A Survey on Improved Filtering Techniques for Multiclass Gene Selection

G. V. Manoharan
Research Scholar
Faculty of Information and Communication Engineering
Anna University, Chennai.

R. Shanmugalakshmi, Ph.D
Associate Professor
Department of CSE
Government College of Technology
Coimbatore, INDIA.

ABSTRACT
In the field of bioinformatics, selection of genes in multiclass sample classification can be done by filtering methods using microarray data. Such approaches usually contribute to bias towards a few classes that are easily recognizable from other classes due to imbalances of strong features and sample sizes of distinct classes in a microarray data. Many methods have been used for the filter methods, as they are very commonly used in gene ranking from microarray data in multiclass problems. In this research, we discuss various methods to decompose multiclass ranking statistics into class specific statistics and then need of Pareto-front analysis for selection of genes. This mitigates the bias induced by class intrinsic characteristics of dominating classes. The need of Pareto-front analysis is to indicate on two filter criteria commonly used for gene selection: F-score and KW-score. A significant development in classification performance and reduction in redundancy among top ranked genes were achieved in experiments with both synthetic and real-benchmark data sets. The following work is analysis over the traditional and improved filter methods used for gene selection of various classes through various mechanisms available in the literature.

Keywords
Aggregation statistics, filter methods, gene selection, multiobjective evolutionary optimization, Pareto-front analysis.

1. INTRODUCTION
The selection of related genes has evolved into integral part of determinate the microarray gene-expression data sets as it not only enhances the recognition of types of samples but also supports curious in biological insights. Approaches to select crucial genes (or marker genes) are generally classified into wrapper, filter or embedded approaches [1], [2], [3]. In this work, we focus on filter techniques, as they are very commonly used in gene ranking from microarray data in multiclass issues because of their simplicity, ease of use, and computational efficiency. Such techniques rank genes by looking at the correlation of gene expressions with respect to class labels. For a comparative review of multiclass classification and various feature selection approaches, the reader is mention in [4], [5].

Many filters test such as F-score [6], Kruskal-Wallis (KW)-score [7], mutual information [8], and entropy [9], and so on, have been projected for multiclass gene selection. Fscore is based on F-statistics and described as the proportion between-within-class sum of squares and class sum of squares [6]. KW-score is a nonparametric filter measure calculated using the ranks of gene expression values; it is the square of variances between within-class average ranks and overall mean of the ranks. Brown-Forsythe criteria statistic and Welch criteria statistic have been presents to relax the assumption of equal variance across the classes [7].

Furthermore, many analytical gene ranking test for multiclass issues are generally for generalization of criterion for two-class issues. Such a generalization regularly includes aggregation of test statistics of individual classes to achieve multiclass ranking statistics. However, class-specific statistics of some classes may manage the overall ranking of the aggregation method that leads to the siren pitfall in gene ranking and eventually adversely infecting the classification performance. In addition, the imbalance in sample sizes of different classes aggravates the problem. This survey discusses a best technique to addresses a general siren-pitfall issue in the context of gene selection. In this paper explain about efficient method of Pareto-Fronts, we first determine how KW-score and F-score are affected by siren pitfalls by utilizing simulated data sets. To overcome such pitfalls in gene selection in a multiclass classification issue, we look at these statistical tests from a multiobjective aspect and present a Pareto-front analysis (PFA)-based method.

We present a PFA [16] method as a general approach to handle a siren-pitfall issue in filter criteria utilizing aggregation statistics for multiclass gene selection. Along with F-score, we additionally demonstrate the efficiency of the PFA algorithm using over KW-scores, and they are distinguished over several simulated. In extension to accuracy, we also inquire into the stability and redundancy of the proposed approaches. Furthermore, we evaluate the differences of outcome with F-score and KW-score based methods using skewness and kurtosis and conclude about when KW-score with PFA (KW-PFA) should be preferred over F-score with PFA (F-PFA).

In this technique to select genes, depending on their class-specific statistics to become better the siren pitfall problem in filter test using aggregation statistics. Each class-specific statistic is assumed as an objective and genes are elected by improving multiple class-specific statistics instead of improving the summation of statistics. This avoids domination of election of genes by a few strong class-specific statistics. In general, each element in the set $\{r_{il}\}_{i=1,l=1}^{n,L}$of class-specific statistics decides the significance of a specific gene in recognize a specific class. Totally, we have $n \times L$ statistics belonging to L classes. To improve a good classification of all the classes, a set of genes achieving individual class-particular statistics is demanded. Therefore, genes are elected by recognizing the set of class-specific statistics that are nondominant by the rest of the class-specific statistics.

