A New Maximum-Relevance Criterion for Significant Gene Selection

Young Bun Kim[1], Jean Gao[1], and Pawel Michalak[2]

[1] Department of Computer Science and Engineering
[2] Department of Biology
The University of Texas, Arlington, TX 76019, USA
{kim, gao}@cse.uta.edu

Abstract. Gene (feature) selection has been an active research area in microarray analysis. Max-Relevance is one of the criteria which has been broadly used to find features largely correlated to the target class. However, most approximation methods for Max-Relevance do not consider joint effect of features on the target class. We propose a new Max-Relevance criterion which combines the collective impact of the most expressive features in Emerging Patterns (EPs) and some popular independent criteria such as t-test and symmetrical uncertainty. The main benefit of this criterion is that by capturing the joint effect of features using EPs algorithm, it finds the most discriminative features in a broader scope. Experiment results clearly demonstrate that our feature sets improve the class prediction comparing to other feature selections.

1 Introduction

In microarray gene expression analysis, identifying the most representative genes (or features) from tens of thousands of genes in experiments, is critical to improve the prediction performance. Feature selection is one of the important and frequently used techniques in data preprocessing for microarray analysis.

There are three general approaches for feature selection algorithms: filters, wrappers [1] and hybrids [2]. Filter approaches use general characteristics of data to select a subset of features without involving any induction algorithm. Wrapper approaches use estimated accuracy of learning method to obtain feature subsets. Generally, wrapper approaches give higher prediction accuracy but they tend to be more computationally expensive than filter approaches. Hybrid approaches try to utilize different evaluation criteria of two approaches in different search stages. In this paper, we focus on the discussion of filter type feature selections which have better generalization property and can be computed easily and efficiently.

The goodness of feature subset is always determined by a certain criterion. Both Max-Relevance and Min-Redundancy have been instinctively used for this criterion. Max-Relevance is to search features which together have the largest correlation to the target class. Some methods based on statistical tests or information gain have been shown in literature [3], [4]. However this criterion

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 71–80, 2006.

could allow rich redundant genes, which jointly do not contribute to the performance of the prediction because they are highly correlated. Min-Redundancy is a criterion to select mutually exclusive features. An effort to reduce "redundancy" among genes has been recently made in gene selection. Some recent methods propose a criterion by combining the above two constraints effectively [5], [6], [7].

Our work in this paper focuses on maximizing Max-Relevance by considering the joint effect of features (or subspace) on the target class. Methods which have been used for Max-Relevance do not consider the dependency between interactions among features and the target class. However, this may be critical in many circumstances. Based on this fact, we propose a new Max-Relevance criterion, combining the emerging pattern (EPs), one of the recent data mining techniques used to identify interactions among features, [8], [9], [10], and the currently used techniques.

The main contribution of this paper is to show the usefulness of employing interactions among features to explicitly maximize relevancy in feature selection via filter approach. Our comparative experiments demonstrate that the proposed method not only gives higher accuracy than other criteria but also provides comprehensive explanation about relevancy and redundancy of features.

The remainder of the paper is organized as follows. The EPs algorithm employed to identify interactions among features is briefly described in the next section. Subsequently, we present a new MAX-Relevance criterion. Then, we introduce Support Vector Machine (SVM), which is a relatively new and promising class prediction method and show the experimental results on widely-used microarray data sets: colon and leukemia tissues. Finally, we discuss the advantages of our criterion comparing to other criteria and future work.

2 Methods

2.1 Emerging Patterns

Emerging Patterns (EPs) were first introduced by [8] as associations of features (conditions involving several genes in microarray data), whose supports increase significantly from one class to another. They have the special advantage of modeling interactions among genes to build powerful classifiers. The followings are the process to get EPs.

Step 1: Discretization
To efficiently explore the most discriminatory features and to remove the noisy features, we discretize data sets using the entropy based discretization method of [11]. The discretization method partitions genes each into two disjoint intervals. In the example of colon cancer dataset, there are two intervals as (-∞, 59.8) and [59.8, +∞) for M26383 gene. For convenience, we index them as the 1st and 2nd items and so on. So, the emerging pattern {2} represents $\{gene_{M26383}@[59.8,+\infty)\}$.

