

A Hybrid Gene Selection Method for Cancer Classification

¹Li-Juan Zhang, ²Zhou-Jun Li and ³XiaoHua Hu

¹National Laboratory for Parallel and Distributed Processing, Changsha, China

²School of Computer Science & Engineering, Beihang University, Beijing, China

³College of Information Science and Technology, Drexel University, Philadelphia, USA

¹nudtzlj@126.com, ²lizj@buaa.edu.cn, ³thu@cis.drexel.edu

Abstract

Due to a large number of genes (features) and a small sample size, gene expression microarray data poses a severe challenge for accurate classification of diseases or phenotypes. Gene selection is a frequently used technique in preprocessing microarray data for successful classification of diseases or phenotypes. Widely used gene selection methods are mainly focused on filter approaches. They have been proved to be efficient and effective. However, we find that some genes discarded by many existing methods are helpful for classification at certain conditions and can not be removed blindly. In this paper we propose a novel hybrid approach, which combines discernibility matrix, the filter strategy and the wrapper method based on Ockham's Razor, thus can select genes by the simplest possible ways and improve classification accuracy to a high enough level. The extensive empirical study using public microarray data shows the superior performance of our proposed approach.

1. Introduction

The rapid advances in microarray technology enable biologists to measure the expression levels of thousands or ten thousands of genes simultaneously [12]. The initial information from microarray experiments goes through various data processing steps including image processing, quality control and normalization [3]. The resulting data set is called gene expression microarray data set, which is a two dimensional array with thousands of columns (genes) but a small number (often less than one hundred) of rows (samples). Such data set poses a very severe challenge for sample classification and usually results in the known problem of “curse of dimensionality” and over-fitting of the training data for traditional sample

classification technologies [5, 10]. Therefore, selecting a small number of discriminative genes from thousands of genes is essential for successful sample classification [5, 6]. In recent years, feature selection has been extensively applied to gene selection for sample classification.

Feature selection is a process that selects a subset of original features by reducing the number of features, removing irrelevant, redundant, or noisy data. It has been proved to be effective in improving classification accuracy, speeding up a classification algorithm and enhancing result comprehensibility [9]. With regard to how to evaluate the goodness (quality) of a subset of features, the feature selection methods fall into two broad categories: the filter approach and the wrapper approach [9]. In the filter approach, a good feature set is selected as a result of pre-processing based on properties of the data itself and independent of the classification algorithm. The wrapper approach requires one predetermined mining algorithm in feature selection and uses its performance to evaluate and determine which features are selected. It tends to give superior performance as it finds features better suited to the predetermined classification algorithm, but it is more computationally expensive than the filter approach [9]. For this reason the filter model is widely used in gene selection for microarray data.

However, the high dimensionality and small sample size of microarray data also poses severe challenges to filter approaches in terms of effectiveness. Some of the recent research efforts have been focused on these challenges [8, 17, 19, 20, 21, 22]. Whereas, we find that selecting some genes from the ones that are discarded by these filter methods can lead to even higher classification accuracy. In this work, we aim to develop a novel hybrid solution for gene selection in sample classification of microarray data which can select discriminative genes and improve classification accuracy more effectively.

The paper is organized as follows. In section 2 we review related work and point out their problems. In section 3 we first introduce some basic notions and techniques used in our approach, then propose a new algorithm. In section 4, we evaluate the performance of our method via extensive experiments. Section 5 concludes our work.

2. Related work and problems

The problem of feature selection is in nature the problem of finding a good subset of the original features. But what is a good subset? Because the foremost goal of feature selection for classification is improving classification accuracy, a good subset is naturally a subset on which the accuracy of a given classifier is maximal. The feature selection algorithm based on this principle must be a wrapper algorithm, which is usually computationally expensive and thus infeasible for high dimensional data.