In [16] C. Jagath presented PFA begins by dividing all the genes into different fronts by using class-specific statistics. A Pareto front is a set of genes that do not dominate one another. A gene i is said to dominate another gene i' if both of the following conditions are satisfied:

- The class-specific statistic value of gene i is no worse than that of gene i' in all classes.

- At least one of the class-specific statistic value of gene i is better than that of gene i'.

If $i < i'$denotes that gene i is dominating gene i', then

$i < i' \Leftrightarrow r_{il} \geq r_{i'l}, \forall l, and\ r_{il} \geq r_{i'l}, \exists l,$ Where l denotes a class label. When neither of the genes dominates the other, a pair of genes is known to be mutually nondominating. In those genes, none of the genes are dominated by any other gene in the set. To decide the Pareto fronts of genes from a given set of class-specific statistics, one can perform all pairwise distinguishes and determine the nondominated genes. The PFA divides genes into ordered sets of Pareto fronts of class specific statistics. If genes in a Pareto front are nondominated by the other genes, they are said to form a Pareto-optimal set. In the context of gene selection, the Pareto-optimal set of genes includes the most important genes for the classification as those genes have at least one (or more) class-specific statistics better than that of genes in the other fronts. More methods have been surveyed based on filter method which is discussed in the following section.

2. METHODS USED FOR FILTER TECHNIQUES

Filtering approaches select features based on discriminating criteria which are relatively independent of classification. Various filtering methods use simple correlation coefficients similar to Fisher's discriminant criterion in [10]. Several filter-based methods which are used earlier evaluated features in isolation and correlation between features are considered below. The Table 1 summarized the related work in terms of using Filter Techniques.

2.1 Conventional Filter Techniques

2.1.1 F-score

In [11] Piyushkumar A. Mundra presented F-score as filtering criteria for selecting and classifying multiclass genes. In [11] F-score is a mainly used for gene selection in multiclass cancer classification. This ranking based criterion might become biased towards classes that have excess of between-class sum of squares which results in inferior classification performance. Fscore is widely based on F-statistics and learned as the distinct ratio of within-class sum of squares and between-class sum of squares.F-test is a test strategy that is widely used to test the hypothesis in which means of different classes are the same by assuming that all the classes have same variance. By Based on such a test strategy, F-score measures the ratio between the interclass and the intraclass distances of gene expression values.

2.1.2 KW-score

In [7] D. Chen presented KW-score (Kruskal-Wall Score) for gene selection which is a nonparametric filter criterion computed using the ranks of gene expression values. These values are considered as the square of differences between overall mean of the ranks and within-class average ranks. In this KW score are decomposed into class components. Then class specific statistics of genes for different classes are defined which actually differtiates the genes. Brown-Forsythe test statistic and Welch test statistic are used in [7] for the assumption of equal variance across the classes. These test statistics were studied to provide the multiclass prediction results of gene selection efficiently.

2.1.3 Mutual information

In [8] C. Ding Author presented minimum redundancy criteria are described by the usual maximum relevance criteria such as maximal mutual information. The class of gene is mainly represented by expressions under different classes in a random or uniform manner. The mutual information within these classes is zero. Larger mutual information between classes depicts genes are strongly differentiated. Thus, the method in [8] used mutual information as a measure of relevance of genes.

2.1.4 Entropy

In [9] X. Liu presents an entropy-based method for gene selection from microarray data. The method in [9] used for these datas classifies various cancer sub-types with high accuracy. Additionally, compact feature set obtained and the redundancy between genes is reduced to a large scope. Thus the result of the discussed method implies that the classifiers can be used with a smaller subset of genes.