Step 2: Generating JEPs
We employed Jumping EPs (JEPs) which are defined as the patterns that are found only in one class and have stronger ability to discriminate different classes than any other types of EPs. For example, let's suppose one of JEPs on cancer class is {2, 3}. It represents {$gene_{M26383}$@[59.8,+∞), $gene_{M63391}$@(-∞, 1700)}. And it can be interpreted as :

the pattern that the expression of M26383 is $\geq$ 59.8 and the expression of M63391 $<$ 1700 was found at least one only in cancer samples.

Step 3: Selecting the Most Expressive Features
Finally, we select the Most Expressive Features (ME-features) which are often participating in JEPs. For example, let's suppose that there are 5 JEPs such as {1, 3, 5}, {2, 6}, {4, 9}, {1, 10} and {2, 4, 10}. The 1st ME-gene is {$gene_{M26383}$@(-∞, 59.8), [59.8,+∞)} because the frequency of $gene_{M26383}$ is 4 (1,2,1,2).

2.2 Maximum Relevance

The aim of Max-Relevance is to find features which mutually have the largest correlation to the target class. In developing an approximation method for Max-Relevance, our goal is to effectively catch the joint effect of features on the target class. For this, we employ emerging patterns which have the strong power of modeling interactions among features. The most used notions are defined as follows:

Definition 1. The jumping emerging patterns (JEPs) in dataset of Class$_+$(C_+), denoted $JEPs(C_+)$, are patterns (P) whose supports in C_- are zero but non-zero in C_+. + and - stand for two class labels.

Definition 2. The most expressive $JEPs(C_i)$, denoted $ME - JEPs(C_i)$, are subsets which have the largest supports of all $JEPs(C_i)$.

Definition 3. The most expressive features of Class$_i$, denoted $ME - features$ (C_i), are features within $ME - JEPs(C_i)$.

Definition 4. The collective impact of $ME - features(C_i)$, denoted $D_{C_i}(F)$, is defined as

$$D_{C_i}(F) = \sum_{P} Supp_{c_i}(F), \tag{1}$$
$$F \in \text{ME-features}(C_i), P \in \text{ME-JEPs}(C_i),$$

where $Supp_{c_i}(F)$ is the frequency of occurrence of features(F).

In our approach, we adopt both parametric and nonparametric approach to select discriminative features: t-test (or f-test for multiple classes) and symmetrical uncertainty (SU). The t-test is a statistic criterion based on the

assumption that data comes from some kind of distribution, while SU based on the information-theoretical concept of entropy is a measure of the uncertainty of a random variable. SU is more used than information gain because it can make good for information gain's bias toward features with more values. In some papers, as the combination of several criteria often outperforms an individual criterion, we attempted to take advantages of both criteria. (e.g., the information gain can compensate for the statistical instability of t-test).

The t-test gives the discriminative power of the ith feature as

$$T(F_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}}, \tag{2}$$

where μ_i^+ and μ_i^- are the means of C_+ and C_- for F_i feature, respectively; σ_i^+ and σ_i^- are the corresponding standard deviations; n^+ and n^- indicate the number of samples contained in each class.

The SU is defined as

$$SU(F_i, C) = 2\Big[\frac{IG(F_i|C)}{H(F_i) + H(C)}\Big], \tag{3}$$

where,

$$H(F) = -\sum_i P(f_i)\mathbf{log_2}(P(f_i)), \tag{4}$$

$$H(F|C) = -\sum_j P(c_j)\sum_i P(f_i|c_j)\mathbf{log_2}P(f_i|c_j), \tag{5}$$

$$IG(F|C) = H(F) - H(F|C), \tag{6}$$

$\mathrm{P}(f_i)$ is the prior probabilities for all values of F_i, and $\mathrm{P}(f_i|c_j)$ is the posterior probabilities of F_i given the values of C.

