Integrating Feature Selection Methods for Gene Selection

Patharawut Saengsiri[1], Sageemas Na Wichian[2], Phayung Meesad[3], Herwig Unger[4]

Department of Information Technology, Faculty of Information Technology[1]
Department of Applied Science and Social, College of Industrial Technology[2]
Department of Teacher Training in Electrical Engineering,
Faculty of Technical Education[3]
Department of Communication Network,
Faculty of Mathematics and Computer Science[4]

KMUTNB, 1518 Pibulsongkram Road, Bangsue, Bangkok, Thailand[123]
10800
Fern Universitatsstr. 27/PRG Hagen, Germany[4]
58084
s5070290035@kmutnb.ac.th[1]
sgm@kmutnb.ac.th[2]
pym@kmutnb.ac.th[3]
Herwig.Unger@gmx.de[4]

Abstract: In fact, cancer is produced for genetic reasons. So, gene feature selection techniques are very important for biological processes which help to find subsets of informative genes. However, the quality of recognition is still not sufficient and leads to low accuracy rates. Hence, this research proposes integrating a feature selection method (IFS). There two phases of IFS: 1) determining feature length by Gain Ratio (GR) and 2) estimating each rank list using a wrapper approach based on K-nearest neighbor classification (KNN), Support Vector Machine (SVM), and Random Forest (RF). Experimental results based on two gene expression datasets, it is found that the proposed method not only has higher accuracy rate than tradition methods, but also reduce many irrelevant features. In addition, most models based on IFS method are more beneficial when working with two or multi-classes.

1 Introduction

Gene expression data is an expression level of gene when DNA (Deoxyribonucleic acid) encoded into protein. At present, expression level of gene is measured by using microarray techniques. This technique is applied to hybridization of nucleic acid for validating gene expression data [HO08]. Actually, several genes are not related to others, this in turn leads to high time consumption in the decision process.

A basic technique of gene feature selection at the moment is to search for genes that have recognition power via filter and wrapper approaches which are the main techniques for feature selection. The first method is to connect individual discriminate power of genes without involving an induction algorithm such as Chi-square (χ^2), Information gain (Info), Gain Ratio (GR) and correlation based feature selection (Cfs). The next method is to associate with induction algorithms (such as Forward Stepwise Feature Selection (FSFS) or Backward Stepwise Feature Selection (BSFS)) [PA07] which determine the power of a pick out subset of genes. In the case of time consumption, the wrapper approach takes more time than the filter approach. On the other hand, the filter method has an accuracy rate less than the wrapper method. Nevertheless, finding a subset of informative genes is beneficial for biology processes that need high accuracy. Hence, the wrapper approach is suitable for supporting these processes. Technique transformation based such as Principle Component Analysis (PCA), Single Value Decomposition (SVD) and Independent Component Analysis (ICA) are not desirable for gene feature selection because they translate less then the original features, cannot move any attribute, and unrelated features [PL04].

As mentioned above, the curse of dimension problems based on gene expression data not only leads to very time consuming but also high miss classification rates. In the traditional feature selection model, filter approach is used first and after that it is sent to the wrapper approach. However, a key problem of this method is that it produces low accuracy rates. Choosing an appropriate algorithm for the wrapper approach with gene expression dataset depends on results from the experiment. Hence, this research proposes IFS methods for gene feature selection that perform 1) determining feature length by a filter approach and 2) evaluating each rank list using a wrapper approach. In this way, the accuracy rate of the proposed integrated method is compared with traditional feature selection methods. The result of the experiment depicts IFS methods not only decreases subsets of genes but also increases accuracy.

The rest of this paper is organized as follows. Section 2 represents a summary of the literature reviews. The proposed IFS method and detail is presented in Section 3. Then, the experiments and comparisons with traditional methodologies are shown in Section 4. Finally, concluding remarks is presented in Section 5.

2 Gene Selection and Classification

The background of gene feature selection is based on statistic methods such as *t*-statistics and difference of means. However, the key definition of *t*-statistics is two samples that must be an equal variant. Significant Analysis of Microarray (SAM) is proposed by [MU04]. SAM can complete over a scope of restriction that recognizes genes. Individual gene has SAM scores which based on combination of gene expression that connects with standard deviation of evaluation iteration.

