2014

A Pairwise Feature Selection Method For Gene Data Using Information Gain

Tian Gui
University of Mississippi

A PAIRWISE FEATURE SELECTION METHOD FOR GENE DATA USING INFORMATION GAIN

A Thesis
presented in partial fulfillment of requirements
For the degree of Master of Science
In the Department of Computer and Information Science
University of Mississippi

by

TIAN GUI

August 2014

ABSTRACT

The current technical practice for doing classification has limitations when using gene expression microarray data. For example, the robustness of Top Scoring Pairs does not extend to some datasets involving small data size and the gene set with best discrimination power may not be involve a combination of genes. Hence, it is necessary to construct a discriminative and stable classifier that generates highly informative gene sets. As we know, not all the features will be active in a biological process. So a good feature selector should be robust with respect to noise and outliers; the challenge is to select the most informative genes. In this study, the Top Discriminating Pair (TDP) approach is motivated by this issue and aims to reveal which features are highly ranked according to their discrimination power. To identify TDPs, each pair of genes is assigned a score based on their relative probability distribution. Our experiment combines the TDP methodology with information gain (IG) to achieve an effective feature set. To illustrate the effectiveness of TDP with IG, we applied this method to two breast cancer datasets (Wang et al., 2005 and van't Veer et al., 2002). The result from these experimental datasets using the TDP method is competitive with the baseline method using Random Forests. Information gain combined with the TDP algorithm used in this study provides a new effective method for feature selection for machine learning.

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my advisors, Dr. Dawn E. Wilkins and Dr. Yixin Chen. Their encouragement, enthusiasm, and faith in me throughout have been extremely helpful. I am also grateful to Dr. H. Conrad Cunningham for being an extraordinary committee member who helped me to get started on the path to my degree and was always available for my questions.

TABLE OF CONTENT

ABSTRACT ii

ACKNOWLEGEMENTS iii

1. INTRODUCTION 1

2. RELATED WORKS 3

2.1 Top Scoring Pairs (TSPs) Classifier 3
2.2 k-TSP Classifier 4
2.3 Hybrid k-TSP+SVM 5
2.4 TSP Decision Tree (TSPDT) 6
2.5 Chisquare-statistic-based Top Scoring Genes (Chi-TSG) 7

3. METHODOLOGY 9

3.1 Datasets 9
3.2 Data Analyzing and Preprocessing 11
3.3 Top Discriminating Pairs (TDPs) 15
3.3.1 Four-Rule based TDP 15
3.3.2 Nine-Rule based TDP 16
3.4 Information Gain 19

4. RESULTS AND DISCUSSIONS 20

4.1 Method of Comparison 20
4.2 Accuracy for Gene Expression Data on Each Approach 21

5. CONCLUSION26

BIBLIOGRAPHY **27**

VITA **29**

LIST OF TABLES

TABLE PAGE

1. Information of gene expression datasets.. 9

2. Feature space reduced cancer gene expression datasets summary.......................... 15

3. Error rate on various classifiers in Wang breast cancer dataset. Perform 5-fold cross-validation on Wang breast cancer dataset and the final accuracy result is averaged from ten 5-fold experiments.. 22

4. Error rate on various classifiers in van't Veer breast cancer dataset. This dataset has separate training and test sets. The error rate on the test set was achieved at the same gene selection level (k) at which the training set obtains the best performance.. 22

5. Comparison of various classifiers in cancer prognosis datasets. This is a direct copy from Shi et al., [2011] paper. 23

LIST OF FIGURES

FIGURE PAGE

1. Comparison of outcome for k-TSP and TSPDT methods. 7
2. Cancer gene expression datasets are not normally distributed. 12
3. Cancer gene expression datasets are normally distributed after preprocessing. 13
4. Distribution of a single gene based on TDP approach across all samples. 14
5. Four-Rule based TDP classifier vs. Nine-Rule based TDP classifier. 18
6. The error rate of *k*-TDP methods on breast cancer datasets. 24

CHAPTER 1

INTRODUCTION

Gene expression classification and feature selection are commonly used techniques to diagnose diseases using microarray technology. In recent years, numerous classifiers have been pursued for correctly identifying cancer tumors based on numerical molecular information. The objective of this study is to find important marker gene pairs to differentiate cancerous samples from non-cancerous ones and build a classifier that can accurately classify the diagnostic cancer subtypes of a sample using microarray expression data. Popular techniques for solving this problem include Support Vector Machine [Vapnik, 1995], Decision Tree [Quinlan, 1993], Prediction Analysis of Microarrays [Tibshirani, 2002], Top Scoring Pair [Geman, 2004], and k-Top Scoring Pair [Tan, 2005]. In fact, there is no evidence to show that there is a single classifier has the best performance over the other methods for all the microarray datasets.

The significance of this paper is to introduce a novel approach that improves classification accuracy of the existing methods with a better selection of informative gene sets. This algorithm is named Top Discriminating Pairs classifiers, simplified as TDP. We can achieve competitive performance by constructing rule-based gene pairs, instead of inspecting individual genes. The classification rules can be constructed using a four or nine set methods. In the next chapter, we discuss and evaluate the existing gene classification methods using microarray datasets. In chapter 3, we introduce our Top Discriminating Pairs (TDPs) classifier. In chapter 4, we discuss the experimental results of our approach on two datasets involving human cancer. Finally, we conclude our results and the advantages and disadvantages of the TDP approach. We start with a brief review of commonly used gene classification techniques.

CHAPTER 2

RELATED WORK

This section introduces the most commonly used related classification methods, which includes Tops Scoring Pairs (TSPs), *k*-TSP, Hybrid *k*-TSP+SVM, TSP Decision Tree, and Chi-TSG. For more detailed tutorials of these methods, we refer readers to Geman et al. [2004], Tan et al. [2005], Shi et al. [2011], Czajkowski et al. [2011], and Wang et al. [2013].

2.1 Top Scoring Pairs (TSPs) Classifier

In gene expression profiles, we consider G genes whose expression levels can be assigned as X = {X_1, X_2, … , X_G}. Each profile X has a true class label in C = {1, 2, … , c}. In our implementation, we only consider two classes (C = 2), either class 1 or class 2. Geman et al. summarized the general process of calculating expression values for each pair of genes – they detected “marker gene pairs” (i, j) under the rule when $X_i < X_j$ from class 1 to class 2 [Geman *et al.*, 2004]. The classification is based on the distinguished pairs and the quantities of interest are,

$$p_{ij}(C) = P\left(X_i < X_j \,\middle|C\right) \qquad (1)$$

The score of each pair of genes is calculated as,

$$\Delta_{ij} = |p_{ij}(1) - p_{ij}(2)| \qquad (2)$$

Then the paired genes are ranked based on the Δ_{ij} values (Eq. (2)) in descending order and the TSP classifier only selects the top scoring pairs.

2.2 *k*-TSP Classifier

The Top Scoring Pairs (TSPs) may change when the training data are perturbed by adding or deleting a few examples [Geman et al., 2004]. In Tan's work, they introduced the *k*-TSP classifier which increases the accuracy of the TSP classifier and generates a more stable classifier. The motivations of using *k*-TSP classifier are: 1) there are many top scoring pairs with the same informative ordering (same score Δ); 2) it combines the discriminating power of many 'weaker' rules; 3) it achieves better combined scores [Tan et al. 2005].

