<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Leung, Y; Hung, Y</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Ieee/Acm Transactions On Computational Biology And Bioinformatics, 2010, v. 7 n. 1, p. 108-117</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2010</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/124745">http://hdl.handle.net/10722/124745</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.; IEEE - ACM Transactions on Computational Biology and Bioinformatics. Copyright © IEEE.; ©2010 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.</td>
</tr>
</tbody>
</table>
A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification

Yukyee Leung and Yeungsam Hung

Abstract—Filters and wrappers are two prevailing approaches for gene selection in microarray data analysis. Filters make use of statistical properties of each gene to represent its discriminating power between different classes. The computation is fast but the predictions are inaccurate. Wrappers make use of a chosen classifier to select genes by maximizing classification accuracy, but the computation burden is formidable. Filters and wrappers have been combined in previous studies to maximize the classification accuracy for a chosen classifier with respect to a filtered set of genes. The drawback of this single-filter-single-wrapper (SFSW) approach is that the classification accuracy is dependent on the choice of specific filter and wrapper. In this paper, a multiple-filter-multiple-wrapper (MFMW) approach is proposed that makes use of multiple filters and multiple wrappers to improve the accuracy and robustness of the classification, and to identify potential biomarker genes. Experiments based on six benchmark data sets show that the MFMW approach outperforms SFSW models (generated by all combinations of filters and wrappers used in the corresponding MFMW model) in all cases and for all six data sets. Some of MFMW-selected genes have been confirmed to be biomarkers or contribute to the development of particular cancers by other studies.

Index Terms—Filters, gene selection, hybrid classification models, microarray data classification, wrappers.

1 INTRODUCTION

The rapid advances of gene expression microarray technology enable the simultaneous monitoring of thousands of genes in a single experiment (referred to as a sample). With a certain number of samples, investigations can be made into whether there are patterns or dissimilarities across samples of different types, such as cancerous versus normal, or even within subtypes of diseases. The problem is referred to as sample classification.

Microarray sample classification has been studied extensively using classification techniques in machine learning and pattern recognition. Classification tools such as weighted voting (WV) [1], k-nearest neighbor (k-NN) [2], support vector machine (SVM) [3], and Fisher’s linear discriminant analysis (LDA) [4] have been used for microarray data classification. However, these tools have not been effective for identifying biomarkers, which are substances (genes in the present context) used for detecting whether a patient has got a particular disease or not [5], [6]. In a microarray chip, the number of genes available is far greater than that of samples, a well-known problem called the curse of dimensionality [7]. However, most genes in a microarray give little benefits to the sample classification problem. Therefore, prior to sample classification, it is important to perform gene selection whereby more interpretable genes are identified as biomarkers, so that a more efficient, accurate, and reliable performance in classification can be expected. These biomarkers may also be useful for assessing disease risk [6] and understanding the basic biology of a disorder [8].

There are, in general, two approaches to gene selection, namely filters and wrappers [9], [10]. The filter approach selects genes according to their discriminative powers with regard to the class labels of samples [11]. Methods such as Signal-to-Noise Ratio (SNR) [1], t-statistics (TS) [11], threshold number of misclassifications (TNoM) score [12], and F-test [13] have been shown to be effective scores for measuring the discriminative power of genes. A comparison of these methods can be found in [14]. In all cases, genes are ranked according to their statistical scores, and a certain number of the highest ranking genes are selected for the purpose of classification [15]. Although gene selection using filters are simple and fast, the method has several shortcomings:

1. The criterion used for gene selection in filters does not necessarily associate with the classifiers to be applied [13], [16].
2. The filter approach does not take into account correlation between genes, which reduces the usefulness of the selected genes for sample classification.
3. Despite reports that classifiers with few genes (less than 15-20) are able to achieve good performances [17], [18], there is little theoretical support for determining how many genes should be chosen for classification, and the number used is somewhat arbitrary [19].
The wrapper approach for gene selection goes some way to address all of the above problems of the filter approach. In the wrapper approach, genes are selected sequentially one by one so as to optimize the training accuracy of a particular classifier [10]. That is, the classifier is first trained using one single gene, and this training is performed for the entire original gene set. The gene that gives the highest training accuracy is selected. Then, a second gene is added to the selected gene and the gene that gives the highest training accuracy for the two-gene classifier is chosen. This process is continued until a sufficiently high accuracy is achieved with a certain gene subset. Contrary to the filter approach which selects genes with no consideration for the classifier, the wrapper approach selects genes that are “tailor-made” for a particular classifier. Furthermore, the process ensures that newly added genes are complementary rather than correlated with genes already selected. The method also provides a stopping criterion when a certain number of genes are sufficient for attaining the required accuracy. However, a major disadvantage of the wrapper approach is that its computation requirement is formidable [5], particularly if the original gene set is large. Because of this, wrappers are not frequently used in microarray data analysis [17], [20].