In both criteria, the more F_i and C is correlated, the larger the result value is (e.g. in SU, if F_i and C is completely correlated, SU(F_i, C) is 1). We use average ranks between above two ranks as follows:

$$Rank_M(F) = \mathrm{AVG}\left(Rank_{t-test}(F), Rank_{SU}(F)\right), \tag{7}$$

where the lower the number of Rank_M is, the stronger the discrimination power is (e.g. the most discriminative feature is the feature whose Rank_M is 1).

Finally, by combining the collective impact of features in $ME - JEPs$ and the merged rank of well-known criteria, our new MAX-Relevance criterion is defined as

$$\begin{aligned}\max D(F) &= argmax D(F) \\ &= \frac{1}{2}\left[\frac{1}{N_k}\sum_{C_i}^{K}\frac{D_{C_i}(F)}{\max(D_{C_i}(F))}\right], \\ &+ \frac{1}{2}\left[\frac{N_f - Rank_M(F) + 1}{N_f}\right],\end{aligned} \tag{8}$$

where N_k is the number of classes and N_f is the number of features.

2.3 Minimum Redundancy

The minimum redundancy condition may be defined in several ways [6], [7], [12], [13]. We use Pearson correlation coefficient which is the most well-known measure of similarity between two random variables. The condition is defined as

$$\min R = argminR = \frac{1}{N_f^2} \sum_{i,j} |c(i,j)|, \tag{9}$$

$$where, \qquad c(i,j) = \frac{cov(i,j)}{\sqrt{var(i)var(j)}}, \tag{10}$$

$var(\cdot)$ denotes the variance of a variable and $cov(\cdot)$ represents the covariance between two variables. N_f is the number of features. And we have assumed that both high positive and high negative correlation mean redundancy, and thus take the absolute value of correlations.

2.4 The Minimum Redundancy Maximum Relevance

Ding and Peng proposed the minimum-redundancy-maximum-relevance (mRMR) criterion to minimize redundancy [5]. The idea is to select the genes such that they are mutually maximally dissimilar. The mRMR criterion ($\Phi(D,R)$) has the following simplest form to optimize D (relevance condition) and R (redundancy condition) simultaneously.

$$\max \Phi(D,R), \Phi = D - R. \tag{11}$$

In this paper, we employ this framework because this is a very simple but efficient method [7]. So, based on Eq. (4) and Eq. (5), our minimum-redundancy-maximum-relevance optimization condition is defined as

$$\max_{F_i \in F} D(F_i) - \frac{1}{N_f - 1} \sum_j |c(F_i, j)|. \tag{12}$$

Table 1. Different conditions to search for the next feature

Acronym	Full Name	Formula
t-test	t-test	$\max_{i\in F}[T(i)]$, (T(i) in Eq.(2))
TCD	t-test correlation difference	$\max_{i\in F}\left[T(i) - \frac{1}{N_f - 1}\sum_j \lvert c(i,j)\rvert\right]$
EPMRCD	EPs and merged rank correlation difference	$\max_{i\in F}\left[D(i) - \frac{1}{N_f - 1}\sum_j \lvert c(i,j)\rvert\right]$, (D(i) in Eq. (8))
MRCD	merged rank correlation difference	$\max_{i\in F}\left[\frac{N_f - Rank_M(i) + 1}{N_f} - \frac{1}{N_f - 1}\sum_j \lvert c(i,j)\rvert\right]$,(Rank$_M$(i) in Eq. (7))

2.5 Class Prediction Method

SVM (support vector machine) is a well-known machine learning algorithm based on the structure risk minimization induction principle. Suppose that there are N training samples, $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$, where $\mathbf{x}_i \in R^d$ is a d-dimensional feature vector representing the i^{th} training sample labeled by $y_i \in \{+1, -1\}$ for $i = 1, \cdots, N$. SVM searches for an optimal hyperplane which maximizes margin between two classes. The hyperplane classifying an input pattern $\mathbf{x}$ can be described as the following function :

$$f(x) = \mathbf{sgn}(\mathbf{w}^T\mathbf{x} + b) = \mathbf{sgn}\Big(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b\Big), \tag{13}$$

where $\mathbf{sgn}(k) = 1$ if $k \geq 0$, otherwise -1, Lagrange multipliers $\alpha_i \in [0, C], i = 1, \cdots, N$ and b is a scalar. In this paper, we use the LIBSVM package in [14] and adopt only a linear kernel to compare our results with other works.