2.2 Improved Filtering criteria for Multiclass gene selection

2.2.1 Ranking Score using Centroid –based method

In [12] Q. Shen Author presented a novel statistical parameter is used for the suitability score to filter gene and classify cancer. This method is similar to classification to nearest centroids, the suitability score method ranked and selected gene for each cancer type and then employed the nearest centroid classification to classify cancer. Gene selection by suitability score and the nearest centroid classifier are all based on class centroid, so the classification result is satisfied. The gene selection is an important feature for class identification and it is significant to determine whether a gene is differentially expressed among different classes. If a gene is appropriate for classification, the appearance of this genes for the samples belonging to the same class are close and the appearance of this genes for the samples belonging to the different class may differ greatly. If a gene is appropriate for classification, the appearance of this genes for the samples belonging to the same class are close and the appearance of this genes for the samples belonging to the different class may differ significantly. The essential idea of this method is to decide that genes vary most appreciably among classes and to choose genes for each cancer type. Utilizing the novel statistical parameter, the principle of selection gene for class k is to find genes that have short distances from centroid of class k and have large distances from other class centroid.

2.2.2 Generalization Approach

In [13] Y.-S. Tsai presents an innovative generalizations of SNR for multiclass cancer discrimination is proposed through the beginning of two indices, one is called Gene Dominant Index and another one is called Gene Dormant Index (GDIs) . These two indices lead to the concepts of dominant and dormant genes with biological significance. Actually, the Signal-to-noise ratio is frequently used for identification of biomarkers for two-class problems and no formal and useful generalization of SNR is available for multiclass problems. By utilizing the two indices, to develop methodologies for discovery of dominant and dormant biomarkers with interesting biological significance. By utilizing the scatter plot of individual gene and 2-D Sammon's projection of the selected set of genes, the dominancy and dormancy of the identified biomarkers and their brilliant discerning power are also established pictorially. Using the information that the Gene Dormant Index based method can identify dominant and dormant genes that play important roles in cancer biology. These biomarkers are also utilized to design diagnostic prediction systems.

2.2.3 Classification of dormant and dominant gene

In [14] K. Kadota presents a new fold-change (FC)-based method is used, the weighted average difference method (WAD), for ranking differentially expressed genes(DEGs). It utilizes the average difference and relative average signal intensity so that highly expressed genes are highly ranked on the average for the different conditions. The methods for ranking genes in accordance with their degrees of differential expression can be classified into *t*-statistic-based methods and fold-change (FC)-based methods. Both types are usually used for choosing DEGs with two classes. Both of them have particular drawbacks. The *t*-statistic-based gene ranking is lacking because a gene with a small fold change can have a very large statistic for ranking, because of the *t*-statistic possibly having a very small denominator. The fold-change ranking is deficient because a gene with larger variances has a higher probability of having a larger statistic.

2.2.4 Redundancy removal

In [15] C. Ooi Author presented Because of the large number of genes in a typical microarray dataset, feature selection looks set to play an important role in reducing noise and computational cost in gene expression based tissue classification while improving accuracy at the same time. Numerous feature selection techniques applied on microarray datasets are either rank-based and hence do not take into account correlations between genes, or are wrapper-based, which require high computational cost, and often yield difficult-to-reproduce results. The two realistically evaluated correlation-based feature selection techniques which incorporate, in addition to the two existing criteria involved in forming a predictor set (relevance and redundancy), a third criterion called the degree of differential prioritization (DDP). DDP functions as a parameter to strike the balance between relevance and redundancy, providing our techniques with the novel ability to differentially prioritize the optimization of relevance against redundancy.

Table 1: Related work for Filter Techniques

Author	Feature
Piyushkumar et al. [11]	Propose an F-score as filtering criteria for selecting and classifying multiclass genes. F-score is a mainly used for gene selection in multiclass cancer classification.
Chen et al. [7]	Propose a KW-score (Kruskal-Wall Score) for gene selection which is a nonparametric filter criterion computed using the ranks of gene expression values. These values are considered as the square of differences between overall mean of the ranks and within-class average ranks.
Ding et al. [8]	Propose a minimum redundancy criteria are described by the usual maximum relevance criteria such as maximal mutual information.
Liu et al. [9]	Propose an entropy-based method for gene selection from microarray data. The method used for these datas classifies various cancer sub-types with high accuracy.
Shen et al. [12]	Propose a novel statistical parameter is used for the suitability score to filter gene and classify cancer. This method is similar to classification to nearest centroids, the suitability score method ranked and selected gene for each cancer type and then employed the nearest centroid classification to classify cancer.
Tsai et al. [13]	Propose an innovative generalizations of SNR for multiclass cancer discrimination is proposed through the beginning of two indices, one is called Gene Dominant Index and another one is called Gene Dormant Index (GDIs).
Kadota et al. [14]	Propose a new fold-change (FC)-based method is used, the weighted average difference method (WAD), for ranking differentially expressed genes(DEGs). It utilizes the average difference and relative average signal intensity so that highly expressed genes are highly ranked on the average for the different conditions.
Ooi et al. [15]	Propose Numerous feature selection techniques applied on microarray datasets are either rank-based and hence do not take into account correlations between genes, or are wrapper-based, which require high computational cost, and often yield difficult-to-reproduce results.