Now, filter approach is the most popular technique based on relationship among attributes. It is also wildly used in feature selection operations. For instance, a gene is merged into a group using relation value of gene that is more than the threshold. After that, it creates top ranked genes from each group. On the other hand, best genes are shown using relation values that are lower than the threshold. Then, Fuzzy Clustering is used to group them for classification [JA03]. A few techniques for gene feature selection are composed of two steps. Firstly, all genes are selected using Cfs technique. Secondly, output of the previous step is chosen by a binary particle swarm optimization algorithm (BPSO). The operation of this technique is the same as the wrapper approach [CH08]. However, some methods are based on distance measurement such as Mahalanobis and Euclidean. Forward Variable Selection Method (FSM) proposed by [HI08] is based on Mahalanobis distance and F-value to find subgroups of genes. This technique compares Simples and S2N techniques.

Combining filter and wrapper approaches are based on Gene Boosting Technique proposed by [JI08]. Firstly, a subgroup of gene is chosen from the top rank using a filter technique. Secondly, subsets of gene via the wrapper approach are based on power of the induction algorithm. The end of the process completes when conditions are met. In contrast, the accuracy rate is not better than the training set. The format of genes on microarray data is small in size and has many dimensionalities. So, the impact of gene feature selection on imbalanced microarray [KA09] is proposed, based on five feature selection methods: Cfs, Chi Square, Info, ReliefF, and Symmetrical Uncertainty integrated four supervised learning: C30-NN, SVM, RF-100, and PART on 11 public microarray data sets. The experiment shows that the SVM approach had higher performance than other techniques. Nevertheless, SVM has low efficiency, if subsets of genes are less than 10. A hybrid model is proposed [RR06] which used subset of genes one by one to separate gene subsets. This model is based on Naïve Bayes, Instance-based Learner (IB1), and Decision Tree (C4.5). But correlation between each gene is ignored. This is the main problem of this model. Comparison of hybrid feature selection models on gene expression datasets is proposed by [PS10]. First of all it selects a subset of gene using Cfs, GR, and Info. Next it transfers output from the previous step into a wrapper approach based on SVM and two heuristic searches which are Greedy Search (GS) and Genetic Algorithm (GA). However, the hybrid model produces a few subsets of gene which leads to decreased accuracy rates.

Classification of Microarrays to Nearest Centroid (CalNC) is proposed by [RD05]. This technique is based on Linear Discriminant Analysis (LDA) and t-statistics value. Subsequently, suitable score is added in CalNC [QS09]. This way, the distance between sample and centroid is evaluated using suitable score that calculates from Mahalanobis distance function.

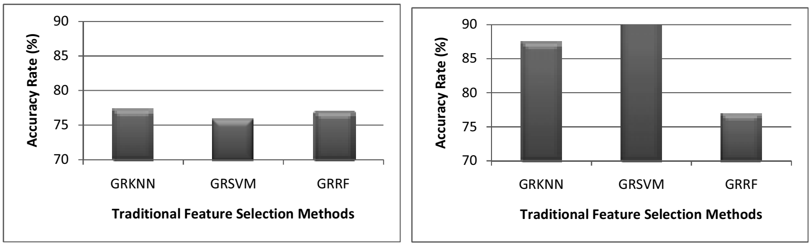
Genetic algorithms are well-known and very beneficial in classification domains. GASVM uses one-objective and the appropriate root to work with genes of 1,000 or less. The concept of Multi-objective strategy comes from a genetic algorithm used to find subgroups of gene. Thus, multi-objective optimization (MOO) concentrates on correlation between assorted objective and a class called MOGASVM [MM09].

One of the major issues in classification is that when dealing with multi-class problems it usually decreases efficiency of classification algorithms. So, [KAA09] proposed a stochastic method called Optimal Feature Weighting algorithm (OFW). This method is based on integrating SVM and CART algorithms. Then a filter and a wrapper approach are used to evaluate public microarray datasets. On the other hand, Pareto-front Analysis can help ranking and measuring of genes [PM09] which calculate restricted classwise between-class and the sum of squares. In addition, it represents more efficiency when working with multi-class of genes on four cancer datasets.