In view of the drawbacks of the filter and wrapper approaches, hybrid filter-wrapper models have been proposed [21], [22], [23] that take advantage of the simplicity of the filter approach for initial gene screening and then make use of the wrapper approach to optimize classification accuracy in final gene selection. In the hybrid model, a filter is first used to screen out a majority of (irrelevant) genes from the original set to give a filtered subset of a relatively small size (say, a few hundred from an original set of several thousands). Then, the wrapper is applied to select genes from the filtered subset to optimize the training accuracy. As the filter efficiently reduces the size of the gene set by an order of magnitude or more, the computations of the subsequent wrapper become acceptable. We will refer to this hybrid model as a single-filter-single-wrapper (SFSW) approach, as it is necessary to select a particular filter and a specific classifier in the process. The SFSW approach however has its own difficulties:

1. Different filters yield different filtered subsets that may leave out some relevant biomarkers which consequently do not have a chance to be considered in the wrapper evaluation.
2. Different wrappers will select different genes from the filtered set despite achieving the same training accuracy.
3. Some SFSW models are better than the others in terms of attaining the required training accuracy.

The last point can be addressed by trying different filter-wrapper combinations in the SFSW model to suit the data set. However, this is not a satisfactory approach, as the need to tune the choice of filter and classifier in the SFSW model suggests that the model is not robust to variations among data sets. Furthermore, the fact that different filters and wrappers would select different genes makes it doubtful whether the genes selected by one particular model are true biomarkers.

To address the above problems, we propose in this paper a multiple-filter-multiple-wrapper (MFMW) approach for the hybrid model. In the MFMW hybrid model, we consider using multiple filters to select genes and then combining them to provide a merged filtered subset of genes. The use of multiple filters with different filter metrics ensures that useful biomarkers are unlikely to be screened out in the initial filter stage. The use of multiple wrappers is intended to enhance the reliability of the classification by establishing consensus among several classifiers. As a result, there is some kind of consensus among the different classifiers in the wrapper step as to which genes should be selected. Hence, the final genes selected can be considered to be more robust with a mixture of characteristics that fit several wrappers, and are therefore better qualified as biomarkers. Furthermore, since the MFMW model already incorporates the characteristics of multiple filters and wrappers, it is no longer necessary to try different filter-wrapper combinations in order to search for a suitable combination that yields the highest classification accuracy. We will show that the proposed MFMW model provides predictive accuracies that are either comparable or better than the best existing results obtained using all available SFSW methods.

This paper is organized as follows: In Section 2, some preliminaries on filter metrics and classifiers are given. Our proposed algorithm MFMW is given in Section 3. In Section 4, the MFMW approach is evaluated by means of six benchmark data sets available in the public domain. Comparisons are made between the results obtained using the MFMW and different SFSW models. In Section 5, discussions are given on how many genes should be selected by each filters, how many filters and classifiers should be chosen in an MFMW model, and biological significance of MFMW-selected genes. Some concluding remarks are given in Section 6.

2 PRELIMINARIES

In a microarray data set $Z$, each sample is an $n$-vector storing the gene expression values, and may be considered as a data point in an $n$-dimensional space, where $n$ may take a value of several thousands. Each sample carries a specific label which is utilized for supervised learning. This means that $Z = [X|Y]$ can be partitioned into two groups $X$ and $Y$ which represent, for example, cancerous and normal samples, respectively. Hence, the $g$th row $Z_g = [X_g|Y_g]$ of $Z$ is the profile of gene $g$ across all samples.

2.1 Filter Metrics

Feature selection is a crucial and critical step in microarray data analysis for removing irrelevant data and reducing computational complexity. One way to do so is by filtering, whereby the “goodness” of a gene is evaluated by measuring the relationship between gene expression and the class label using statistical techniques [14]. Three of the most commonly used metrics are 1) SNR, 2) TS, and 3) Pearson Correlation Coefficient (PC).