3. CONCLUSION

In the above survey various filtering methods for multiclass gene selection has been discussed. The approach mentioned above are F.Score, KW-Score, mutual information, Entropy and some improved filtering methods such as centroid based, Generalization approach, Classification of dormant and dominant gene and Redundancy removal approach has been discussed. Each of the above surveyed methods proves and shows better in some categories and not in some other categories. In order to give better performance in the selection of gene, proposed method of PFA outperforms in selection category rather than surveyed methods. The PFA simultaneously looks for optimal class specific statistics nondominated by others. This not only improves the classification accuracy but also keeps the diversity among selected genes in the classes. The PFA method has been used for gene ranking in multiclass tissue classification problem. By decomposing F-score and KW-score into class-specific statistics, ranking is achieved by simultaneously finding the dominating class specific statistics. The PFAs render minimally redundant gene subsets with achieving the performance in classification while compromising the stability.

4. REFERENCE

[1] Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

[2] K.-B. Duan, J.C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data," IEEE Trans. Nanobiosciences, vol. 4, no. 3,pp. 228-234, Sept. 2005.

[3] P.A. Mundra and J.C. Rajapakse, "SVM-RFE with MRMR Filter for Gene Selection," IEEE Trans. Nanobiosciences, vol. 9, no. 1, pp. 31-37, Mar. 2010.

[4] 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," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 9, no. 4, pp. 1106-1119, July/Aug.2012.

[5] T. Li, C. Zhang, and M. Ogihara, "A Comparative Study of Feature Selection and Multiclass Calssification Methods for Tissue Classification Based on Gene Expression," Bioinformatics, vol. 20, no. 15, pp. 2429-2437, 2004.

[6] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," J Am. Statistical Assoc., vol. 97, no. 457, pp. 77-86, 2002.

[7] D. Chen, Z. Liu, X. Ma, and D. Hua, "Selecting Genes by Test Statistics," J. Biomedicine and Biotechnology, vol. 2, pp. 132-138, 2005.

[8] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," J. Bioinformatics Computational Biology, vol. 3, pp. 185-205, 2005.

[9] X. Liu, A. Krishnan, and A. Mondry, "An Entropy-Based Gene Selection Method for Cancer Classification Using Microarray Data," BMC Bioinformatics, vol. 6, article 76, 2005.

[10] Golub,T. *et al*. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286, 531–537.

[11] Piyushkumar A. Mundra, Jagath C. Rajapakse," F-score with Pareto Front Analysis for Multiclass Gene Selection , Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics Lecture Notes in Computer Science Volume 5483, 2009, pp 56-67.

[12] Q. Shen, W.-M. Shi, and W. Kong, "New Gene Selection Method for Multiclass Tumor Classification by Class Centroid," J. Biomedical Informatics, vol. 42, no. 1, pp. 59-65, 2009.

[13] Y.-S. Tsai, C.-T. Lin, G. Tseng, I.-F. Chung, and N. Pal, "Discovery of Dominant and Dormant Genes from Expression Data Using a Novel Generalization of SNR for Multi-Class Problems," BMC Bioinformatics, vol. 9, article 425, 2008.

[14] K. Kadota, Y. Nakai, and K. Shimizu, "A Weighted Average Difference Method for Detecting Differentially Expressed Genes from Microarray Data," BMC Bioinformatics, vol. 3, article 8, 2008.

[15] C. Ooi, M. Chetty, and S. Teng, "Differential Prioritization between Relevance and Redundancy in Correlation-Based Feature Selection Techniques for Multiclass Gene Expression Data," BMC Bioinformatics, vol. 7, article 320, 2006.

[16] 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.