The *k*-TSP algorithm is similar to TSP method. In the prediction of TSP classifier (h_{TSP}), we suppose $p_{ij}(1) > p_{ij}(2)$ and X_{new} is a new sample, Then, the decision rule is [Tan et al. 2005],

$$h_{TSP}(X_{new}) = \begin{cases} C = 1, & if\, X_{i,new} > X_{j,new}, \\ C = 2, & otherwise \end{cases} \qquad (3)$$

The *k*-TSP classifier selects K-top disjoint pairs of genes in prediction according to (3). It simply chooses the class receiving the majority votes and consists of a list of ranked TSPs genes from largest scores to smallest scores in equations (4) and (5),

$$h_{k\text{-}TSP}(X_{new}) = arg \max_{C=C_1,\ C_2} \textstyle\sum_{i=1}^{k} I\,(h_i(X_{new}) = C) \qquad (4)$$

and

$$I(h_i(X_{new}) = C) = \begin{cases} 1, & if\ hi(Xnew) = C, \\ 0, & otherwise \end{cases}, C = \{1, 2\} \qquad (5)$$

Ties are broken by sorting the pairs that achieve the same score Δ using the secondary ranking score Γ (Gamma) [Tan et al. 2005], which is based on the ranking differences in each sample in each class,

$$\Gamma_{ij} = |\gamma_{ij}(1) - \gamma_{ij}(2)| \qquad (6)$$

where the 'average rank difference' is,

$$\gamma_{ij}(C) = \frac{\sum_{n \in S_C}(X_{i,n} - X_{j,n})}{S_c} \qquad (7)$$

S_c denotes the number of samples in class C and the score of the pair of genes is defined in Eq. (6). The k disjoint pairs of genes with the largest score values Γ are selected from those pairs with the highest value Δ_{ij} in TSP classifier (Eq. (2)). Both original TSP and k-TSP techniques perform competitively with prediction analysis of microarrays (PAM) and support vector machine (SVM) classifiers. However, the TSP-family classifiers are easier to interpret and involve fewer genes.

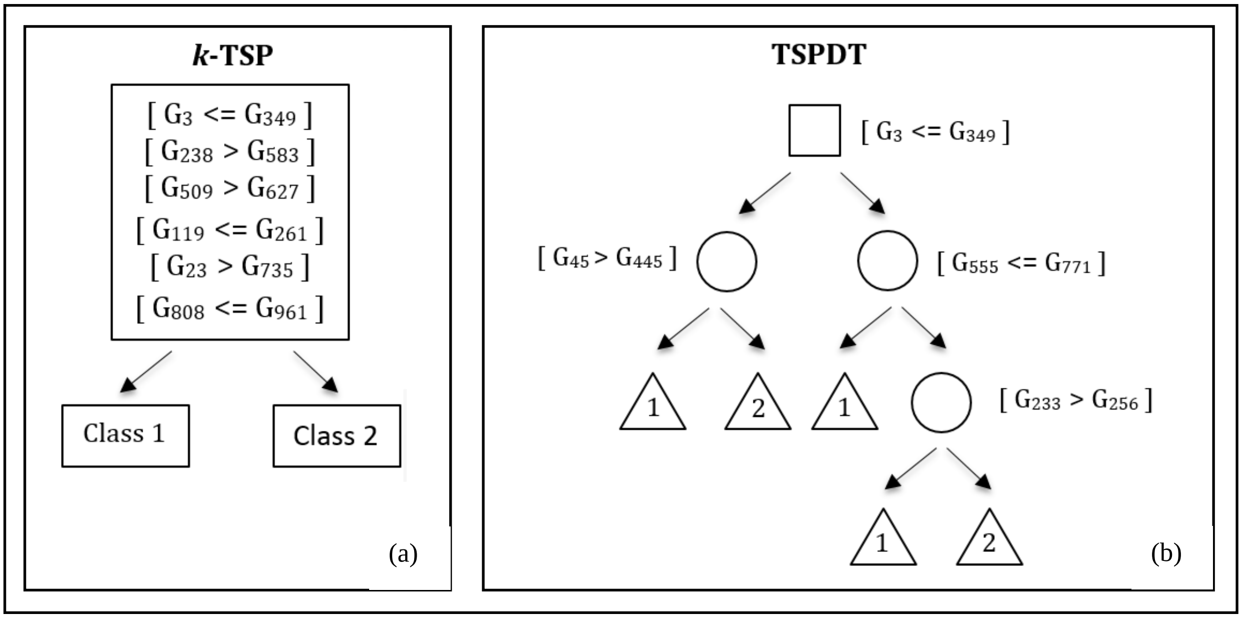
2.3 Hybrid *k*-TSP+SVM

The *k*-TSP technique is computationally efficient and enhances performance for feature selection in machine learning. However, it does not extend to some difficult datasets due to its relatively simple voting scheme [Shi et al. 2011]. For solving this issue, a powerful classifier such as the support vector machine (SVM) is needed. Support vector machines are powerful and elegant linear classifiers [Vapnik, 1998] and also can be extended to nonlinear cases. The examples are represented as points in space by SVM model and mapped with each associated category, thus the examples can be separated as wide as possible with a clear gap. SVMs can efficiently perform linear and nonlinear classification and map the examples into high-dimensional feature spaces [Cortes, 1995]. Shi et al. implemented the hybrid scheme *k*-TSP + SVM, which integrate the k-TSP algorithm with multivariate classifier, SVM. They compared the classification performance of the hybrid scheme with other TSP-family methods involving human cancer datasets. The experiments were repeated 50 times to generate averaged test error rates as they reported in their previous paper [Shi et al. 2011]. The results show the hybrid *k*-TSP+SVM achieves better performances compared with the original TSP, *k*-TSP and SVM techniques on four cancer prognosis datasets.

2.4 TSP Decision Tree (TSPDT)

Czajkowski et al. borrowed the idea of decision trees (DT), which are also known as classification trees and represent one of the main techniques for classification analysis in data mining and knowledge discovery [Czajkowski, 2011]. The approach is based on top-down greedy search. The name of this newly presented approach is TSPDT, which the test attribute is known as the decision node when put all gene information in a tree format. Then each value is separated based on the decision rules as event nodes and each subset goes to the corresponding branches after qualifying each rule and reach the endpoints of the decision tree. The endpoints are known as terminal nodes and each terminal node has an associated terminal value. In the TSPDT method, the terminal value is either 1 (Class 1) or 2 (Class 2). Figure 1 compares the individual performance between the original *k*-TSP algorithm and the TSPDT approach. The comparison is shown using a flow chart which is easier to read and understand. In Figure 1(b), the decision nodes, event nodes and terminal nodes are represented by squares, circles and triangles, respectively [Quinlan, 1986]. This approach is a combination of TSP technique with decision trees, which splits the sample based on pairwise comparisons of its gene expression values [Czajkowski, 2011]. It has been tested on 11 public domain gene expression datasets and the results are promising compared with the original TSP and decision trees classifiers

Figure 1. Comparison of outcome for k-TSP and TSPDT methods.