3 Experimental Results

We used the well-known datasets, the colon tumor set of [15] and the Leukemia set of [16] to demonstrate the robustness of our new approach. The colon data set contains 40 tumor and 22 normal colon tissue samples of 2,000 genes with highest minimal intensity across the samples. In the leukemia dataset, the target classes are AML and ALL which are subtypes of leukemia and there are 72 samples of 7,129 genes. For the leukemia data, we merged training and test samples together for the purpose of leave-one-out cross validation. In our experiments, we used two different formats as input. We first discretized the data using the entropy based discretization method [11]. This preprocessing step efficiently explores the most discriminatory features with EPs algorithm as well as removes many of the noisy features. Then we normalized the original data so that each gene has zero mean value and unit variance and classified them using SVM. 132 genes in colon dataset and 1026 genes in leukemia dataset were used after discretization.

We measured the classification error rate using Leave-One-Out Cross Validation (LOOCV) to compare the results with Ding and Peng's [5]. Given n samples, LOOCV method constructs n classifiers, where each one is trained with $n-1$ samples, and is tested with the remaining one sample. The final classification accuracy is the average of each classifier.

In our experiments, we compared feature subsets by using three different mRMR optimization conditions in Tab. 1 against the feature sets obtained using t-statistic ranking to pick the top m features. We referred the results of t-test and TCD in [5] to demonstrate the robustness of our proposed criterion. The reason to select TCD (t-test correlation difference) instead of TCQ (t-test correlation quotient) is that TCD is the same scheme as ours.

Table 2. LOOCV errors of colon and leukemia datasets

		The number of features (top m features)													
Data	Method	1	2	3	4	5	6	7	8	10	15	20	30	40	50
Colon	t-test	14	10	9	11	10	9	9	9	10	10	13	10	9	8
	TCD	14	10	8	7	7	7	6	7	8	8	8	8	13	14
	EPMRCD	9	10	9	9	9	10	7	7	**6**	8	8	7	7	7
	MRCD	9	10	9	9	9	10	10	10	8	7	8	7	7	7
Leukemia	t-test	9	3	2	2	2	3	3	4	2	3	3	3	4	1
	TCD	9	3	2	3	3	3	2	4	2	3	5	1	1	1
	EPMRCD	12	6	7	5	3	5	3	2	2	2	1	**0**	**0**	**0**
	MRCD	12	6	6	7	2	5	4	4	3	5	2	1	1	2