2.2 Classifiers

The aim of supervised classification is to develop a decision rule to discriminate between samples of different
classes based on the gene expression profile. In this study, three widely used classifiers are considered, namely 1) WV, 2) \( k \)-NN, and 3) SVM.

WV define the vote of gene \( g \) as

\[
v_g = w_g \left( x_g - \frac{\mu_1 + \mu_2}{2} \right),
\]

where the weighting factor \( w_g \) is the SNR score of gene \( g \), \( x_g \) is the expression level of gene \( g \) in a sample \( S \), and \( \mu_1 \) and \( \mu_2 \) are the mean expression levels of gene \( g \) for samples of class 1 and class 2, respectively [1]. Assuming \( \mu_1 > \mu_2 \), \( v_g > 0 \) (respectively < 0) indicates a vote for class 1 (respectively 2). Let \( V_{\text{win}} \) and \( V_{\text{lose}} \) be the sums of votes for the winning and losing class, respectively. The prediction strength (PS) provides a measure of the margin of victory:

\[
PS = \frac{V_{\text{win}} - V_{\text{lose}}}{V_{\text{win}} + V_{\text{lose}}},
\]

If PS exceeds a certain threshold, the sample \( S \) is assigned to winning class, otherwise, sample \( S \) is unclassified.

\( k \)-NN classifies samples based on closest training examples in the gene space. All training samples are mapped into a multidimensional gene space which is partitioned into regions by class labels of the training samples. A point in the space is assigned to the class \( c \) if it is the most frequent class label among the \( k \) nearest training samples [24]. The choice of \( k \) depends upon the data—larger values of \( k \) reduce the effect of noise, but make boundaries between classes less distinct. We choose \( k = 3 \) in our study.

SVM—Vapnik [25] proposed the SVM for performing classification by locating the hyperplane that maximally separates two sets of points in an \( n \)-dimensional feature (e.g., genes) space \( R^n \). The SVM classifier then assigns a class label to a new sample according as which side of the hyperplane the sample lies.

### 3 Multiple-Filter-Multiple-Wrapper (MFMW) Model

Since the MFMW hybrid model is a generalization of the SFSW model, we will begin by introducing the structure of the SFSW model.

In the SFSW model shown in Fig. 1, the full set of genes \( G \) is first filtered by a filter (\( FT \)) using a particular filter metric to extract a subset of genes \( F \) that is much reduced in size compared to \( G \). The genes in \( F \) are then selected in an incremental manner using a chosen classifier (\( CF \)) with an aim to optimize the training accuracy. To start with, each gene in \( F \) is considered as a candidate for a single-gene classifier and the gene(s) giving the highest accuracy is/are identified and retained as the first gene in the set \( P \). Then, the remaining genes in \( F \) are combined with the first gene in \( P \) to give a two-gene classifier, and the second gene that together with the first selected gene yield the highest accuracy is identified. This process of adding genes to \( P \) is repeated until \( P \) contains a sufficient number of genes to achieve the required training accuracy. The final gene set \( P \) is then evaluated using either an independent data set or by performing Leave-One-Out Cross-Validation (LOOCV). In LOOCV, one sample from the data set is excluded, and the rest of the samples are used to train the classifier. The trained classifier is then used to predict the class label of the left-out sample, and this is performed for each sample in the data set. The LOOCV estimate of classification accuracy is defined as the overall number of correct classifications divided by the total number of samples in the data set. If independent testing and training data sets are available, the classifier can first be trained using the training data set and then applied to the testing data set. In this case, the testing accuracy represents the unbiased predictive power of the gene subset \( P \). Note that the above LOOCV strategy means that only the classification (as opposed to the entire process of gene filtering) is cross validated [26], [27].

In order to overcome the weaknesses of using SFSW as discussed in Section 1, we now propose the MFMW model shown in Fig. 2.
In the MFMW model, several (e.g., three in Fig. 2) filters \( FT_i(i = 1, 2, 3) \) are employed, each for selecting a pre-defined number of (e.g., a few hundred) genes. Different filters have their own characteristics and there is typically a fair proportion of both overlapping and nonoverlapping genes among the lists of genes selected by two different filters. The filtered gene subset \( F \) is formed by taking the union of the lists of genes obtained by all the filters. After that, the genes in \( F \) are selected by means of a wrapper consisting of multiple (e.g., three in Fig. 2) classifiers. Since different classifiers may give different classification labels for the same sample, there is a need to resolve this conflict when it arises. It is natural to resort to some kind of voting scheme among the classifiers. Two possibilities are majority voting and unanimous voting. The advantage of majority voting is that there will always be a decision if the number of classifiers is odd. However, the “winning” class label may win by only a small margin, in which case the classification output may not be reliable. For this reason, we have chosen to use unanimous voting to decide on the overall classification output based on the outputs of the classifiers. In the case where a unanimous vote cannot be reached, the classification output is regarded as indecisive (denoted by “X”). Table 1 illustrates the voting result for all possible combinations of the outputs of three classifiers in a two-class (with labels “A” and “B”) classification problem where the sample has a true class label “A.” Out of these combinations of classifier outputs, only two will produce a unanimous vote of class A or B, one of which is right and the other is wrong (with prediction status \( R \) and \( W \), respectively), and the other six produce an indecisive outcome (denoted as \( I \)).