k-TSP
[G_3 <= G_{349}]
[G_{238} > G_{583}]
[G_{509} > G_{627}]
[G_{119} <= G_{261}]
[G_{23} > G_{735}]
[G_{808} <= G_{961}]
Class 1
Class 2
(a)
TSPDT
[G_3 <= G_{349}]
[G_{45} > G_{445}]
[G_{555} <= G_{771}]
1
2
1
[G_{233} > G_{256}]
1
2
(b)

2.5 Chisquare-statistic-based Top Scoring Genes (Chi-TSG)

One of the challenges in feature selection of cancer expression data is to establish an effective method that can accurately diagnose disease. The existing pairwise classification methods always use an even number of genes and the gene set with the best discriminating power may not be the selected marker gene pairs. An improved classifier, Chisquare-statistic-based Top Scoring Genes (Chi-TSG) is introduced by Wang et al. [2013] and it works for both binary and multi-class classification. Consider a gene expression data of M genes and N samples. The data can be expressed as a matrix of dimension N by M. The expression value of the jth gene in the ith sample can be represented as x_{ij}. To assess whether the marker gene pairs i and j are informative for classification of disease diagnosis, this method redefines the scoring function for gene pairs and the classification rules by incorporating the sample size information in equation (8).

$$x_{ij}^2 = \left(\sum_{q=1}^{2} \sum_{p=1}^{P} \frac{(f_{qpij} - n_p T_q/N)^2}{n_p T_q/N}\right) \qquad (8)$$

Above, the f_{qpij} is the frequency counts of the samples in each class for each pair of genes i and j; n_p is the row totals from the pth row and T_q is the column totals from the qth column. These changes can lead to a better feature selection algorithm and eliminates the concern about bias on preprocessing different samples [Wang et al., 2013].

CHAPTER 3

METHODOLOGY

In this chapter, we introduce our Top Discriminating Pairs (TDPs) classifier starting with data preprocessing and then use it to analyze two human cancer datasets [Wang *et al.*, 2005; van't Veer *et al.*, 2002]. Then we briefly describe the design and implementation strategies for defining classification rules and marker gene pairs. Finally, we adapt the purity measurement, Information Gain (IG), to select the top ranked marker gene pairs with largest information gain.

3.1 Datasets

The datasets we use are from published resources [Wang *et al.*, 2005; van't Veer *et al.*, 2002; Shi *et al.*, 2010] and have been pre-processed by Shi et al. 2010. The sample size, number of genes, number of samples in each class and source are summarized in Table 1

Table 1. Information of gene expression datasets.

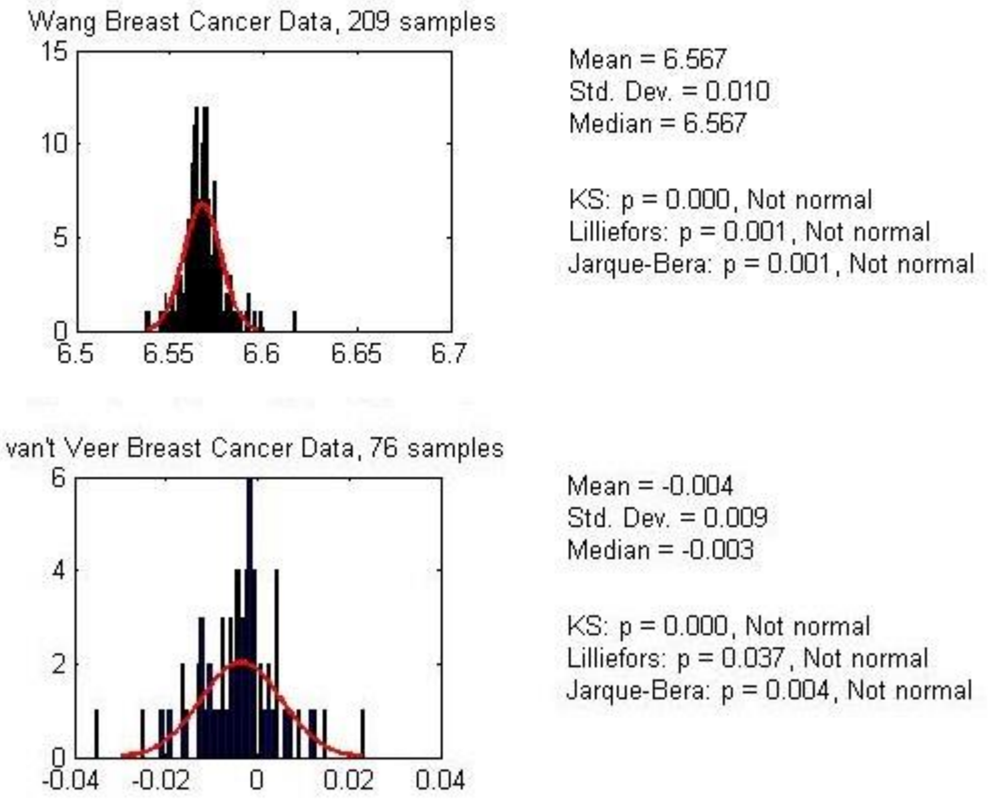
Dataset	No. of samples	No. of genes	Good/Poor prognosis samples	Source
Wang Breast Cancer	209	22283	138/71	Wang *et al.*, 2005
van't Veer Breast Cancer	78/19	23624	51/46	van't Veer *et al.*, 2002

1. *Wang Breast Cancer*: the original dataset is derived from Wang at al. [2005], which contains estrogen-receptor-positive and lymph node-negative patients without receiving any adjuvant treatment. Shi *et al*. [2011] preprocessed the raw intensity Affymetrix CEL files and normalized the data by Robust Multi-array Average (RMA) procedures. The pre-processed expression matrix comprises 209 samples and 22283 features [Shi *et al*., 2010]. It is available at http://math.bu.edu/people/sray/software/prediction.
2. *van't Veer Breast Cancer*: the second dataset is originally obtained from Rosetta Inpharmatics and is also available at http://math.bu.edu. The dataset is already partitioned into training and test sets. In our work, we first apply the training data consisting of 78 samples, 34 have poor prognosis (died) and 44 have good prognosis (remain healthy) for an interval of 5 years after treatment. Shi *et al*. [2011] normalized the raw training data using a log-transformed ratio and removed two samples that contained more than 50% missing values. The final matrix contains 76 samples and 23624 features for the training dataset. The test dataset contains a total of 19 samples with 12 poor diagnosis patients and 7 good diagnosis patients.