Filter and wrapper approaches gain the most achievement for feature selection but, two examples below show that sometimes they make inferior decisions when picking out feature power.

Example 1. This example uses "Colon Tumor" gene expression dataset that has 61 instances and 2001 attributes. Traditional feature selection steps are as follows: 1) A gene is ranked using GR algorithm 2) the result is sent to a wrapper approach by KNN, SVM and RF 3) output validation using 5-folds cross validation measurement. Original feature selection method can remove 1,865 dimensions and represents as shown in Figure 1. Nevertheless, this method develops correct classification of KNN, SVM, and RF, 77.42%, 75.81%, and 77.08%, respectively (Fig. 1(a)).

Example 2. In the same way, Fig.1 (b) represents accuracy rate of KNN, SVM, and RF, 87.5%, 91.67%, and 77.08% as follows. The influence of traditional feature selection takes away 2,241 attributes from 96 instances and 4027 attributes of "Lymphoma" gene expression dataset.



(a) (b)

Figure 1. The result of classification based on traditional feature selectioin (a) Colon Tumor dataset (b) Lymphoma dataset

3 Integrating Feature Selection Methods (IFS)

At present there are many feature selection techniques in the data mining domain for selecting feature. However, improving performance of feature selection is very essential. In this part, the research proposed IFS methods (Fig. 2) base on combining GR and KNN, SVM, and RF (GRKNN, GRSVM, and GRRF). Then, all models are applied to be selected as a few dimensions based on two public gene expression datasets. The components of proposed IFS method, composed of Phases 1 and 2, are described below:

Phase 1: Filter Approach

Gene expression data determines feature length using the Gain Ratio (GR) method based on forward selection searches. In that case, each subset of rank attribute (one to ten attributes) is created.

GR technique improves the problem of information Gain (Info) algorithm. The structure of this method is created by using to-down design. GR was developed by Quinlan in 1986 and based on evaluation of information theory [RG90]. Typically, probability of v_i answers the ($P(v_i)$) probabilities. So, the information substances (i) of the real answers are given in (1) [PA05]. Info based on Entropy measurement reduces because of the separate method. Entropy at a given node t is given in (2):

$$I(P(v_i),...(P(v_n)) = \sum_{i=1}^{n}(-P(v_i)\log_2 P(v_i)) \qquad (1)$$

$$Entropy(t) = -\sum_{i} p(j \mid t)\log_2 p(j \mid t) \qquad (2)$$

- $p(j|t)$ is associated with frequency of category j at node t.

$$\Delta INFO = Entropy(p) - \left(\sum_{i=1}^{k}\frac{n_i}{n}Entropy(i)\right) \qquad (3)$$

- Info is shown in (3); the parent node, p, is split into k partition; n_j is the number of records in partition i. Nevertheless, bias splits can happen with large partitions. For instance, Student ID attribute is not appropriate for predicting because it contains particular values for individual data. Therefore, it leads to large amounts of results because the number of data is connected with partitions (one by one), which is very difficult to believe predictions.
- *SplitINFO* is presented in (4) to resolve bias in Info. In (5), Info is adapted by using the entropy of the partitioning (*SplitINFO*). Thus, attribute contain high Entropy is adjusted.

$$SplitINFO = -\left(\sum_{i=1}^{k}\frac{n_i}{n}\log\frac{n_i}{n}\right) \qquad (4)$$

$$GR = \frac{\Delta INFO}{SplitINFO} \qquad (5)$$

Steps in Phase 1

Step 1 Let dataset D with features $x_1, x_2,\ldots,x_n$.

Step 2 Calculate $\Delta INFO$ value of attribute x using equation (3)

Step 3 An attribute x is said to be connected with class (C) if GR value of x is > 0.

Step 4 Ranked all attribute value by descending sort; $x_1, x_2,\ldots,x_n$.

Step 5 Define interval index of rank attribute, *i*, is added into each one to ten attributes. So, feature subset D_j with features $x_{i1}, x_{i2},\ldots,x_{i10}$; $j = 1,\ldots,n$.

Phase 2: Wrapper Approach

Next, each subset of rank attribute is evaluated using a wrapper approach such as KNN, SVM, and RF. In this way, discriminating attributes are selected by integrating filter and wrapper approaches together.

