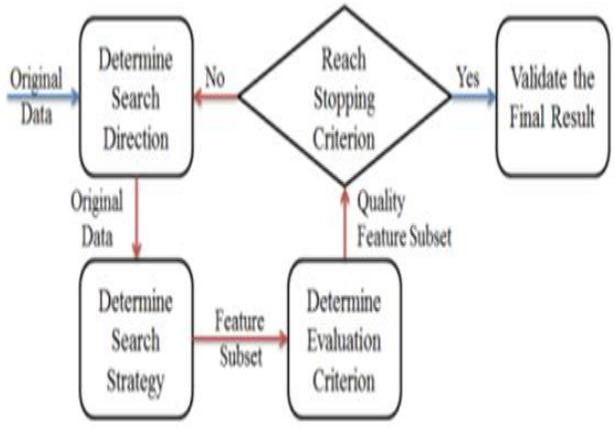
2. OVERVIEW OF FEATURE SELECTION

Feature selection provides many benefits as it improves prediction performance, understandability, scalability, and generalization capability of the classifiers. It also reduces computational complexity and storage, provides faster and more cost-effective model, and plays an important role in knowledge discovery. Moreover, it offers new insights for determining the most relevant or informative features.

In machine learning, a feature vector is an n dimensional vector of numerical values that represents one sample. The vector space associated with these vectors is often called the feature space. In order to reduce the dimensionality of the feature space, feature extraction or feature selection techniques can be

employed. Feature selection can be considered as the special case of feature extraction. Feature extraction is a technique that transforms the original feature space into a distinct space with different set of axes in order to reduce the dimensionality of the data. Feature selection reduces the original feature space into a subspace without transformation.

Feature selection aims to select a feature subset from the original set of features based on feature's relevance and redundancy. The process of selecting a subset of relevant and informative features from the original set of features consists of five main stages [1].



Stage 1: Determine search direction

The first stage is to determine the starting point and the search direction. The search process can be started with an empty set and then is followed by successively adding new features into the set in each iteration. This strategy is called forward search. In contrast, the search process can be started with a full set and then the features are eliminated consecutively from the set in each iteration. This strategy is called backward elimination search. Another alternative is to start with both ends by simultaneously adding and removing features in each iteration. This strategy is called bi-directional search. The search process may also begin somewhere in the middle by randomly selecting features to form the subset.

Stage 2: Determine search strategy

According to Gheyas and Smith [5], a good search strategy should provide good global search capability, rapid convergence to near optimal solution, good local search ability, and high computational efficiency. Search strategies can be categorized into four groups: exponential, sequential, and randomized.

Exponential search, also called complete search, is the most exhaustive global search strategy. It starts from the original feature set and guarantees to find the optimal result. However, this strategy is generally impractical and computationally intensive especially for high dimensional data sets, and prohibitive and intractable for all but a small initial number of features. An example of this strategy is exhaustive search, a search that evaluates all possible subsets to find the optimal subset.

Sequential search, also called greedy hill climbing search, adds or removes one feature at a time. The most common sequential strategies are sequential forward selection (SFS) and sequential backward selection (SBS). It is relatively simple to implement, its complexity is polynomial with respect to the number of features, and it is robust to multi col-linearity problems. However, these methods perform poorly on non-monotonic indices and may cause nesting effect [1] because once a feature is added (or deleted), it is not allowed to be deleted (or added) latter. Moreover, they are sensitive to feature interaction, so that they can easily be trapped into local minima [10]. Sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS) [1] were developed to overcome these problems by providing mechanisms to re-select the deleted features and delete the already added features. Some other examples of sequential search strategy are best first search, beam search (an optimized solution of best first search), and plus l take-away r algorithm (PTA) [1].

Randomized search strategy starts by randomly selecting the features and then proceeds with two different search strategies. The first uses the

classical sequential or bi-directional search, e.g., simulated annealing [1] and random hill climbing [1]. And the second uses strategies that have no regular movements, e.g., genetic algorithm (GA) [1], Las Vegas algorithm [1], and Tabu search [1]. The second strategies can escape local optima, but they have a greater chance of producing incorrect results due to no mechanism for capturing the relationship between the features.