3.2 Data Analysis and Preprocessing

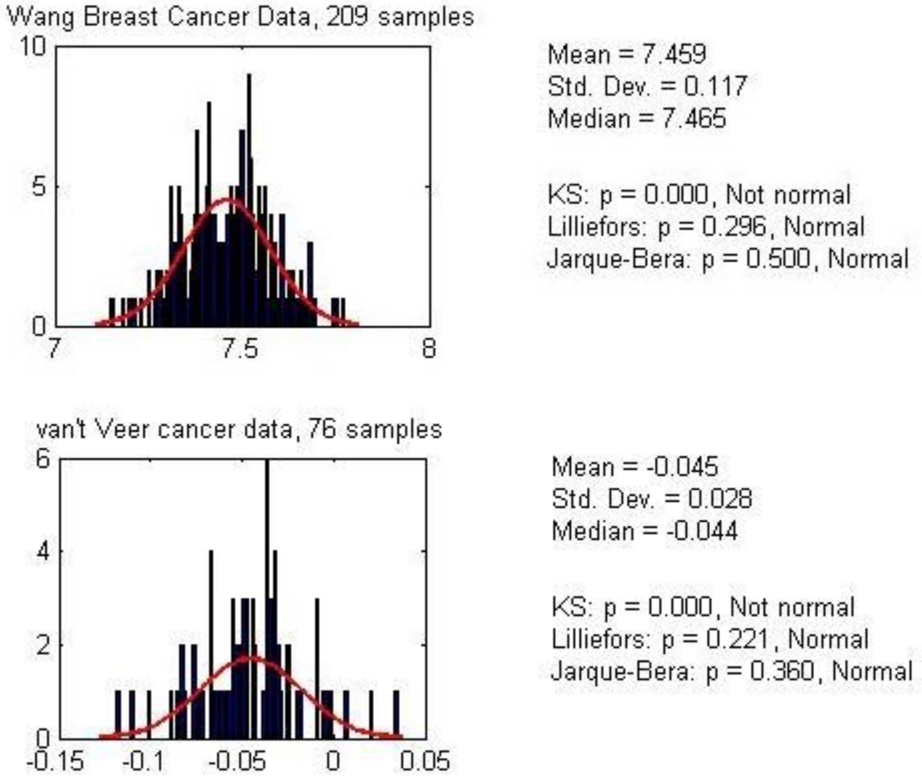
Microarray-based assays of gene expression have become a mainstay of basic and translational cancer research [Nicholas *et al*., 2012]. Scientists commonly assume the gene expression data is distributed normally; this assumption has both empirical [Giles *et al*., 2003; Irizarry *et al*., 2003] and theoretical support. However, the possibility of non-normal distribution for gene expression data presented has been discussed in recent publications [Hardin and Wilson, 2009]. Before we apply our Top Discriminating Pairs classifier to the above two cancer datasets, we first examine the distributions of the entire expression dataset as a whole. Nicholas *et al.* [2012] introduce two related types of expression datasets under the assumption of normality. The first dataset examines the distributions of the complete set of individual expression values across all genes and all samples, which is useful for downstream clustering and class discrimination analyses. The second dataset considers a single gene across the entire range of experimental samples. It is advantageous to provide descriptive behavior of this specific gene over multiple samples [Nicholas *et al*., 2012]. In our pre-processing step, we examine the distributions of a single gene across all samples, which is known as the individual gene level. In Figure 2, the source data for these graphs are averaged gene (the average value of all genes in each sample) across all samples from each cancer dataset. The histograms display the mean, standard deviation, median values on each data. The p-value represents whether the test considers the data do not follow a normal distribution based on the significance level of alpha = 0.05. In other words, if p > 0.05, there is no presumption against the null hypothesis and the data is considered as normal distribution.

Figure 2. Cancer gene expression datasets are not normally distributed. Red curve represents the best fit normal distribution for comparison with the non-normal distribution for each of the datasets on the histogram. Three normality tests are Kolmogorov-Smirnove (KS), Lilliefors, and Jarque-Bera (JB).



The distributions of Wang and van't Veer breast cancer datasets are not normal, as shown in Figure 2. The p-value of the KS, Lilliefors and JB tests are small, which states that the observed data are inconsistent with strong assumption against the null hypothesis. In Figure 3, the graph displays the averaged gene across all samples of preprocessed datasets. After data preprocessing, both the cancer datasets tend to be normally distributed. Two of the three normality tests claim that the observed data is with the assumption that the null hypothesis is true. Only Kolmogorov-Smirnove (KS) contains very strong presumption against null hypothesis. Thus we look closely at the distribution of a single gene, not the averaged gene, across all samples. This single gene is selected based on our Top Discriminating Pair (TDP) algorithm, which is the most informative gene among others (Figure 4). We detect the single gene has a larger standard deviation compared with the averaged gene of both original and preprocessed datasets. We believe the features that differentiate between the two classes should be relatively sparse. Thus, in our approach, we consider the genes with high variance and are known as outliers.

Figure 3. Cancer gene expression datasets are normally distributed after preprocessing. Red curve represents the best fit normal distribution for comparison with the non-normal distribution for each preprocessed datasets on the histogram. Three normality tests are Kolmogorov-Smirnove (KS), Lilliefors, and Jarque-Bera.



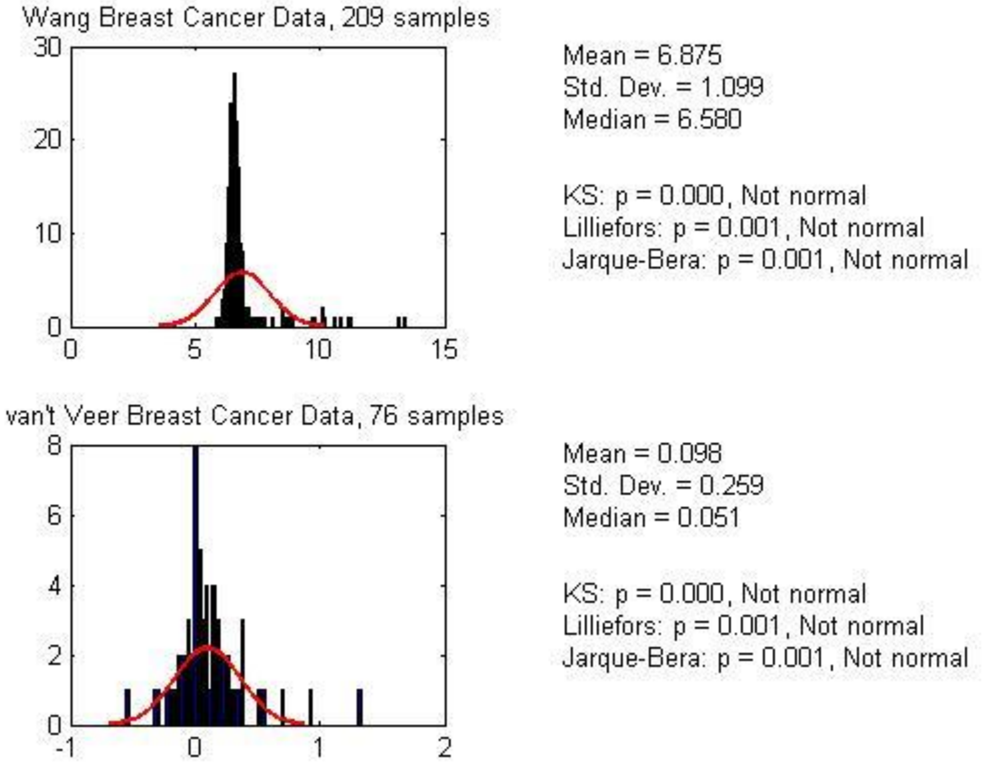
Mean = 7.459
Std. Dev. = 0.117
Median = 7.465

KS: p = 0.000, Not normal
Lilliefors: p = 0.296, Normal
Jarque-Bera: p = 0.500, Normal

Mean = -0.045
Std. Dev. = 0.028
Median = -0.044

KS: p = 0.000, Not normal
Lilliefors: p = 0.221, Normal
Jarque-Bera: p = 0.360, Normal

Figure 4. Distribution of a single gene selected based on TDP approach across all samples.