Steps in Phase 2

Step 6 Feature subset $D_j = (x_{i1}, x_{i2},\ldots,x_{i10})$ is evaluated using wrapper approach. In particular, heuristic method is used to perform feature selection searches based on each feature subset D_j.

Step 7 Induction algorithm is used to estimate feature subset D_j using cross validation. In practices, induction algorithms are similar to the black box process; users do not need to known procedure details but must be familiar with the interface.

Step 8 Each remained feature subset D_j can be computed repeatedly using definition 6.

For example, the suitable attribute, the suitable attribute plus the next suitable attribute, etc. The final steps, groups of suitable attribute are calculated. For the SVM classifier, this research selects Radial Basis Function (RBF) because it has higher performance than other kernel functions [5]. Moreover, measurement methods focus on precision, recall, F-measure and accuracy rate. Fig. 2 shows the experiment design of the research.

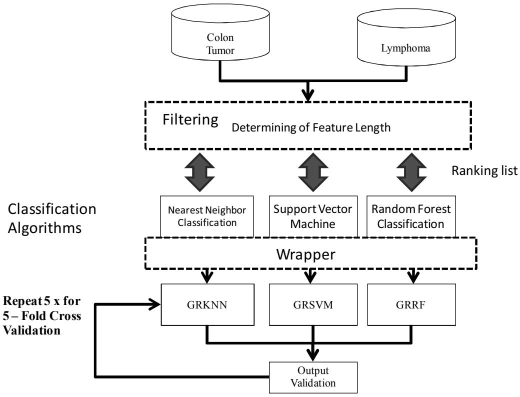

Colon Tumor
Lymphoma
Filtering
Determining of Feature Length
Ranking list
Classification Algorithms
Nearest Neighbor Classification
Support Vector Machine
Random Forest Classification
Wrapper
Repeat 5 x for 5 – Fold Cross Validation
GRKNN
GRSVM
GRRF
Output Validation

Figure 2. Experiment design for comparison of IFS methods for gene feature selection

TABLE I. DETAIL OF GENE EXPRESSION DATASET

Dataset	Instances	Attributes	Class Values
Colon Tumor	62	2001	Positive and Negative
Lymphoma	96	4027	GCL, ACL, DLBCL, GCB, NIL, ABB, RAT, TCL, FL, RBB, and CLL

4 Experiments

In this section, the experiment results depicts that all of IFS methods provide higher value than traditional feature selection methods. First of all, testing based on two class dataset with IFS method and traditional technique. Secondly, verify based multi-class dataset with proposed method. Finally, IFS method and conventional methods are measured using Precision, Recall, F-Measure, and Accuracy rate.

1) Experiment on Colon Tumor Dataset

The details of two public gene expression datasets are shown in Table I. The experimental results on Colon Tumor gene expression dataset are shown in Table II. Using traditional feature selections (GRKNN, GRSVM, and GRRF) there were 135 attributes chosen. The IFS method shows significant results; it is better than old feature selections. As we can see from Table II that it can select much less number of attributes, i.e., 12, 42, and 9, based on proposed model GRKNN, GRSVM, and GRRF, respectively.

2) Experiment on Lymphoma Dataset

Similarly, experiment on traditional dimension reduction with Lymphoma dataset is shown in Table II. In same way, several dimensions are selected, which were 1785 from 4027 attributes. On the contrary, few attributes are chosen using IFS method such as 202, 1622, and 682 based on proposed model GRKNN, GRSVM, and GRRF, respectively.