Stage 3: Determine evaluation criterion

Originally evaluation methods in feature selection are classified into four types: filter, wrapper, embedded, and hybrid [6]. In recent years, another kind of evaluation method is developed, i.e., ensemble feature selection [7].

Filter or also called open-loop method is the earliest method. It examines the features based on the intrinsic characteristics prior to the learning tasks. A filter algorithm principally measures the feature characteristics based on four types of evaluation criteria, i.e., dependency, information, distance, and consistency. Most filter methods in literature are univariate. They are known to be very efficient and computationally faster hence more easily scale up to huge databases than wrapper methods. Filter methods are independent of any learning algorithm, therefore it can provide general solutions for various classifiers. Also the bias in the feature selection does not correlate with the bias in the learning algorithm, so it has a better generalization property [8]. However, filter methods ignore the interactions between classifiers and the possible interaction among features (combined features may have net effect that is not necessarily reflected by the individual features in that group). It also leads to varied prediction performance when the selected features are applied to different learning algorithms [9]. For reference, Lazar [7] reviewed the filter methods for feature selection in the gene microarray analysis.

Wrapper or close-loop method wraps the feature selection around the learning algorithm and utilizes classification error rate or performance accuracy as feature evaluation criterion. It selects the most discriminative subset of features by minimizing the prediction error of a particular classifier. This method often gives better performance results compared to the filter method because it takes into account the feature dependencies and directly incorporates bias in the learning algorithm. However, it is less general than the filter method because it must be re-executed if another learning algorithm is utilized. So, there is no guarantee that the solution is optimal for other learning algorithms. Furthermore, wrapper method is more prone to over-fitting than the filter method because the classifier is repeatedly called to evaluate each subset. The majority of wrapper methods are multivariate, hence they require extensive computation times to achieve the convergences and can be intractable for large data sets.

Embedded method is a built-in feature selection mechanism that embeds the feature selection in the learning algorithm and uses its properties to guide feature evaluation. Embedded method is more efficient and computationally more tractable than wrapper method while maintaining similar performance. This is because the embedded method avoids the repetitive execution of classifier and examination of every feature subset. Moreover, this method has lower risk to over-fitting compared to wrapper method. Like wrapper, embedded method takes into account the dependencies among features, but is only specific to a given learning algorithm [10]. However the computational complexity is the major issue, especially in high-dimensional data.

Hybrid and ensemble methods represent the latest developments in feature selection. Hybrid method can be either formed by combining two different methods (e.g. filter and wrapper), two methods of the same criterion, or two feature selection approaches. Hybrid method attempts to inherit the advantages of both methods by combining their complementary strengths [11]. It uses different evaluation criteria in different search stages to improve the efficiency and prediction performance with better

computational performance. The most common hybrid method is the combination of filter and wrapper methods [9].

Ensemble method is a method that aims to construct a group of feature subsets and then produce an aggregated result out of the group [12]. It is purposely designed to tackle the instability and perturbation issues in many feature selection algorithms. This method is based on different subsampling strategies where a particular feature selection method is run on a number of subsamples and the obtained features are merged to form a more stable subset. The performance of feature selection is no longer depending on a single selected subset, thus it is more flexible and robust when dealing with high dimensional data. Moreover, ensemble method provides a better approximation to the optimal subset or ranking of features by aggregating the outputs of several feature selectors. A detailed discussion on ensemble feature selection can be found in [13].

3 .A REVIEW ON GENE SELECTION

Gene selection is equivalent to feature selection in pattern recognition and machine learning. Many feature selection methods can be easily adapted to select genes. Among the many methods, feature selection based on large margin nearest neighbor has been an active research area in the past decade

Yijun2010 [14] provided a principled way to perform feature selection for classification problems with complex data distributions and very high data dimensionality. In this paper, a hybrid approach of filter and wrapper method was adopted wherein the number of features is first reduced by using a filter method and then a wrapper method is applied to the reduced feature set. Therefore, it is capable of processing many thousands of features within minutes on a personal computer while maintaining a very high accuracy that is nearly insensitive to a growing number of irrelevant features.