To overcome this problem, Hall [7] proposed a practical principle: Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other. This principle makes use of the properties of the data itself to evaluate the goodness of features and is independent of the classification algorithm. It is simple and efficient for high dimensional data. Therefore, most of the existing gene selection methods are based on this principle. According to this principle, these existing methods usually define feature relevance and redundancy by different means, and then use the definitions to select features. With regard to how to handle feature relevance and redundancy, these methods can be divided into four categories. Below, we discuss the advantages and shortcomings of methods in each category and show the need of a new method.

Earlier gene selection methods are gene ranking. They usually evaluate each gene individually by assigning a discriminative score to each gene in the original set and then select the top ranked genes according to their individual discriminative power [6, 11, 20]. These methods are efficient for high-dimensional data due to linear time complexity in terms of dimensionality. However, without considering the interactions among genes, they can only identify relevant features to the target concept according to different relevance criteria, but cannot effectively discover redundancy among genes and require the user to determine the threshold for the number of selected genes.

The second category methods are based on clustering. They often firstly apply clustering

algorithm to group similar genes into clusters, genes in the same cluster are considered to be highly correlated, and then select the most relevant genes from each cluster [1, 21] or create a gene prototype for each cluster to represent this cluster [8]. The third category methods integrate the metric for measuring the gene-class relevance and that for measuring the gene-gene redundancy into a single criterion function and then select genes so that the criterion function is optimized [4, 7, 16]. These two categories of methods can remove redundant genes to certain extent but are all time-consuming. The last category uses a new framework that decouples relevance analysis and redundancy analysis. They usually measure the gene-class relevance to obtain relevant genes, and then use a methodology called redundant cover to remove redundancy. Methods in this category have been proved to be effective and efficient on many high dimensional data sets [18, 19, 22].

The above filter approaches use feature relevance and redundancy to select features. There are a number of different definitions on feature relevance and redundancy in machine learning literature [2, 9]. But these definitions are usually independent of the specific classification algorithm being used, and none of them has been accurately and rigorously defined in common agreement. So the methods based on these definitions usually remove some useful features by error. To overcome this problem, we develop a novel hybrid approach which can effectively select features and improve classification accuracy.

3. A novel hybrid approach

In this section, we introduce our new approach. We attempt to take advantage of all of the different approaches by exploiting their best performances.

3.1 Basic idea and definition

According to the discussion in section 2, the principle used by the wrapper model is too strict to be practical for high dimensional data. While the principle adopted by the filter model is too loose. We here give a moderate principle.

Definition 1 (good subset)

Given a sample of data D , a classification algorithm L , a feature subset S is a good subset if the accuracy of the hypothesis that L produces using S is high enough, and the procedure obtaining S is simple enough to implement in a limited time.

A good subset need not be unique because it may be possible to achieve the same accuracy using

different subsets of features. A good feature selection algorithm should endeavor to find one good subset using definition 1. But this definition is qualitative. How to find a good subset that satisfies the definition 1? A good subset contains one or many good features. A good feature must be selected based on a simple evaluation measure (for example information gain) or a classification method as evaluation measure.

Based on the degree of dependency selecting good features on evaluation measures and classification algorithms, we divide a good subset into three parts: the first part contains the features that are absolutely necessary for classification. Selection of them is independent of any evaluation measure and classification algorithm, we call them independent features, they are in nature the core of all good subsets. The second part includes the features that can be chosen based on properties of the data itself and independent of the classification algorithm, we call them half-independent features. The third one involves the features that only can be found using a classification algorithm as evaluation measure, we name them dependent features.

The selection of independent features is the simplest to some extent, because it is independent of any evaluation measure and classification algorithm. The selection of dependent features is the most complicated, because it requires a classification algorithm as the evaluation measure.

According to Ockham's Razor, we should adopt the simplest possible method to select good features. Only when the simple method fails, the complicated approach is used. So our approach should select firstly the independent features, then the half-independent features and lastly select the dependent features.