By the unanimous voting scheme, each sample can be categorized as either \( R \), \( I \), or \( W \). This information will be used for gene selection in the MFMW model. The number of \( I \) and \( W \) prediction statuses across all samples will be used to determine the usefulness of the set of genes. A prediction status \( W \) implies that all classifiers misclassify the sample under consideration, and is therefore particularly undesirable. Our first objective in gene selection is therefore to minimize the number of \( W \), preferably to zero. Beyond that, the next goal will be to reduce the number of \( I \) (i.e., the indecisive cases). The procedure is summarized in the following algorithm:

**MFMW Algorithm:**

**FILTER**
1. Choose \( m \) = the number of genes to be selected by each filter.
2. For each filter \( FT_i(i = 1, 2, 3) \),
   a) Calculate the statistical scores for all genes and rank the scores from the highest to the lowest.
   b) Select \( m \) genes with top ranking scores in each list.
3. Take the union of the list of genes obtained by \( FT_i(i = 1, 2, 3) \) to give \( F \).

**WRAPPER**
1. Initialize \( P = \Phi \) (the set of selected genes) and \( n = 0 \) (number of genes in \( P \)).
2. Put \( n = n + 1 \); Repeat for all genes \( g_n \in F\setminus P \),
   a) Using \( P \cup \{g_n\} \) as the gene set, classify all samples using \( CF_i(i = 1, 2, 3) \) individually.
   b) Based on unanimous voting, determine the number of samples that receive \( W \) or \( I \) prediction statuses.
   Select the gene subset with the smallest number of \( W \).
   If more than one gene subset has the same number of \( W \), the one(s) with the smallest number of \( I \) is/are chosen. Update \( P \) by adding the selected gene to the set.
3. Repeat step 2 until the number of \( W \) and number of \( I \) are below a prescribed threshold, or cannot be reduced further.
4. Output the selected genes set(s) \( P \).

Ideally, both the number of \( W \) and the number of \( I \) should be zero at the end of the MFMW algorithm. In practice, this may not be achievable as the data set may contain outliers or corrupted samples, in which case a sufficiently small number of \( W \) and number of \( I \) will be regarded as acceptable. In any case, the algorithm will be terminated when first the number of \( W \) and then the number of \( I \) cannot be reduced further by adding more genes to the set \( P \). Hence, the number of iterations in the algorithm (i.e., the number of loopings between steps 2 and 3) is given by the number of genes selected in the final MFMW model.

**4 EXPERIMENTAL RESULTS**

The proposed MFMW model was evaluated by means of six DNA microarray data sets that have been used in the diagnosis of different cancers, namely LEU (LEU38 and LEU72) [1], COL62 [28], BR-ER49 [29], LYM77 [30], PROS102 [31], and LUNG181 [32]. A brief description of these data sets is given below with a summary in Table 2:
1. **LEU (LEU38 and LEU72) data set**—Seventy-two samples were analyzed with Affymetrix oligonucleotide arrays. The training set (LEU38) contains 38 samples (27 Acute Lymphoblastic Leukemia (ALL), 11 Acute Myeloid Leukemia (AML)), with 7,129 probes from 6,817 genes. The testing set contains an independent group of 34 samples: 20 ALL and 14 AML. LEU72 data set was formed by combining the training set and testing set [1].

2. **COL62 data set**—Gene expression in 40 tumor and 22 normal colon tissue samples were analyzed with Affymetrix oligonucleotide arrays [28]. Two thousand out of around 6,500 genes were selected based on the confidence in the measured expression levels.