Our data preprocessing step uses the MATLAB Bioinformatics Toolbox http://www.mathworks.com. We filter out the genes that exhibit variance less than the 75th percentile in their profiles and obtain a final expression data comprising 5665 features for the Wang breast cancer dataset and 6005 features for the van't Veer breast cancer data (Table 2).

Table 2. Feature space reduced cancer gene expression datasets summary.

Dataset	Total no. of genes	No. of genes remained	Percentage remaining
Wang Breast Cancer	22283	5665	25.422%
van't Veer Breast Cancer	23624	6005	25.419%

3.3 Top Discriminating Pairs (TDPs)

For generality, we describe the method in terms of marker gene pairs, which represent the most informative paired genes. Consider a training dataset of M genes whose expression levels can be assigned as X = {X_1, X_2, …, X_M} and a total of N samples {1, …, N}. The data can be represented as a matrix of M by N dimension in which the i^{th} gene expression value of the k^{th} sample is denoted by X_{ik}. Each profile X has a true class label in C = {1, 2, …, C}. In our method, we only consider two classes (C = 2).

$$X_{ik} = \begin{bmatrix} X_{11}, & X_{12}, & \dots, & X_{1N} \\ X_{21}, & X_{22}, & \dots, & X_{2N} \\ \vdots & \vdots & \dots & \vdots \\ X_{M1}, & X_{M2}, & \dots, & X_{MN} \end{bmatrix}$$

3.3.1 Four-Rule based TDP

For each single gene expression value, we define labeling rules based on two conditions first. If X_{ik} is less than or equal to the mean value of individual gene (i^{th} gene) across all samples, then we label X_{ik} as Low, represented by symbol L. Otherwise, X_{ik} is High, represented by symbol H. We call this the Four-Rule based TDP approach.

Before calculating expression values for each marker gene pair, we clarify the comparison rules for every pair of genes i and j,

$$R_{ij}(X) = \begin{cases} LL, & X_{ik} \leq \bar{x}_i,\ X_{jk} \leq \bar{x}_j, \\ LH, & X_{ik} \leq \bar{x}_i,\ X_{jk} > \bar{x}_j, \\ HL, & X_{ik} > \bar{x}_i,\ X_{jk} \leq \bar{x}_j, \\ HH, & X_{ik} > \bar{x}_i,\ X_{jk} > \bar{x}_j \end{cases} \qquad (9)$$

The classification is based on the probability of the distinguished marker gene pairs and the quantities of interest,

$$p_{ij}(C) = \begin{cases} P(R_{ij} = \mathrm{LL}|C), \\ P(R_{ij} = LH|C), \\ P(R_{ij} = HL|C), \\ P(R_{ij} = HH|C) \end{cases},\ C = \{1, 2\} \qquad (10)$$

3.3.2 Nine-Rule based TDP

Similarly, we extend the Four-Rule based TDP approach to a Nine-Rule based method by plugging in the variance and standard deviation. We believe the best informative marker genes would involve the genes overly expressed and down-regulated. In this approach, we are interested in outlier genes and detect those genes based on nine comparison rules in Equation (11),

$$\mathrm{R_{ij}}(X) = \begin{cases} LL, & X_{ik} < \mu - \sigma,\ X_{jk} < \mu - \sigma \\ LN, & X_{ik} < \mu - \sigma,\ \ \mu - \sigma \le X_{jk} \le \mu + \sigma \\ LH, & X_{ik} < \mu - \sigma,\ X_{jk} > \mu + \sigma \\ NL, & \mu - \sigma \le X_{ik} \le \mu + \sigma,\ \ X_{jk} < \mu - \sigma \\ NN, & \mu - \sigma \le X_{ik} \le \mu + \sigma,\ \ \mu - \sigma \le X_{jk} \le \mu + \sigma \\ NH, & \mu - \sigma \le X_{ik} \le \mu + \sigma,\ \ X_{jk} > \mu + \sigma \\ HL, & X_{ik} > \mu + \sigma,\ X_{jk} < \mu - \sigma \\ HN, & X_{ik} > \mu + \sigma,\ \ \mu - \sigma \le X_{jk} \le \mu + \sigma \\ HH, & X_{ik} > \mu + \sigma,\ X_{jk} > \mu + \sigma. \end{cases} \qquad (11)$$

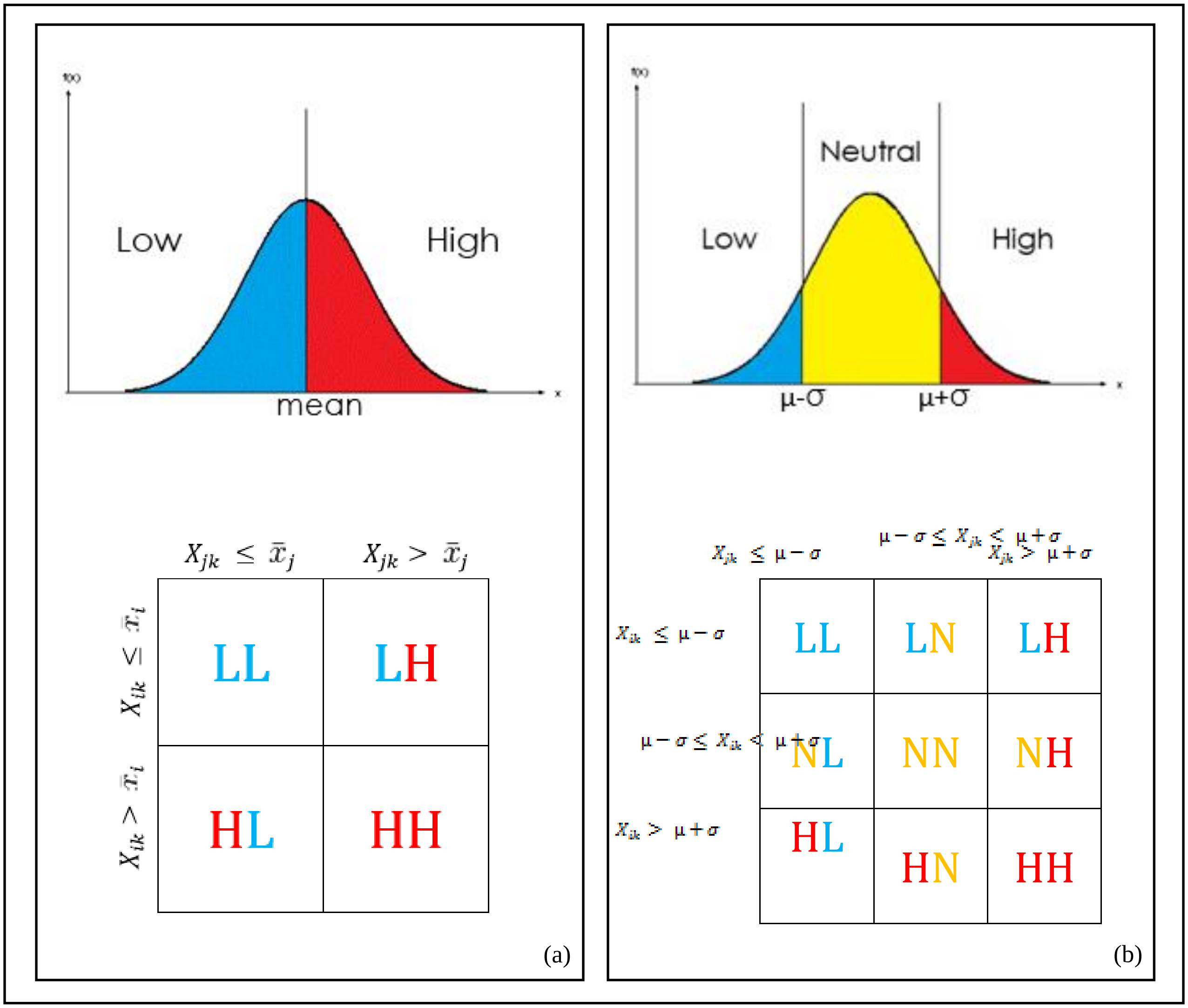
Above, the Low (L) area contains the genes whose expression values are less than the difference of mean and standard deviation. The expression values that fall between the difference and sum of the mean and standard deviation are labeled as neutral (N). The rest of the genes whose expression values are greater than the sum of these two numbers are labeled as high (H).