Jagath2013 [15] proposed to decompose multiclass ranking statistics into class specific ranking statistics and then use Pareto-front analysis for selection of genes. . The use of Pareto-front analysis is demonstrated on two filter criteria commonly used for gene selection: F-score and KW-score. F-score and KW-score are two popular gene ranking methods used in microarray data analysis, which are based on F statistics and KW statistics. F-score measures the ratio between the intraclass and the interclass distances of gene expression values. KW-score is based on the nonparametric KW statistic that uses the rankings of gene expressions instead of their values. The PFA simultaneously looks for optimal class specific statistics non-dominated by others. A gene i is said to dominate another gene j if the class-statistic value of gene i is better than that of gene j. This approach only improves the classification accuracy but also keeps the diversity among selected genes in the classes. With level of class dominance 2, the performance of F-PFA weakens because overall mean across all classes vary and, therefore, class-wise statistics are affected. However, KW-PFA is able to select features from other classes even at high class dominance because of its independence from class distributions.

Liao2014 [16] proposed a method for gene expression analysis called locality sensitive Laplacian score (LSLS). The LSLS method belong to three families, namely, local margin-based feature selection method, spectra graph-based feature selection method (SPFS). In LSLS, two parameters should be initialized: the number of nearest neighbor k and weight parameter α and a cross-validation approach. They evaluate the performance of LSLS with different values of k and α by leave-one-out cross validation, namely, k varying from 3 to 20 with an increment value of 1 and α varying from 0.1 to 0.9 with an increment value of 0.1, respectively. Apparently, LSLS can achieve higher predictive accuracy on various data set as k increases. However, the computational complexity of two

support vector machine (SVM) based on wrapper methods is significantly more expensive than that of filter methods.

Narges2015 [17] presented an effective and practical method for local feature selection for application to the data classification problem. Unlike most feature selection algorithms which pick a "global" subset of features which is most representative for the given data set, the proposed algorithm instead picks "local" subsets of features that are most informative for the small region around the data points. The cardinality and identity of the feature sets can vary from data point to data point. The process of computing a feature set for each region is independent of the others and can be performed in parallel. The local weight is assigned for each sample is based on the corresponding distance where, in order to concentrate on neighboring samples and reduce the effect of remote samples on the objective functions, higher weights are assigned to the closer samples. Weights decrease exponentially with increasing distance from samples. However, measuring sample distances from sample is a challenging issue since these distances should be measured in the local co–ordinate system defined by indicator vector f (i), which is unknown at the problem outset

Jian2016 [18] propose a new group-oriented and mutual information (MI)-based feature selection approach for microarray data. They use a strategy called relevance boosting, which requires a desirable feature to show substantially additional relevance to class labeling beyond the already selected features, which means it measure the MI shared between to gene .And the second strategy is called Feature Interaction Enhancing, which probabilistically compensates for feature interaction missing from simple aggregation. This strategy gives an additional piece of information to what both genes are already knows about each other. This additional information may contain a portion called interacting information.

Shuai2016 [19] proposed a new feature weighting approach which is based on the local nearest neighbors for gene selection, called LNNFW. The goal of LNNFW is to shrink the distances between the target neighbors and magnify the distances between the local differently labeled instances. This goal is achieved by minimizing the cost function as well as indirectly maximizing the margin of nearest neighbor rule. Experimental results on the UCI and the open microarray data sets show that LNNFW outperforms other classic feature selection methods on the excellent robustness to noisy features, the good ability to select informative features and the high classification accuracy. Feature selection based on large margin nearest neighbor has attracted more attention in recent years.