3. **BR-ER49 data set**—Breast tumors are either positive or negative for the estrogen receptors (ER). The collection of tumors consists of 25 ER+ tumors and 24 ER- tumors. All samples were tested using HuGeneFL Affymetrix microarray, with 6,817 genes included in each chip [29].

4. **LYM77**—Samples were obtained from 77 patients with 58 of them belonging to diffuse large B-cell lymphoma (DLBCL) group, and the remaining 19 belong to follicular lymphoma (FL). These biopsies were subjected to transcriptional profiling using oligonucleotide microarrays containing probes from 6,817 genes [30].

5. **PROS102**—Samples were divided into two groups: with 50 normal and 52 tumor prostate specimens included. A total of 12,600 are present on the Affymetrix chips used for these experiments [31]. The preprocessing steps are specified in the website [33].

6. **LUNG181**—Samples came from two different types of lung cancer. Thirty-one of them belong to malignant pleural mesothelioma (MPM) and the remaining 150 belong to adenocarcinoma (ADCA). These experiments were performed on Affymetrix human U95A oligonucleotide probe arrays. There are altogether 12,600 genes for each sample [32].

In the MFMW model, the three filters we used are SNR, TS, and PC (see Section 2). Each filter is used to generate a list of genes from the full set. Taking the union of the three lists consolidates overlapping genes and reduces the size of the combined list of filtered genes for the respective data sets to: 484 for LEU38, 467 for LEU72, 451 for COL62, 463 for BR-ER49, 473 for LYM77, 455 for PROS102, and 473 for LUNG181. Then, three classifiers WV, k-NN (k = 3), and SVM (see Section 2) are used to determine the gene subset. The training accuracy is defined as

\[
T - N_W - N_I \times 100\% ,
\]

where \( T \) is the total number of samples in the data set and \( N_W(N_I) \) is the number of samples with prediction status \( W(I) \).

Since it is not appropriate to use the training accuracy as defined in (3) for evaluating the classification performance of MFMW, to get a more realistic estimate of the classification accuracy, the MFMW model is evaluated using LOOCV (only R outputs contribute toward the accuracy). Table 3 shows the comparison of MFMW against the best reported existing SFSW models on the six data sets [21], [22], [23], [34], [35]. Although LEU is a benchmark data set, we are not aware of any studies applying the SFSW model to the LEU38 data set. For ease of comparison, LEU72 data set was used. Results given in Table 3 show that MFMW achieves the same accuracy, namely 100 percent, as in [23] for the LEU72 data set and 98.34 percent, as in [34] for the LUNG181 data set. MFMW performs better in all existing SFSW models for the remaining four data sets.

Details on the best set of selected genes \( P \) using MFMW for the six different data sets are shown in Table 4. Accession numbers of the Affymetrix chip, as well as the symbols and names of our selected genes, are given.

Next, we examine the results for LYM77 in more details. Table 5 shows the ranking of the MFMW-selected genes in the original lists of genes generated by SNR, TS, and PC. The first two genes selected by MFMW are either present in both the SNR and TS lists, or in the PC list only. This shows that using any one of these filters alone cannot generate this best subset of genes that yields 100 percent classification accuracy.

To further compare the MFMW with SFSW approaches, nine SFSW models are built by considering all possible combinations of the three filters and three classifiers used in our MFMW model. Of the six data sets we used, only LEU contains an independent testing data set. We first perform gene selection on LEU38 (using SFSW and MFMW, respectively). Out of the nine SFSW models, eight of them achieve perfect training classification accuracies, with different number of selected genes (see Table 6). When the chosen gene lists of the nine SFSW models are compared, few overlaps can be found (results not shown). This is undesirable as the genes selected depend on which filter and wrapper are used, and hence the selected genes can hardly be justified as
biomarkers. Prediction accuracies were obtained on the 34 testing samples (see Table 6), and it can be observed that the testing accuracies vary significantly from 67.65 percent to 91.18 percent even though the training accuracies are mostly 100 percent (except one at 94.73 percent). In contrast (see Table 7), a total of 21 pairs of genes were selected using MFMW, each giving a perfect training accuracy (defined as in (3)). These 21 MFMW models are more robust in their testing accuracies, ranging from 91.18 percent to 100 percent.

Similar experiments using all nine possible combinations of SFSW models were performed for the remaining five data sets (detailed results of individual SFSW models not shown), and LOOCV was used for evaluation. For all these data sets, the estimated classification accuracies of the proposed MFMW model (given in Table 3) consistently outperform those obtained from every one of the nine individual combinations. The results are summarized in Table 8.