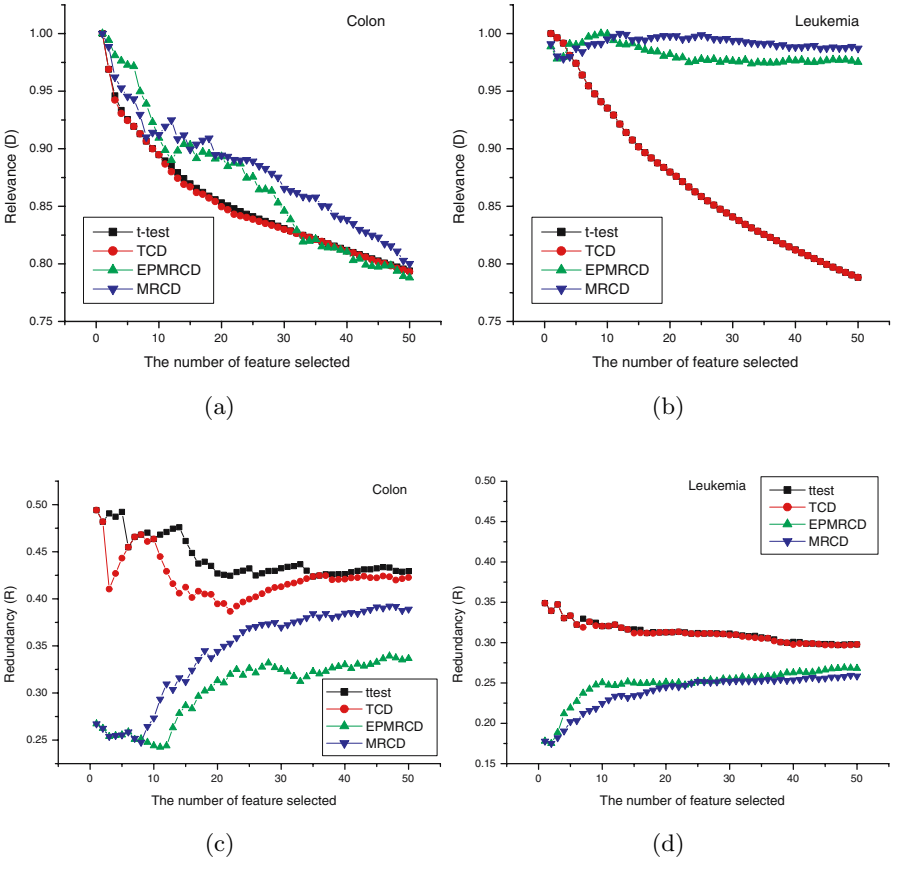
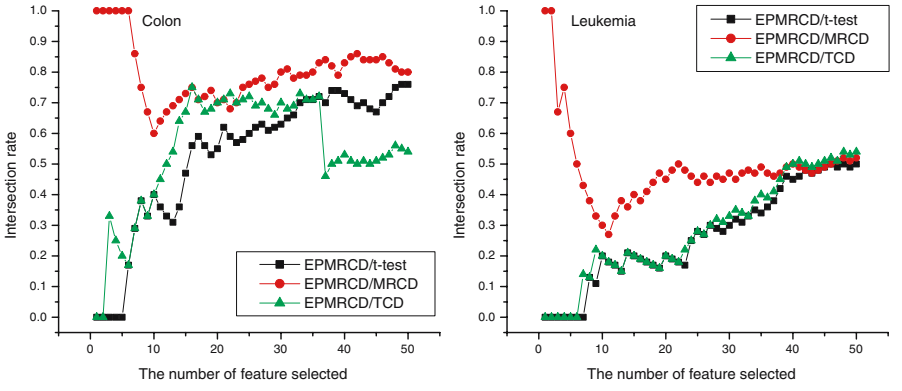


Fig. 1. (a) Relevance on Colon dataset, and (b) Relevance on Leukemia dataset, and (c) Redundancy on Colon dataset, and (d) Redundancy on Leukemia dataset



(a) (b)

Fig. 2. Intersection of features selected using different conditions. (a) Colon dataset (b) Leukemia dataset.

The results of the LOOCV error are shown in Tab. 2. Generally, EPMRCD (EPs and merged rank correlation difference) features outperformed other features. For instance, for colon, EPMRCD leads to 6 errors while t-test leads to 8 errors and MRCD (merged rank correlation difference) leads to 7 errors. And for leukemia, EPMRCD leads to 0 errors while t-test leads to 2 errors and TCD leads to 1 error. However, LOOCV classification error does not provide enough evidence for our efficient criterion.

To demonstrate the effectiveness of our proposed approach, we showed the average relevance (D) and the average redundancy (R) of feature sets in Fig. 1 (refer to Eqs. (7) and (8)) . For colon, although the relevance for EPMRCD reduced as compared to the others, the redundancy also reduced dramatically. Note that both t-test and TCD feature sets show relatively high redundancy. This is more clearly observed in leukemia. For leukemia, the relevance of EPMRCD least reduced relative to others, while the redundancy reduced considerably. In the case of t-test feature set, even relevance reduced impressively according to increase the number of features and within top 10 features, it also shows relatively high redundancy. This results show that the EPMRCD feature set is the most effective one satisfying the Max-Relevance-Min-Redundancy condition.