3.2 Primary techniques

To find a *good subset*, it is enough for us to find all the independent, half-independent and dependent features. For independent features, selection of them is required to be independent of any evaluation measure and classification algorithm. We here use the notion of discernibility matrix [13] from rough set theory to find them. The definition of discernibility matrix is described as follows:

Given a data set S , feature set $F=\{f_1, f_2, \dots, f_n\}$, class label C , sample set $U=\{x_1, x_2, \dots, x_m\}$, let $f_i(x_j)$ denote the value of the sample x_j for the feature f_i , then the discernibility matrix of data set S is a symmetric $|U| \times |U|$ matrix with entries c_{ij} defined

as $\{f_k \in F \mid f_k(x_i) \neq f_k(x_j)\}$ if $C(x_i) \neq C(x_j)$, \emptyset otherwise. The following example can help understand the definition of discernibility matrix.

Example 1: Suppose we have a collection of 4 persons with information about the Stature, Hair, Eye and Race. Stature, Hair and Eye are the features and Race is the class label. The example data set is presented in Table 1.

Table 1 **Example Data Set**

	Stature	Hair	Eye	Race
1	low	black	blue	no
2	low	yellow	blue	yes
3	high	yellow	brown	no
4	high	red	blue	yes

Based on the definition of discernibility matrix, we construct the discernibility matrix of the example data as shown in Table 2.

Table 2 **Discernibility Matrix of Example Data**

	1	2	3	4
1	\emptyset	Hair	\emptyset	Stature, Hair
2	Hair	\emptyset	Stature, Eye	\emptyset
3	\emptyset	Stature, Eye	\emptyset	Hair, Eye
4	Stature, Hair	\emptyset	Hair, Eye	\emptyset

According to the definition of discernibility matrix, if there is only one element in an entry c_{ij} , it must be an independent feature because it is the only feature that can differentiate sample x_i and x_j . Based on this observation, we can examine each entry in the discernibility matrix to obtain all the independent features. For example, in Table 2, only the entry c_{12} contains the only one feature "Hair", which means the values of sample 1 and 2 on the feature "Hair" are different. Namely, only the feature "Hair" can differentiate sample 1 and 2 and thus the feature "Hair" is an independent feature. It is possible that no entry contains only one feature, so the independent feature subset may be an empty set. In addition, construction of discernibility matrix requires the value of features discrete. Therefore, we must discretize the features. We adopt Fayyad & Irani's MDL method[14].

As for the half-independent features, selection of them is independent of classification algorithm. But it still relies on some evaluation measures to evaluate the goodness of features. These evaluation measures usually try to evaluate the goodness of a feature or feature subset by exploiting the intrinsic characteristics of the training data without involving any classification algorithm. Some popular evaluation measures are distance measure, correlation measure, consistency measure and dependency measure. Among them, the correlation measure has been widely used and shown effective [18]. We here adopt the correlation measure to evaluate the goodness of features.

We denote the correlation value of a feature f_i and the class C as $Corr(f_i, C)$, and let $F_S \subseteq F$ be a feature subset, then we can use $Corr(f_i, C)$ to define the correlation between F_S and class C as follows:

$$Corr(F_S, C) = \frac{1}{|F_S|} \left(\sum_{f_i \in F_S} Corr(f_i, C) \right)$$

Given F_S is the current selected feature subset, E is the selected correlation measure. A feature $f_i (f_i \notin F_S)$ should be selected iff $Corr_E(\{f_i\} \cup F_S, C) > Corr_E(F_S, C)$.

As for the dependent features, it is difficult or even infeasible for us to select them only based on some simple evaluation measures. We need a classification algorithm L to evaluate their goodness.

Given F_S is the current selected feature subset, L is the selected classification algorithm. We denote the accuracy of L using F_S as $Acc_L(F_S, C)$. A feature $f_i (f_i \notin F_S)$ should be selected iff $Acc_L(\{f_i\} \cup F_S, C) > Acc_L(F_S, C)$.

In addition, we must also decide the search starting point and search strategy for the process of feature selection. Because the independent features are absolutely necessary, we start search with it to prevent losing some independent features in the successive process of search. If the independent feature subset is an empty set, search will start with an empty set. It is important for high dimensional microarray data to search feature space quickly. We therefore rank features according to the selected correlation measure E and adopt sequential forward search strategy, which is simple to implement and fast.