The probability is estimated by the relative frequencies of occurrences of each classification rules,

$$\mathrm{p_{ij}}(C) = P\big(R_{ij}(X)\big|C\big),\ \ C = \{1, 2\} \qquad (12)$$

In Figure 5, we visualize these two approaches in distribution graphs and table charts. Figure 5(a) is four-rule based TDP classifier. Each gene expression value is compared with the mean value across all samples for an individual gene and labeled into two categories: low or high. Figure 5(b) represents nine-rule based TDP classifier. Each gene expression value is compared with the value of variance ± standard deviation and falls into either low or neutral or high area.

Figure 5. Four-Rule based TDP classifier vs. Nine-Rule based TDP classifier.



	$X_{jk} \le \bar{x}_j$	$X_{jk} > \bar{x}_j$
$X_{ik} \le \bar{x}_i$	LL	LH
$X_{ik} > \bar{x}_i$	HL	HH

	$X_{jk} \le \mu - \sigma$	$\mu - \sigma \le X_{jk} < \mu + \sigma$	$X_{jk} > \mu + \sigma$
$X_{ik} \le \mu - \sigma$	LL	LN	LH
$\mu - \sigma \le X_{ik} < \mu + \sigma$	NL	NN	NH
$X_{ik} > \mu + \sigma$	HL	HN	HH

3.4 Information Gain (IG)

In machine learning, Information Gain (IG) is a measurement of the amount of information in bits about the class prediction [Roobaert *et al.*, 2006]. We use entropy to measure the level of *impurity* and information gain to determine which pair of genes is *most useful* for discriminating between the classes. Equation (13) is known as Shannon entropy [Shannon, 1951], Entropy(X) is defined as,

$$\text{Entropy(X)} = -\sum_{i=0}^{n} p(x_i) log p(x_i) \qquad (13)$$

where $p(x_i)$ is the probability mass function and the overall impurity is the sum of the individual impurities. Information gain measures the expected reduction in entropy [Kullback *et al.*, 1951] and is defined as,

$$\text{Gain} = \text{Entropy (X)} - \text{Entropy (X|Y)} \qquad (14)$$

For each cancer dataset we select the marker gene pairs with highest information gain; basically, it is a score in the range from 0 to 1. The score is the amount of bits of information we have gained about the dataset by choosing each marker gene pair. Thus, the higher the information gain the more effective the marker gene pairs in classifying.

CHAPTER 4

RESULTS AND DISCUSSIONS

In this chapter, we present the experimental performance and obtain the comparison results of all approaches. In section 4.1, we recall the two datasets mentioned in chapter 3. In section 4.2, we apply our proposed approaches to both of the datasets, four-rule based *k*-TDP and nine-rule based *k*-TDP and compare them with baseline method. Finally, we discuss the performance based on the accuracy results in section 4.3.

4.1 Method of Comparison

The performance of our proposed Top Discriminating Pair (TDP) classification method is evaluated on binary class gene expression data. We consider the two breast cancer datasets that were used for assessment of TSP, *k*-TSP and *k*-TSP+SVM classifiers in Shi et al. 2011. The number of classes is 2; class 1 represents the good diagnosis samples and class 2 is poor diagnosis samples. The number of samples per class ranges from 50 to 138.

First, we consider comparison of the baseline method and *k*-TDP for the Wang breast cancer dataset based on 5-fold cross-validation, where each class is partitioned into

5 subsets, the training set is formed from 4 subsets of each class and the remaining subset serves as test data. Another breast cancer dataset is derived from van't Veer et al. 2005, that contains both training and test sets. Beacause, the van't Veer dataset has its own individual test set, the results presented in Table 4 are the error rate on the test set. The model is built from the van't Veer traning set, which contains 76 samples. We only perform 5-fold cross-validation on the Wang breast cancer dataset and the final accuracy result is averaged from ten 5-fold experiments.

Both of the datasets are tested on four learning algorithms, ADTree, BFTree, SVM and Random Forests. Overall, the Random Forests (RF) algorithm has the best performance and ran efficiently. Hence, we compare the baseline method of all features and TDP for feature selection using Random Forests as the class predictor. The Random Forests algorithm uses a combination of tree predictors which generate a random number of trees with the same distribution [Breiman, 2001].

4.2 Accuracy for Gene Expression Data on Each Approach

The classification results for the proposed datasets are shown in Table 3. The parameter *ntree* for RF is optional and we range the number of trees to generate from 10 to 500 based on the number of features we use on each experiment. The parameter *mtry* for RF is the number of variables in each split and should not be larger than the number of features. It is chosen according to the default setting which in the MATLAB Bioinformatics Toolbox is the nearest integer to the square root of the number of total features of the dataset. The randomForest package we use is developed in R by Andy Liaw et al. [2012].

The results below in Table 3 and Table 4 demonstrate the competitive performance of the four and nine rules *k*-TDP against the other approaches. In the Wang breast cancer dataset, the *k*-TDP approach significantly improves the performance, achieving an error rate of 30.9% and 28.8%. In the case where sample size is small or moderate, the *k*-TDP approaches on van't Veer breast cancer dataset achieves similar performance, achieving an error rate of 29.3% and 31.9% with only half the number of the samples of Wang breast cancer dataset. The improvement of four rules *k*-TDP and nine rules *k*-TDP appears constant no matter whether the sample size is small or moderate or large.

Table 3. Error rate on various classifiers in Wang breast cancer dataset.