In order to show how different the EPMRCD feature set is from other features, we also present the rates of intersecting features for the top m ($1 \leq m \leq 50$) features selected as shown in Fig. 2. Features selected using EPMRCD have less overlap with those selected by using t-test or TCD when $m \leq 20$, while they are frequently found in the features selected via MRCD when $m \leq 5$.

Above experiment results demonstrated that even though LOOCV classification error rates were comparable, our criterion found great features, which are dissimilar to those selected by other criteria and are sufficiently satisfying the Max-Relevance-Min-Redundancy condition.

4 Conclusions

In this paper, we presented a new Max-Relevance criterion applied to the minimum redundancy-maximum relevance (mRMR) framework. This criterion is independent of class prediction methods, and thus does not guarantee the best results for any prediction method. The main benefit of proposed criterion is to capture the class characteristics in a broader scope by identifying the joint effect of features and reducing mutual redundancy within the feature set at the same time. Our experiment results showed that proposed criterion generated features which have better generalization property and improve prediction. For example, we achieved 100%, 90.32% LOOCV accuracy in leukemia and colon, respectively, even though we just used the top m features without considering any kind of selection mechanisms. These features also were sufficiently satisfying the Max-Relevance-Min-Redundancy condition relative to other criteria on the same mRMR framework. In the future work, we will apply the EPMRCD feature selection method on multi-class datasets using several prediction methods and verify that it can outperform consistently regardless of the class prediction methods and the number of classes.

References

1. R. Kohavi and G. John: Wrapper for feature subset selection. Arti.Intel. **97(1-2)** (1997) 273–324
2. S. Das, Filters: Wrappers and a Boosting-Based Hybrid for Feature Selection. Proc. 18th Intl Conf. Mach. Learn. (2001) 74–81
3. H. Liu and L. Yu: Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Tran. on Know. and Data engi. **17(4)** (2005) 491–502
4. H.Y. Chung, H. Liu, S. Brown, C. McMunn-Coffran, C.Y. Kao, and D. frank Hsu: Identifying Significant Genes from Microarray Data. Proc. of the fourth IEEE symp. on BIBE. **358** (2004).
5. C. Ding and H. Peng: Minimum redundancy feature selection from microarray gene expression data. J. of Bioinfo. and Comp. Bio. **3(2)** (2005) 185–205
6. L. Yu and H. Liu: Efficient Feature Selection via Analysis of Relevance and Redundancy. J. of Mach. Learn. Rese. **5** (2004) 1205–1224
7. H. Peng, F. Long, and C. Ding. Feature Selection Based on Mutual Information:Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans. on Patt. anal. and mach. intel. **27(8)** (2005) 1226–1238
8. G. Dong and J. Li: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. Proc. of the fifth ACM SIGKDD Inter. Conf. on Know.e Disc. and Data min. (1999) 43–52
9. J. Li, G. Dong and K. Ramamohanarao: Making Use of the Most Expressive Jumping Emerging Patterns for Classification. Know. and Info. Sys. **3(2)** (2001) 131–145
10. J. Li and L. Wong: Identifying good diagnostic genes or genes groups from gene expression data by using the concept of emerging patterns. Bioinformatics. **18** (2002) 725–734, 1407–1408
11. U. Fayyad and K. Irani: Multi-interval discretization of continuous-valued attributes for classification learning. Proc. of the 13th Inter. J. Conf. on Arti. Intel. (1993) 1022–1029

12. P. Mitra, Murthy, and S.K. Pal: Unsupervised Feature Selection Using Feature Similarity. IEEE Tran. of Patt. anal. and mach. intel. **24(2)** (2002) 301–312
13. L. Yu and H. Liu: Redundancy Based Feature Selection for Microarray Data. KDD'04. (2004) 22–25
14. C.W. Hsu, C.J. Lin: A comparison of methods for multi-class support vector machines. IEEE Trans. of Neural Networks. **13** (2002) 415–425
15. U. Alon, N. Barkai, D. Notterman et al: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. of the Nat. Acad. of Sciences. **96(10)** (1999) 6745–6750
16. T.R. Golub, D.K. Slonim et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. **286** (1999) 531–537