3) Evaluation measurement between IFS and conventional methods

The efficiency comparing between IFS and conventional methods are based on precision, recall, F-measure, and accuracy rate. IFS method indicates that the performance of this technique is suitable for feature selection. In contrast, the GRKNN model of IFS method has low improvement of performance when selecting a subset of gene. Nevertheless, GRRF model based on traditional feature selection techniques is the best when changed into an IFS method. Moreover, all of model based on the IFS method can reduce feature and still generate better accuracy rates than traditional techniques, shown in Fig.3, Fig. 4 and Table III

TABLE II. COMPARISON SUBSET OF GENE BETWEEN IFS AND TRADITIONAL METHODS

Datasets	Traditional Methods			IFS Methods		
	GRKNN	GRSVM	GRRF	GRKNN	GRSVM	GRRF
Colon	135	135	135	12	42	9
Lymphoma	1785	1785	1785	202	1622	682

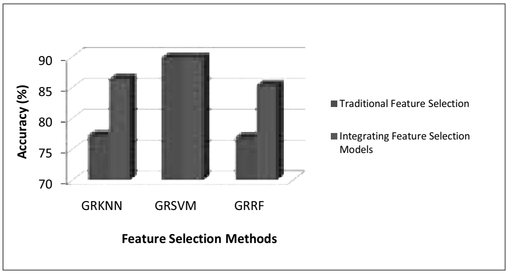


Figure 3. Comparing accuracy rate of traditional techniques and IFS methods on Lymphoma

TABLE III. COMPARISON OF ACCURACY RATE BETWEEN TRADITIONAL AND IFS METHODS

Datasets	Traditional Methods (%)			IFS Methods (%)		
	GRKNN	GRSVM	GRRF	GRKNN	GRSVM	GRRF
Colon	77.42	75.81	77.08	87.1	87.1	88.71
Lymphoma	87.5	91.67	77.08	86.46	93.75	85.42
Average	82.46	83.74	77.08	86.78	90.43	87.07

TABLE IV. COMPARING PRECISION (P), RECALL (R), AND F-MEASURE (F)

Datasets	Efficiency	Traditional Models (%)			IFS Methods (%)		
		GRKNN	GRSVM	GRRF	GRKNN	GRSVM	GRRF
Colon Tumor	P	77.4	75.6	77.5	87.1	87	88.6
	R	77.4	75.8	77.1	87.1	87.1	88.7
	F	77.4	75.7	74.3	87.1	86.9	88.6
Lymphoma	P	87.7	87.2	77.5	84.9	91	84.5
	R	87.5	91.7	77.1	86.5	93.8	85.4
	F	86.6	88.9	74.3	85.5	92.3	84.1

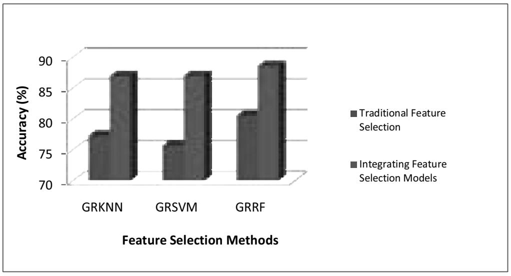


Figure 4. Comparing accuracy rate of traditional techniques and IFS methods on Colon Tumor

Many features are found when using a proposed GRSVM model compared to others. In the the meantime, a few features are chosen when using proposed GRKNN. In Table II, proposed GRSVM model shows higher average accuracy than others but proposed GRRF model shows the best improvement and efficiency when compared to traditional GRRF models. Precision, recall, and F-measure are presented in Table IV. This also reveals that proposed GRSVM model has better performance while proposed GRRF model is enhanced more than other models. Most of the IFS methods can improve accuracy rate for both two class and multi-class, this means it is more flexible when working with a lot of datasets.

5 Conclusion

This research proposed an IFS method that integrate filter and wrapper approaches together. The crucial difference between traditional feature selection and IFS methods are working process. IFS methods determine feature length based on the GR algorithm and forward selection searches. Meanwhile, individual subset of rank feature (zero to ten attributes) is created and then evaluated using a wrapper approach. Proposed model based on IFS methods such as GRKNN, GRSVM, and GRRF is better than the existing methods because they decrease dimension and recognize cancer related feature, also increases accuracy rates. Thus, IFS method is useful and appropriate for implementing a software package model. From the experiments, proposed GRSVM model has higher efficiency over the other models. These IFS methods can eliminate dimensions of Colon Tumor data from 2001 to 11, 41, and 8, respectively and Lymphoma data from 4026 to 201, 1621, and 681, respectively. Furthermore, IFS methods conduct higher average of accuracy rate in which GRKNN, GRSVM, and GRRF: 86.78%, 90.43%, and 87.07% as follows. The trend toward IFS methods generates high accuracy rates over traditional methods. For example, proposed GRSVM and GRRF models. On the other hand, the proposed GRKNN model has an average accuracy rate of 86.78% which is lower than the other IFS methods. In this way, GRKNN model contained 11 and 201 attributes in the Leukemia and Colon Tumor datasets.