Error Rate on 10X 5-Fold Cross-Validation (%)							
Dataset	**Random Forests**	**TSP**	**k-TSP**	**SVM**	**k-TSP+SVM**	***k*-TDP (Four)**	***k*-TDP (Nine)**
Wang Breast Cancer	32.6±3.1	41.4±2.5	37.3±2.8	30.1±1.8	32.9±3.0	30.9±2.9	**28.8±2.1**

Table 4. Error rate on various classifiers in van't Veer breast cancer dataset. This dataset has separate training and test sets. The error rate on the test set was achieved at the same gene selection level (*k*) at which the training set obtains the best performance.

Error Rate on the Test Dataset (%)							
Dataset	**Random Forests**	**TSP**	**k-TSP**	**SVM**	**k-TSP+SVM**	***k*-TDP (Four)**	***k*-TDP (Nine)**
van't Veer Breast Cancer	39.2	42.9	28.6	31.6	**10.5**	29.3	31.9

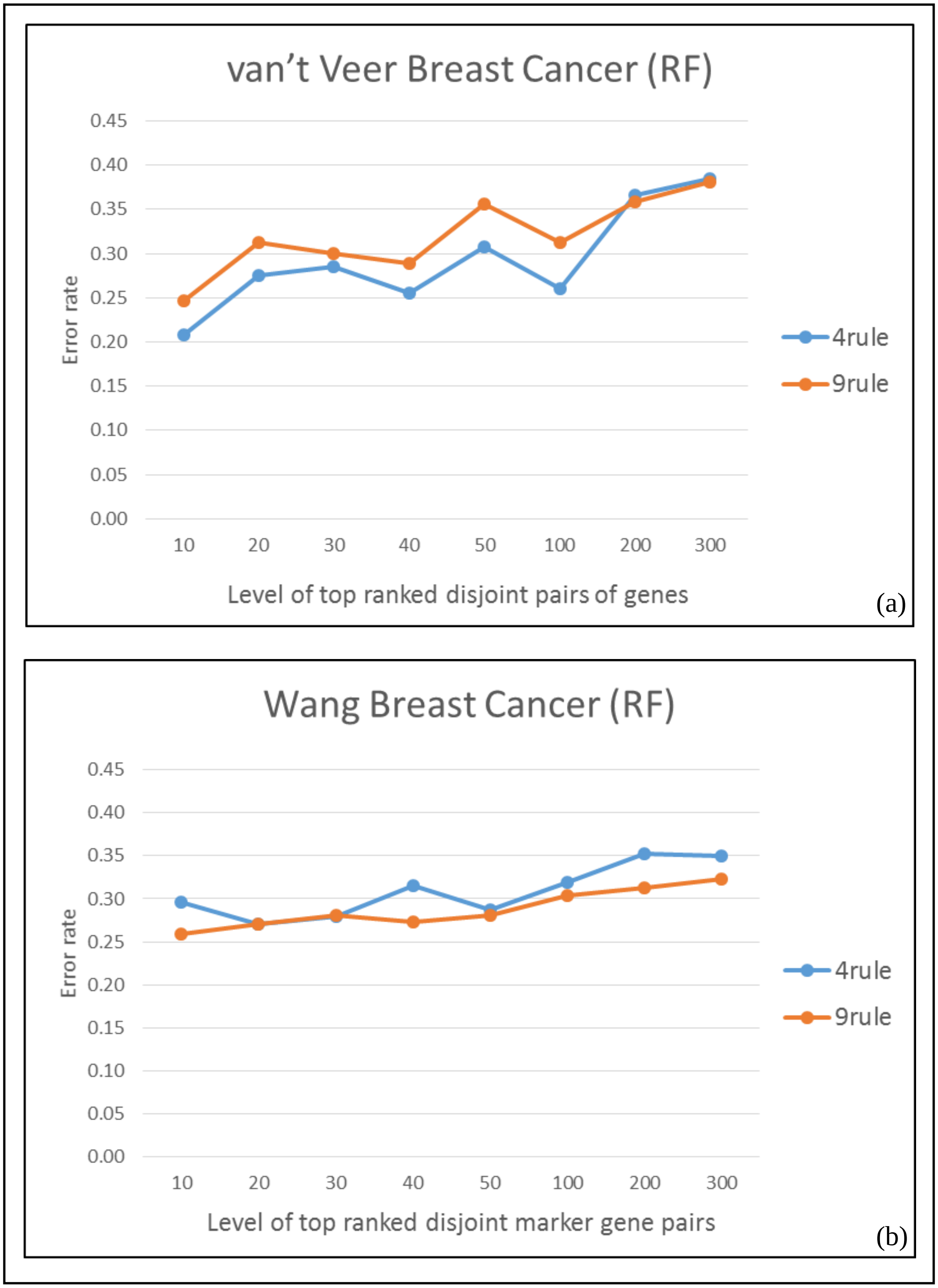
Furthermore, we compare the *k*-TDP approach against other methods such as Random Forests, Top Scoring Pair (TSP), *k*-TSP, SVM, and *k*-TSP+SVM. Random Forests, TSP, *k*-TSP, SVM and *k*-TSP+SVM methods using both the human breast cancer datasets, which are available at http://math.bu.edu. The *k*-TDP method uses the same gene expression data with extra pre-processing, the description of pre-processing given in Chapter 3 section 2. Since our prognosis datasets are directly from Shi *et al.* [2011], we observe that the original results [Table 5] from Shi et al. [2011] publication on *k*-TSP+SVM outperforms *k*-TSP in most cases and the performance on van't Veer breast cancer dataset using TSP is low. However, we ran the TSP technique using the same dataset provided by Shi et al. [2011]; it achieves a better error rate of 42.9% and 28.6%, as compared to 68.4% and 47.3% in Table 5, which was reported in Shi et al. [2011].

Table 5. Comparison of various classifiers in cancer prognosis datasets. In the van't Veer breast cancer dataset where there is an independent test set, the error rate on the test set was obtained at the gene selection level at which the training set achieves its minimum LOOCV error rate. In the other datasets where there is no separate test set, the error rates (mean ± SE) were obtained from two experiments of five-fold cross validation.

Dataset	**Error rate on 2X 5-fold CV (%)**				**Error rate on the test set (%)**			
	TSP	***k*-TSP**	**SVM**	***k*-TSP+SVM**	**TSP**	***k*-TSP**	**SVM**	***k*-TSP+SVM**
van't Veer Breast cancer					68.4	47.3	31.6	10.5
Wang Breast cancer	41.4 ± 2.5	37.3 ± 2.8	30.1 ± 1.8	32.9 ± 3.0				

The number of genes used by each classifier is important when the number of samples is finite [Wang et al. 2013]. The classifier with a small number of genes tends to be more preferred in microarray studies. Hence, we restrict the rest of the discussion to our four rules *k*-TDP and nine rules *k*-TDP classifiers.

Figure 6. The error rate of *k*-TDP methods on breast cancer datasets. The x-axis is the number of top ranked pairs of genes. (a) It shows the error rate on Wang breast cancer dataset at various level of top ranked pairs of genes. (b) It shows the error rate on van't Veer breast cancer dataset at various level of top ranked pairs of genes.



Meanwhile, we plot the error rate of these classifiers with different selection level (k) of disjoint pairs of genes (k = 10, 20, 30, …, 300) in Figure 6. As shown in Figure 6(a), all the classifiers improve their performance when the selection level (k) is small. The best performance of both four rules and nine rules k-TDP classifiers occur when the size of k is less than 50 on Wang breast cancer dataset. Similarly, in the other dataset [van't Veer et al., 2002] the classifiers achieve best performance when the number of marker gene pairs is 10, shown in Figure 6(b).