3.3 Algorithm and analysis

Based on the above analysis, we can easily obtain a deterministic procedure that can effectively identify discriminative genes in microarray data set. Our approach can be summarized by an algorithm HFW (Hybrid of Filter and Wrapper) shown in Figure 1.

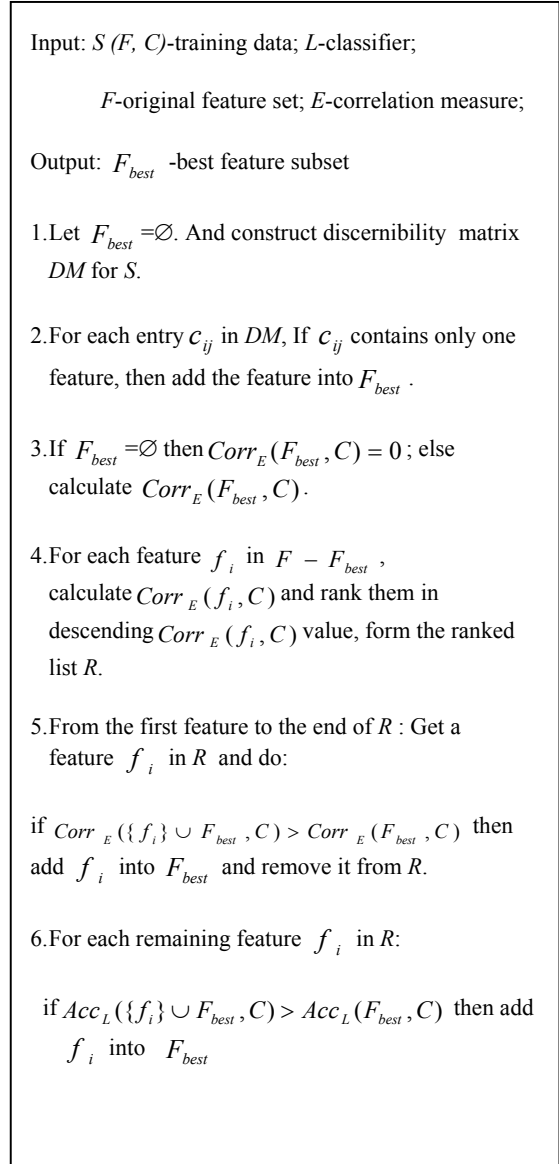


Figure 1. **A Novel Hybrid Algorithm**

As in Figure 1, given a microarray data set S with n genes, m samples and a class C , the algorithm consists of four parts. In the first part (tag 1-2), it constructs a discernibility matrix from the input data set, finds the independent gene subset based on the discernibility

matrix as the starting set. In the second part (tag 3-4), it calculates the correlation between the starting set and the class C and ranks the rest genes using the predefined correlation measure E . In the third part (tag 5), it iteratively chooses half-independent genes from the ranked list R using the predefined correlation measure E . The iteration starts from the first gene in the ranked list R and continues until the last gene in the list R is reached. In the fourth part (tag 6), it selects more accurately dependent genes from the genes discarded by the previous steps using the predefined classifier L .

The first part of the above algorithm has the complexity $O(m^2n)$, because discernibility matrix DM is symmetrical, constructing it needs $\frac{(m+1)mn}{2}$

times. The second part and the third part has respectively linear time complexity $O(n)$ in terms of dimensionality n . In the fourth part, the time complexity depends on the classification algorithm chosen.

4. Empirical study

In this section, we empirically evaluate our proposed algorithm in terms of running time, degree of dimensionality and classification accuracy on selected genes. For gene selection in sample classification, it is perfect to select small enough genes which can lead to high enough classification accuracy.

4.1 Experimental setup

We perform extensive experiments using twelve microarray data sets¹. The main characteristic of these data sets is the great number of genes and the relatively small number of samples. The details of these data sets are summarized in Table 3.