5 DISCUSSION

5.1 Study on Varying the Number of Genes Selected by Each Filter

To study the effect of varying the number of genes selected by each filter, we repeat all the experiments of Section 4 for each data set with $m = 200$ and 350, to ascertain whether $m = 300$ is a reasonable choice. The LOOCV performance results for different values of $m$ are summarized in Table 9. Note that those in bold are the best LOOCV performance for each data set.

Estimated Classification Accuracies of the MFMW Model versus the Best of SFSW Models, for the COL62, BR-ER49, LYM77, PROS102, and LUNG181 Data Sets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Best SFSW Model</th>
<th>MFMW Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Filter</td>
<td>$t$ of genes</td>
</tr>
<tr>
<td>COL62</td>
<td>SNR</td>
<td>10</td>
</tr>
<tr>
<td>BR-ER49</td>
<td>PC</td>
<td>12</td>
</tr>
<tr>
<td>LYM77</td>
<td>PC</td>
<td>8</td>
</tr>
<tr>
<td>PROS102</td>
<td>TS</td>
<td>4</td>
</tr>
<tr>
<td>LUNG181</td>
<td>SNR</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No. of Pairs</th>
<th>Accession # (First gene)</th>
<th>Accession # (Second gene)</th>
<th>Training (35 samples)</th>
<th>Testing (34 samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>X951578_at</td>
<td>J06157_at</td>
<td>100%</td>
<td>94.12%</td>
</tr>
<tr>
<td>3</td>
<td>X951578_at</td>
<td>M16972_at</td>
<td>100%</td>
<td>95.47%</td>
</tr>
<tr>
<td>4</td>
<td>X951578_at</td>
<td>M3660_at</td>
<td>100%</td>
<td>97.06%</td>
</tr>
<tr>
<td></td>
<td>X951578_at</td>
<td>M3660_at</td>
<td>100%</td>
<td>97.06%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Accession</th>
<th>Rank in SNR list</th>
<th>Rank in TS list</th>
<th>Rank in PC list</th>
</tr>
</thead>
<tbody>
<tr>
<td>U264832_at</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>D178134_at</td>
<td>1</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>D55716_at</td>
<td>101</td>
<td>103</td>
<td>3</td>
</tr>
</tbody>
</table>

NA means that the corresponding gene is "Not Available" in the list.
For all the data sets, the best accuracy can be achieved with \( m = 300 \), except for COL62, where the accuracy has decreased from the case of \( m = 200 \) to that of \( m = 300 \) due to the increase of one misclassification. Theoretically, the use of a larger gene subset should always increase the accuracy. However, the MFMW algorithm is not guaranteed to converge to the global optimal solution, and the results of Table 9 suggest that the use of an unnecessarily large gene set may cause the algorithm to be trapped at a local minimum, as are the cases when \( m \) is increased from 300 to 350 for the COL62 and BR-ER49 data sets. Hence, for the data sets under consideration, \( m = 300 \) would be the best choice of number of genes to be retained by each filter.

### 5.2 Study on Varying the Number of Filters and Classifiers in the MFMW Model

The idea of using multiple filters is to include as many relevant genes of different characteristics as possible using different filters, so as not to leave out any potential biomarkers at the filtering stage. In this regard, it would be beneficial to include more filters. Apart from the three statistical filters used in our study, most of the other filter metrics are based on information theory, which do not quite match the classifiers in our study. Hence, by using SNR, TS, and PC, we have effectively included all filters relevant to our classification methods. The remaining question is whether a smaller number of filters can serve the same purpose as the three selected filters. Our study shows that for “good” data sets such as LEU38, two filters may generate sufficient genes for achieving the same accuracy as three filters. However, for more difficult data sets such as LYM77, some two-filter combinations fail to include some under-expressed genes.

### 5.3 Biological Interpretation of the Selected Genes

We would like to explore the biological significance of the MFMW-selected genes on all data sets. The contributions to each cancer types of these genes have been confirmed by other existing studies. With the help of the Oncomine database [36], where \( p \)-values obtained by performing student \( t \)-tests between two classes in a data set are given for each gene, we can determine how significantly over- or under-expressed the MFMW-selected genes are.