In general, prognostic datasets are more challenging than the regular diagnostic datasets. The samples with poor and good prognosis usually share the same pathophysiological characteristics [Shi *et al.*, 2011] and the features are relatively sparse to distinguish between the two classes. Our experiments show that compared to other feature selection methods, the TSP family techniques seem not to be successful in all real microarray datasets. This may be caused by the relatively simple voting scheme in choosing the marker genes and the datasets involving small sample size. Hence, we believe that in such cases performance can be improved constantly with k-TDP technique among various sizes of datasets.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In current microarray studies, an effective and stable gene classification method is critical in disease diagnosis. In this work, we integrated the feature selection method of k-TDP with Information Gain and evaluated this combination approach in real human breast cancer datasets. We compared the four rules and nine rules *k*-TDP methods with the baseline method using Random Forests. We also tested the performance of these two approaches with different levels of *k*, the number of disjoint marker gene pairs. In terms of the number of genes used, TDP uses many fewer genes than the baseline method. Also, the error rate increases as the number of genes being selected increases.

The most challenging problems in this work are stabilization and scalability when dealing with large-scale datasets and multi-class classification. Additional work is needed to extend the idea of TDP method and generate a family of TDP algorithms that can handle multiple classes.

BIBLIOGRAPHY

Breiman L. (2001). "Random forests." *Machine Learning* 45: 15-32.

Cortes C., and Vapnik, V. (1995). "Support-vector networks". *Machine Learning* 20 (3): 273.

Czajkowski M., and Marek K. (2011). "Top scoring pair decision tree for gene expression data analysis*." Software Tools and Algorithms for Biological Systems*. Springer New York. 27-35.

Geman D, et al. (2004). "Classifying gene expression profiles from pairwise mRNA comparisons." *Statistical Applications in Genetics and Molecular Biology3.1*.

Giles P.J., and Kipling D. (2003). "Normality of oligonucleotide microarray data and implications for parametric statistical analyses". *Bioinformatics* 19: 2254–2262.

Hardin J., and Wilson J. (2009). "A note on oligonucleotide expression values not being normally distributed". *Biostatistics* 10: 446–450.

Herrero J., Dıaz-Uriarte R., and Dopazo J. (2003). "Gene expression data preprocessing." *BIOINFORMATICS* 19.5: 655-656.

Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., et al. (2003) "Exploration, normalization, and summaries of high density oligonucleotide array probe level data". *Biostatistics* 4: 249–264.

Jarque, C. M., and Bera A. K. (1987). "A test for normality of observations and regression residuals." *International Statistical Review*. Vol. 55, No. 2, pp. 163–172.

Kullback S.; Leibler R.A. (1951). "On information and sufficiency". *Annals of Mathematical Statistics* 22 (1): 79–86.

Liaw A., and RColorBrewer (2012). "Package 'randomForest'." URL http://cran.r-project.org/web/packages/randomForest.

Lilliefors, H. W. (1976). "On the Kolmogorov-Smirnov test for normality with mean and variance unknown." *Journal of the American Statistical Association*. Vol. 62, pp. 399–402.

Massey F. J. (1951). "The Kolmogorov-Smirnov test for goodness of fit." *Journal of the American Statistical Association*. Vol. 46, No. 253, pp. 68–78.

Roobaert D., Karakoulas G., and Chawla. V.N. (2006). "Information gain, correlation and support vector machines." *Feature Extraction*. Springer Berlin Heidelberg. 463-470.

Quinlan J. R. (1986). "Induction of decision trees." *Machine learning* 1.1 (1986): 81-106.

Shannon C. E. (1951). "Prediction and entropy of printed English." *Bell system technical journal* 30.1 50-64.

Shi P., et al. (2011). "Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction." *BMC bioinformatics* 12.1: 375.

Tan A. C., et al. "Simple decision rules for classifying human cancers from gene expression profiles." *Bioinformatics* 21.20 (2005): 3896-3904.

van't Veer Laura J., et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer." *Nature* 415.6871: 530-536.

Vapnik V.N. (1998). "Statistical learning theory". Wiley, New York;

Wang Haiyan, et al. (2013). "TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection." *BMC medical genomics* 6.Suppl 1: S3.

Wang Yixin, et al. (2005). "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." *The Lancet* 365.9460: 671-679

VITA

Tian ‘Tina’ Gui

801 Frontage Rd., Apt. 411, Oxford, MS 38655
(626)319-0896 • tgui@go.olemiss.edu • http://turing.cs.olemiss.edu/~tgui

EDUCATION

May 2016 (expected)	**University of Mississippi, MS** Ph.D. in Computer Science, Overall GPA: 3.70/4.0
May 2014 (expected)	**University of Mississippi, MS** M.Sc. in Computer Science, Overall GPA: 3.70/4.0
Aug 2011	**California State University, CA** B.Sc. in Computer Science, Overall GPA: 3.29/4.0

WORK EXPERIENCE

Jun 13 – Present **Graduate Research Assistant**, University of Mississippi

- Graduate research assistant for Dr. Dawn Wilkins and Dr. Yixin Chen
- Designed and implemented a pairwise feature selection approach for microarray data to improve gene classification accuracy. The results will present at the McBios and ISMB conferences. Programs developed in MATLAB

Aug 12 – May 13 **Graduate Instructor**, University of Mississippi

- Teach one undergraduate course (100 students) in Office Applications
- Developed and implemented all aspects of the courses: syllabi, lectures, homework assignments, exams and grading

Jun 12 – Jul 12 **Graduate Research Assistant**, University of Mississippi

- Summer research assistant for Dr. Dawn Wilkins and Dr. Yixin Chen
- Worked on extending the ability of top scoring pair algorithm using a hybrid method. Designed and tested the accuracy,

sparsity, efficiency and robustness of our model and produced a guaranteed simultaneously selection of highly correlated gene sets.

Aug 11 – May 12 **Teaching Assistant**, University of Mississippi

- Laboratory teaching assistant for fundamental programming classes in Java

Jan 11 – May 11 **Undergraduate Research Assistant**, California State University

- Undergraduate research assistant for Dr. Melissa Danforth;
- Worked on using machine learning and evolutionary computation to derive and analyze attack graphs. Literature reviewed on basic concepts of network security and graphs attack and worked with the group to apply for funding

AWARDS & ACHIEVEMENTS

2014	Travel Grant, 22nd annual international conference on ISMB
2013	First Place, Poster Award, 10th annual McBios conference
2011 – 2015	Graduate Fellowship, University of Mississippi
2009 – 2011	Outstanding and Dean List Student, California State University

HONORS

- Member of Upsilon Pi Epsilon International Honor Society (UPE)
- Member of Association for Computing Machinery (ACM)
- Member of Institute of Electrical and Electronics Engineers (IEEE)
- Member of Society of Women Engineers (SWE)
- Member of Experimental Program to Stimulate Competitive Research (EPSCoR)
- Former Vice President of Chinese Students and Scholars Association (CSSA)