Three representative feature selection algorithms are chosen in comparison with HFW. One algorithm representing feature ranking methods is ReliefF [11], which searches for nearest neighbors of instances of different classes and ranks features according to their importance in differentiating instances of different classes. Another algorithm is a variation of CFS [7], denoted by CFS-SF (Sequential Forward), which used some correlation measure and sequential forward search to obtain optimal subset. A third one is FCBF [18], which used the correlation measure *symmetrical uncertainty* to obtain relevant genes and to remove redundancy. To test how the selection of dependent

genes affects the performance of HFW, we also use half-HFW to represent the version of HFW without selecting dependent genes (namely, removing the fourth part of the algorithm HFW). In addition, two widely used classification algorithms, C4.5 and NaiveBayes, are adopted to evaluate the predictive accuracy of the selected genes. The experiments are conducted using WEKA's implementation of all these existing algorithms and our algorithm is also implemented in the WEKA environment [15].

Table 3 Summary of Microarray Data

Title	#genes	#samples	#class
colonTumor	2000	62	2
breastCancer	24481	97	2
lungCancer harvard	12533	181	2
lungCancerOntario	2880	39	2
ALL-AML_Leukemia	7129	72	2
MLL_Leukemia	12582	72	3
DLBCL-Stanford	4026	47	2
DLBCLTumor	7129	77	2
DLBCLOutcome	7129	58	2
prostate	12600	136	2
centralNervousSystem	7129	60	2
Hyperdip50	12558	64	2

For each data set, we first run all the feature selection algorithms, and obtain the running time and selected genes for each algorithm. Note that in HFW, the correlation measure E is set to *symmetrical uncertainty*. The classifier is respectively set to C4.5 and NaiveBayes. We differentiate the two versions by $HFW_{C4.5}$ and HFW_{bayes} . For ReliefF, we use 5 neighbors and 30 instances throughout the experiments, and to compare the performance of ReliefF and HFW, the number of genes selected by ReliefF is set to be the same as that of HFW. Because HFW has two versions $HFW_{C4.5}$ and HFW_{bayes} , we respectively use $ReliefF_{C4.5}$ and $ReliefF_{bayes}$ to differentiate corresponding ReliefF. We then apply C4.5 and NaiveBayes respectively on each original data set and each newly obtained data set only containing the selected genes, and obtain the overall classification accuracy by leave-one-out cross-validation. The experiments are conducted on an AMD 64 PC with 512 MB RAM.

4.2 Results and discussions

Table 4 reports the running time for each feature selection algorithm. As we can see, time needed for $HFW_{C4.5}$ and HFW_{bayes} are different on each data set, because they wrap different classification algorithms in the fourth part of the algorithm respectively. Time needed for half-HFW, ReliefF and FCBF are

¹ <http://sdmc.lit.org.sg/GEDatasets/Datasets.html>

approximately the same on each data set, because they do not depend on the classifier used. In addition, we can see that CFS-SF is not available on most of data sets due to its $O(n^2)$ complexity in terms of the number of genes n .

Table 5 records the number of genes selected by each feature selection algorithm. We can see that half-HFW on average selects the smallest number of genes.

Table 6-7 reports the leave-one-out accuracy by C4.5 and NaiveBayes respectively. For most of the data sets, we can observe that, (1) CFS-SF is not available due to its $O(n^2)$ complexity in terms of the number of genes n . (2) Both half-HFW and HFW can increase or maintain the accuracy of C4.5 and NaiveBayes. But HFW is better than half-HFW in improving classification accuracy, because it selects useful genes from the ones discarded by half-HFW and thus improves classification accuracy to a higher level. (3) None of other three algorithms can enhance the accuracy of C4.5 and NaiveBayes to the same level as HFW does. In summary, the above experimental results suggest that HFW is effective in gene selection and is practical for use in sample classification of high dimensional microarray data.

5. Conclusions

In this work, we have provided a moderate principle for feature selection and proposed a novel hybrid solution which has combined discernibility matrix, the filter strategy and the wrapper method based on Ockham's Razor. Extensive experiments on microarray data have demonstrated the superior performance of our approach.