1. **LEU data set**—First, genes chosen from LEU38 data set are discussed (see Table 7). The first selected gene in our gene subsets is X95715_at (with symbol ZYX), which is also selected in [1] as under-expressed in the ALL class (with \( p \)-value: \( 4.1E^{-10} \)). The gene may be a component of a signal transduction pathway that mediates adhesion-stimulated changes in gene expression. Note that MFMW is capable of consistently identifying ZYX as the first gene in all 21 gene sets. Furthermore, of the 21 genes given in Table 7 which, when added to ZYX give a perfect training set, \( 3 \) out of the 21 (= \( 7C2 \)) cases of two-classifier MFMW model cannot achieve the same accuracy as the three-classifier MFMW model, whereas the use of four classifiers do not improve the accuracy any further. Hence, the use of three classifiers is optimal for this data set.

Clearly, it is not possible to conclude that the use of three filters and three classifiers is optimal in general—even testing this hypothesis just for all the data sets considered in this paper is computationally prohibitive. However, our experiences based on a wide range of experiments (many of which are not reported here) suggest that the choice of three filters and three classifiers in an MFMW model incorporating the kind of popular filtering and classification methods considered in this paper should be close to an optimal compromise between accuracy and computational efficiency.

### Functions of the Genes

<table>
<thead>
<tr>
<th>Accession ID</th>
<th>Over/Under-expressed (class)</th>
<th>Dataset: ( p )-value</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>M92357_at (CCND3)</td>
<td>Over-expressed (ALL)</td>
<td>Goelub [1]: 1.3E ( ^{-10} )</td>
<td>Essential for the control of the cell cycle at the G1/S transition.</td>
</tr>
<tr>
<td>Z15115_at (TOP2A)</td>
<td>Over-expressed (ALL)</td>
<td>Goelub [1]: 1.2E ( ^{-11} )</td>
<td>Control of topological states of DNA; widely used in the treatment of AML and ALL</td>
</tr>
<tr>
<td>U23736_s_at (MYB)</td>
<td>Over-expressed (ALL)</td>
<td>Armstrong [38]: 7.5E ( ^{-6} )</td>
<td>Involved in hematopoiesis of both AML and ALL</td>
</tr>
<tr>
<td>M63838_s_at (H116)</td>
<td>Over-expressed (ALL)</td>
<td>Armstrong [38]: 1.2E ( ^{-6} )</td>
<td>Controls cellular proliferation by modulating the functions of cell cycle regulatory factors, e.g. p53</td>
</tr>
<tr>
<td>M19507_at (MPO)</td>
<td>Under-expressed (ALL)</td>
<td>Goelub [1]: 7.6E ( ^{-4} )</td>
<td>Involved in phagocytosis</td>
</tr>
<tr>
<td>M36222_s_at (KIT)</td>
<td>Under-expressed (ALL)</td>
<td>Armstrong [38]: 2.7E ( ^{-6} )</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accession ID</th>
<th>Over/Under-expressed (class)</th>
<th>Dataset: ( p )-value</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>M92357_at (CCND3)</td>
<td>Over-expressed (ALL)</td>
<td>Goelub [1]: 1.3E ( ^{-10} )</td>
<td>Essential for the control of the cell cycle at the G1/S transition.</td>
</tr>
<tr>
<td>Z15115_at (TOP2A)</td>
<td>Over-expressed (ALL)</td>
<td>Goelub [1]: 1.2E ( ^{-11} )</td>
<td>Control of topological states of DNA; widely used in the treatment of AML and ALL</td>
</tr>
<tr>
<td>U23736_s_at (MYB)</td>
<td>Over-expressed (ALL)</td>
<td>Armstrong [38]: 7.5E ( ^{-6} )</td>
<td>Involved in hematopoiesis of both AML and ALL</td>
</tr>
<tr>
<td>M63838_s_at (H116)</td>
<td>Over-expressed (ALL)</td>
<td>Armstrong [38]: 1.2E ( ^{-6} )</td>
<td>Controls cellular proliferation by modulating the functions of cell cycle regulatory factors, e.g. p53</td>
</tr>
<tr>
<td>M19507_at (MPO)</td>
<td>Under-expressed (ALL)</td>
<td>Goelub [1]: 7.6E ( ^{-4} )</td>
<td>Involved in phagocytosis</td>
</tr>
<tr>
<td>M36222_s_at (KIT)</td>
<td>Under-expressed (ALL)</td>
<td>Armstrong [38]: 2.7E ( ^{-6} )</td>
<td></td>
</tr>
</tbody>
</table>
that ZYX is not among one of the four selected genes. This may be due to the inconsistencies present in the training and testing data sets. Of the four genes, two of them (namely MME and CD33) are biologically useful. MME is an important cell surface marker in the diagnostic of human ALL [37]. It is over-expressed in ALL of three data sets: p-value: 7.5E−20 [38], p-value: 4.9E−15 [39], and p-value: 5.1E−5 [1]. CD33 is under-expressed in ALL with a p-value of 1.2E−20, 9.9E−15, and 2.4E−5 in data sets of Andersson et al. [39], Golub et al. [1], and Armstrong et al. [38], respectively. It induces apoptosis in AML (in vitro) [40].