CONCLUSION

This paper provides a review on the current and relevant feature selection researches in gene expression microarray analysis. A very large number of gene selection approaches have been designed by researchers, yet this paper implies that there are still many open opportunities for further improvement as discussed below. Another promising direction for gene selection is the development of hybrid and ensemble frameworks to enhance the robustness of the selected feature subsets. Hybrid method is developed by combining two or more evaluation criteria. And ensemble method is developed by aggregating the results out of the group. The characteristics of these two methods are specifically more flexible and efficient in dealing with high dimensional data and help to achieve high classification accuracy. Unfortunately, there are not many theoretical or empirical works that study the hybrid or ensemble approaches in gene expression analysis.

REFERENCE

1. Ang Jun Chin, Andri Mirzal, Habibollah Haron, Senior Member, IEEE, Haza Nuzly Abdull Hamed "Supervised, Unsupervised and Semisupervised Feature Selection: A Review on Gene Selection ". Citation information: DOI 10.1109/TCBB.2015.2478454, IEEE/ACM Transactions on Computational Biology and Bioinformatics
2. Y. Saeys, I. Inza, P. Larranaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp.2507–2517, 2007.
3. S. Maldonado, R. Weber, "A wrapper method for feature selection using Support Vector Machines, " Information Sciences, vol. 179, pp. 2208-2217, 2009.
4. P. Bermejo, J.A. Gamez, J.M. Puerta, "A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high dimensional datasets, " Pattern Recognition, vol. 32, pp. 701-711, 2011.
5. I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," Pattern Recognition., vol. 43, no. 1, pp. 5– 13, Jan. 2010.
6. Y. Leung and Y. Hung, "A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification," IEEEACM Trans Computational Biology Bioinformatics. vol. 7, no. 1, pp. 108–117, Jan. 2010.
7.] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis," IEEEACM Trans Computational Biology Bioinformatics., vol. 9, no. 4, pp. 1106–1119, Jul. 2012.
8. C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," J. Computational Biology Bioinformatics. vol. 03, no. 02, pp. 185–205, Apr. 2005.
9. Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification," J. Biomed. Inform., vol. 43, no. 1, pp. 15–23, Feb. 2010.
10. Y. Saeys, "Feature selection for classification of nucleic acid sequences," Phd Dissertation, Ghent University, 2004.
11. M. Monirul Kabir, M. Monirul Islam, and K. Murase, "A new wrapper feature selection approach using neural network," Neurocomputing, vol. 73, no. 16–18, pp. 3273–3283, Oct. 2010.
12. Q. Shen, R. Diao, and P. Su, "Feature Selection Ensemble," in Turing-100, 2012, vol. 10, pp. 289–306.
13. W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in 2012 IEEE 13th International Conference on Information Reuse and Integration (IRI), 2012, pp. 356–363.
14. Yijun Sun, Sinisa Todorovic, and Steve Goodison." Local-Learning-Based Feature Selection for High-Dimensional Data Analysis." IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 32, NO. 9, SEPTEMBER 2010
15. Jagath C. Rajapakse and Piyushkumar A. Mundra." Multiclass Gene Selection Using Pareto-Fronts". IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 10, NO. 1, JANUARY/FEBRUARY 2013
16. Bo Liao∗, Yan Jiang, Wei Liang, Wen Zhu, Lijun Cai, Zhi Cao." Gene selection using locality sensitive Laplacian score". IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. XXX, NO. XXX, XXX 2014
17. Narges Armanfard, James P. Reilly, Majid Komeili. "Local Feature Selection for Data Classification". Citation information: DOI 10.1109/TPAMI.2015.2478471, IEEE Transactions on Pattern Analysis and Machine Intelligence
18. Jian Tang and Shuigeng Zhou." A New Approach for Feature Selection from Microarray Data Based on Mutual Information", Citation information: DOI 10.1109/TCBB.2016.2515582, IEEE/ACM Transactions on Computational Biology and Bioinformatics
19. Shuai An, Jun Wang, and Jinmao Wei." Local-Nearest-Neighbors-Based Feature Weighting for Gene Selection". IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 2016