6. Acknowledgments

This work is supported by the National Science Foundation of China under Grants No. 60573057, 60473057 and 90604007.

References

[1] Wai-Ho Au, Keith C.C.Chan, Andrew K.C.Wong, and Yang Wang, Attribute Clustering for Grouping, Selection, and classification of gene expression data, IEEE/ACM transactions on computational biology and bioinformatics VOL 2, NO.2, APRIL-JUNE 2005
[2] David A. Bell and Hui Wang, A Formalism for Relevance and Its Application in Feature Subset Selection, Machine Learning, 41,175-195, 2000.
[3] G. V. Bobashev, S. Das, and A. Das. Experimental design for gene microarray experiments and differential expression analysis. Methods of Microarray Data Analysis II, 2001.

[4] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. In Proceedings of the Computational Systems Bioinformatics Conference, pages 523-529, 2003.
[5] E. R. Dougherty. Small sample issue for microarray-based classification. Comparative and Functional Genomics, 2:28-34, 2001.
[6] T. R. Golub et al. Molecular classifications of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286(5439):531-537, 1999.
[7] M. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the 17th International Conference on Machine Learning, pages 359-366, 2000.
[8] Blaise Hanczar et al. Improving Classification of Microarray Data using Prototype-based Feature Selection, SIGKDD Explorations Volume 5, Issue 2, pages23-30, 2003
[9] R. Kohavi, G. John, Wrappers for feature subset selection, Artif. Intell. 1-2 (1997) 273-324.
[10] S. Ramaswamy, P. Tamayo, et al. Multiclass cancer diagnosis using tumor gene expression signatures. PNAS, 98(26):15149-15154, 2001.
[11] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of Relief and ReliefF. MachineLearning, 53:23-69, 2003.
[12] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science,270: 467-470, 1995.
[13] Skowron A, Rauszer C. The discernibility matrices and functions in information systems. In: Slowinski R ed. Intelligent Decision Support—Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, 1992: 331-362.
[14] Usama M. Fayyad and Keki B. Irani, On the Handling of Continuous-Valued Attributes in Decision Tree Generation, Machine Learning, vol.8, page 87-102, 1992.
[15] I. Witten and E. Frank. Data Mining – Practical Machine Learning Tools and Techniques with JAVA Implementations. Morgan Kaufmann Publishers, 2000.
[16] Y. Wu and A. Zhang. Feature selection for classifying high-dimensional numerical data. In IEEE Conference on Computer Vision and Pattern Recognition 2004, volume 2, pages 251-258, 2004.
[17] E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. In Proceedings of the 18th International Conference on Machine Learning, pages 601-608, 2001.
[18] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learning Res. 5 (2004) 1205-1224.
[19] L. Yu and H. Liu. Redundancy based feature selection for microarray data. In Proceedings of the Tenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 737-742,2004.
[20] Li-Juan Zhang and Zhou-Jun Li, Gene Selection for classifying microarray data using grey relational analysis. Proceedings of Discovery Science'2006, Lecture Notes in Computer Science Vol.4265: 378-382, 2006.

[21] Li-Juan Zhang, Zhou-Jun Li, Huo-Wang Chen and Jian Wen, Minimum Redundancy Gene Selection based on Grey Relational analysis. In Workshops Proceedings of ICDM'2006 pages 120-124, IEEE Computer Society, 2006.

[22] Li-Juan Zhang, Zhou-Jun Li and Huo-Wang Chen, An Effective Gene Selection Method Based on Relevance Analysis and Discernibility Matrix. Accepted by PAKDD 07.