2. **COL62 data set**—Since no genes were identified to distinguish between cancerous and normal samples in [28], no comparison can be made. Our selected genes include CD36, GSN, and MORF4L2. CD36 was reported to increase the risk of predisposing colon cancer [41] by functioning as a cell adhesion molecule. GSN is likely an effector of morphologic change during apoptosis, and is found to be over/under-expressed in cancerous samples by Alon et al. [28] (p-value: 1.1E−4) and Noterman et al. [42] (p-value: 5.2E−8). MORF4L2 is responsible for the activation of transcriptional programs associated with oncogene mediated growth induction.

3. **BR-ER49 data set**—None of our six MFMW-selected genes are in the list of genes claimed to be differentially expressed between ER+ and ER- samples in the original paper [29]. However, three potential biomarkers are found: PIM1, PG5M, and E2F2. PIM1 contributes to both cell proliferation and potential biomarkers are found: PIM1, PGM5, and samples in the original paper [29]. However, three genes are in the list of genes claimed to be mediated growth induction.

4. **LYM77 data set**—All three MFMW-selected genes, namely KHSRP, CIRBP, and MCM7, are good biomarkers for lymphoma. The first selected gene is KHSRP, a type of KH-type splicing regulatory protein, which plays a role in mRNA trafficking. This gene is also found to be over-expressed (p-value: 6.8E−8) in DLBCL group of Alizadeh’s lymphoma cDNA data set (IMAGE: 1340925) [46]. The second selected gene, CIRBP, is significantly over-expressed (p-value: 9.4E−5) for the FL group [30]. This cold inducible RNA binding protein plays an essential role in cold-induced suppression of cell proliferation. The last one, MCM7, was selected by Shipp et al. [30]. It was under-expressed (p-value: 8.1E−11) for the FL group. It is required for DNA replication and cell proliferation.

5. **PROS102 data set**—Genes selected by our MFMW model have little correlation with existing knowledge on the differentiation between prostate cancer and normal tissues. Only gene PAIP2B is significantly over-expressed in the cancerous tissue in [47] (p-value: 7.6E−7).

6. **LUNG181 data set**—Only two of our selected genes are good at differentially between the two types of lung cancers (MPM and ADCA). The first gene, SLC7A5, is significantly over-expressed in ADCA of three data sets: p-value: 8.6E−8 [48], p-value: 3.3E−7 [49], and p-value: 3.3E−5 [32]. The second gene is OGT, which is also significantly over-expressed in ADCA in [50] (p-value: 2.3E−6).

### 6 Conclusion

Although combining filters and wrappers for gene selection appears to be a sensible approach, SFSW models yield results that are dependent on the specific choice of filter and wrapper. In this paper, an MFMW approach has been proposed to overcome the difficulties of SFSW models, to improve the classification accuracy, and possibly to identify biomarker genes. By combining genes selected by different filters, an expanded feature subset is created to ensure that relevant genes are not left out. Improvement in classification accuracy is achieved through the use of multiple filters with a unanimous voting scheme.

The proposed MFMW model was tested on six microarray data sets containing samples of cancerous versus normal, or subtypes of diseases. The experimental results show that the LOOCV estimated classification accuracies obtained by the MFMW model outperforms that of SFSW models generated from all possible combinations of filters and wrappers used in the MFMW model. Many of the MFMW-selected genes have been confirmed to be biomarkers or contribute to the development of particular cancers by other studies.

We have not compared the results of the MFMW with classification results using popular methods based on PCA because of the different nature of the two classes of methods. PCA methods perform gene selection by feature extraction, whereas MFMW is designed to optimize the classification accuracy. Hence, MFMW is expected to outperform the feature extraction methods, as can be readily shown by comparing the results presented in this paper with existing ones based on feature extraction methods.

### Acknowledgments

The authors would like to thank the reviewers for their valuable comments which helped improve the manuscript in many ways. This work was supported by a HKU CRCG research grant.

### References