In the future, clustering techniques based on proposed method should be developed because most of them produce low accuracy when working with many dimensions; this may deliver higher exactness.

References

[CH08] Cheng-San, Y., C. Li-Yeh, et al, "A Hybrid Approach for Selecting Gene Subsets Using Gene Expression Data," Soft Computing in Industrial Applications, SMCia '08. IEEE Conference, 2008; p.159-164.

[HI08] Hikaru Mitsubayashi, Seiichiro Aso, Tomomasa Nagashima, and Yoshifumi Okada, "Accurate and Robust Gene feature selectionfor Disease Classification Using a Simple Statistic," ISSN 0973-2063 (online) 0973-2063 (print), Bioinformation 3(2), 2008; p.68-71.

[HO08] Hongbo Xie, Uros Midic, Slobodan Vucetic, and Zoran Obradovic, "Hand Book of Applied Algorithms," John Wiley & Sons, 2008; p. 116-117.

[JA03] Jaeger J., R. Sengupta , W. L. Ruzzo, "Improved Gene feature selection for Classification of Microarrays," Pacific Symposium on Biocomputing 8, 2003; p.53-64.

[JI08] Jin-Hyuk H. and C. Sung-Bae, "Cancer classification incremental gene feature selectionbased on DNA microarray data," Computational Intelligence in Bioinformatics and Computational Biology, IEEE Symposium, 2008; p.70-74.

[KA09] Kamal A., X. Zhu, A. Pandya, S. Hsu, and M. hoaib, "The Impact of Gene feature selectionon Imbalanced Microarray Expression Data," Bioinformatics and Computational Biology, 2009; p.259-269.

[KAA09]K.-A. L. Cao, A. Bonnet, and S. Gadat, "Multiclass Classification and Gene feature selectionwith A Stochastic Algorithm," Computational Statistics & Data Analysis, vol. 53, 2009; p.3601-3615.

[MM09] M. Mohamad, S. Omatu, S. Deris, M. Misman, and M. Yoshioka, "A Multi-Objective Strategy in Genetic Algorithms for Gene feature selectionof Gene Expression Data," Artificial Life and Robotics, vol. 13, 2009; p.410-413.

[MU04] Mukherjee, S. and S. J. Roberts. "A Theoretical Analysis of Gene Selection," Computational Systems Bioinformatics Conference, Proceedings, 2004; p.131-141.

[PA05] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining. Addison Wesley," 2005; p.146,207-208.

[PA07] Pádraig Cunningham, "Dimension Reduction," Technical Report UCD-CSI-2007-7, August 8, 2007; p.1-4.

[PL04] P. Lance, H. Ehtesham, and L. Huan, "Subspace Clustering for High Dimensional Data: A Review," SIGKDD Explor. Newsl. 1931-0145, vol. 6, 2004; p.90-105.

[PM09] P. Mundra and J. Rajapakse, "F-score with Pareto Front Analysis for Multiclass Gene Selection," Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, 2009; p.56-67.

[PS10] P. Saengsiri, S.N. Wichian, P.Meesad, and U. Herwig, "Comparison of hybrid feature selection models on gene expression data," in 8th International Conference on ICT and Knowledge Engineering, , 2010; p. 13-18.

[QS09] Q. Shen, W.-m. Shi, and W. Kong, "New Gene feature selectionMethod for Multiclass Tumor Classification by Class Centroid," Journal of Biomedical Informatics, vol. 42, 2009; p.59-65.

[RD05] R. Dabney, "Classification of Microarrays to Nearest Centroids," Bioinformatics vol. 21(22), 2005; p.4148-4154.

[RG90] R. Gray, "Entropy and Information Theory," Springer, 1990; p. 12-18.

[RR06] R. Ruiz, et al., "Incremental Wrapper-based Gene feature selectionfrom Microarray Data for Cancer Classification," Pattern Recognition, vol. 39, 2006, pp.2383-2392.