Table 4 Running time(s) for each feature selection algorithm

	HFW _{C45}	HFW _{bayes}	half-HFW	FCBF	CFS-SF	ReliefF
colonTumor	59.83	22.41	0.91	1.14	63.78	1.39
breastCancer	4163.98	1858.34	12.97	14.66	N/A	20.3
lungCancer harvard	564.5	453.67	17.23	15.09	N/A	17.92
lungCancerOntario	52.36	39.3	0.89	1.08	36.12	1.24
ALL-AML_Leukemia	146.58	121.47	3.59	3.3	N/A	4.97
MLL_Leukemia	436.77	298.27	9.34	7.47	N/A	9.41
DLBCL-Stanford	56.19	39.59	1.3	1.66	N/A	1.86
DLBCLTumor	229.86	108.16	3.41	3.34	N/A	5.03
DLBCLOutcome	395.27	140	2.34	2.5	N/A	3.75
prostate	713.83	495.47	10.11	7.8	N/A	13.97
centralNervousSystem	325.2	135.14	3.25	2.8	N/A	4.19
Hyperdip50	2985.64	1388.56	24.13	24.66	N/A	28.7

Table 5 Number of genes selected by each feature selection algorithm

	Full set	HFW _{C45} ReliefF _{C45}	HFW _{bayes} ReliefF _{bayes}	half-HFW	FCBF	CFS-SF
colonTumor	2000	11	14	5	14	26
breastCancer	24481	36	66	7	90	N/A
lungCancer harvard	12533	7	12	7	128	N/A
lungCancerOntario	2880	12	18	5	12	13
ALL-AML_Leukemia	7129	9	15	6	51	N/A
MLL_Leukemia	12582	11	13	5	97	N/A
DLBCL-Stanford	4026	8	9	6	60	N/A
DLBCLTumor	7129	12	12	5	73	N/A
DLBCLOutcome	7129	22	19	5	27	N/A
prostate	12600	9	29	6	38	N/A
centralNervousSystem	7129	20	29	4	28	N/A
(Hyperdip50)	12558	25	36	5	244	N/A

Table 6 Leave-one-out cross-validation accuracy of C4.5 on selected genes for each feature selection method (%)

	Full set	HFW _{C45}	half-HFW	FCBF	CFS-SF	ReliefF _{C45}
colonTumor	80.65	90.32	85.48	88.71	87.10	82.26
breastCancer	57.73	78.35	78.35	67.01	N/A	67.01
lungCancer harvard	96.13	97.24	97.24	98.90	N/A	98.90
lungCancerOntario	84.62	89.74	76.92	79.49	66.67	61.54
ALL-AML_Leukemia	73.61	87.5	87.5	81.94	N/A	86.11
MLL_Leukemia	86.11	93.06	90.28	93.06	N/A	87.5
DLBCL-Stanford	82.98	89.36	89.36	82.98	N/A	82.98
DLBCLTumor	80.52	94.81	90.91	85.71	N/A	90.91
DLBCLOutcome	37.93	75.86	65.52	46.55	N/A	43.10
prostate	77.94	95.59	93.38	83.82	N/A	86.03
centralNervousSystem	50	75	75	66.67	N/A	50
Hyperdip50	88.07	95.41	93.88	89.30	N/A	89.60

Table 7 **Leave-one-out cross-validation accuracy of bayes on selected genes for each feature selection method (%)**

	Full set	HFW _{bayes}	half-HFW	FCBF	CFS-SF	ReliefF _{bayes}
colonTumor	75.81	90.32	90.32	77.42	80.65	83.87
breastCancer	71.13	87.63	83.72	55.67	N/A	65.98
lungCancer harvard	98.34	100	100	99.45	N/A	98.34
lungCancerOntario	69.23	89.74	79.49	61.54	61.54	64.10
ALL-AML Leukemia	97.22	98.61	95.83	95.83	N/A	93.06
MLL Leukemia	91.67	100	95.83	94.44	N/A	94.44
DLBCL-Stanford	97.87	97.87	93.62	97.87	N/A	93.62
DLBCLTumor	85.71	97.41	92.21	92.21	N/A	89.61
DLBCLOutcome	51.72	86.21	85.38	53.45	N/A	53.45
prostate	63.97	64.71	63.97	65.44	N/A	61.76
centralNervousSystem	65	86.67	80	56.67	N/A	56.67
Hyperdip50	96.64	96.94	93.58	94.50	N/A	93.27