

Microarray Data Integration and Machine Learning Techniques For Lung Cancer Survival Prediction

Daniel P. Berrar*, Brian Sturgeon*, Ian Bradbury, Werner Dubitzky
School of Biomedical Sciences, University of Ulster at Coleraine, BT52 1ST, Northern Ireland
{dp.berrar, b.sturgeon, i.bradbury, w.dubitzky}@ulster.ac.uk

*both authors contributed equally to this paper.

ABSTRACT

The advent of high-throughput technologies such as gene microarrays and protein chips are currently revolutionizing biology and medicine. Machine learning techniques play a pivotal role in analyzing the generated data. Recent studies have reported on the successful application of supervised machine learning approaches to prediction of cancer subclasses, treatment outcome, and drug response. Novel diagnostic tools promise the development of patient-tailored cancer treatment. However, the major step towards individualized therapy is expected to use a combination of various data sources, e.g. transcriptional data, proteomic data, clinical data. In this paper, we have integrated two lung cancer data sets, the Harvard Data Set and the Michigan Data Set. We developed a relational database for the data management and data preprocessing. From the integrated data, we selected 211 lung cancer patients and their transcriptional profile of 3,588 selected genes. Using Kaplan-Meier analysis, we divided the patients into two survival risk groups that are motivated by clinical relevance with respect to survival prediction (5 year survival). We applied six machine learning techniques to predict the survival risk groups. Based on quantitative and qualitative evaluation criteria, we chose decision trees as the most relevant technique for this data set. The comparative study of six machine learning techniques and different subsets of the clinical and transcriptional data revealed that the best overall classification accuracy on unseen test data is obtained by an ensemble of decision trees using boosting (test set accuracy of 77.5%). Integrating clinical and transcriptional data results in a remarkably improved classification performance of the support vector machine. Integrating heterogeneous data did not improve the prediction accuracy of the decision trees (best prediction accuracy is achieved based on anamnestic data alone), but led to the identification of a set of eight potential marker genes that are most important for lung cancer outcome prediction. All these eight genes have been previously reported as marker genes in malignant neoplasia, e.g., breast cancer, glioblastomas, pancreatic cancer, and fibrosarcoma.

Categories and Subject Descriptors

I.5 [Pattern Recognition] I.5.2 [Design Methodology]: Classifier design and evaluation, feature evaluation and selection, pattern analysis.

General Terms

Algorithms, Performance, Design, Experimentation, Standardization, Verification.

Keywords

Microarray, lung cancer, data integration, survival analysis, classification, machine learning.

1. INTRODUCTION

Modern high-throughput technologies produce growing amounts of biomedical data. Transcriptional profiling using microarray technology promises to derive unprecedented insights into the pathogenesis of complex diseases such as cancer. Recent studies on cancer profiling have demonstrated that gene expression patterns of cancer can be successfully used for cancer subtype classification (e.g., leukemias [GOL00], melanoma [BIT00], breast cancer [PER00], prostate cancer [DHA01]), survival prognosis (e.g., childhood leukemia [YEO02], lung cancer [BEE02, BHA01], breast cancer [VEE02]), and response to drug treatment (e.g., [SCH00]). Delineating cancers based on their specific expression profiles may provide the breakthrough required to develop a patient-tailored therapy. Currently, it is unclear how individual patients respond to chemotherapy. Existing chemotherapies have in general severe side effects for the patients, but sometimes low efficacy.

Supervised machine learning techniques are a promising approach for analyzing microarray data in the context of patient outcome prediction. For example, Shipp et al. reported on the successful survival prediction of patients suffering from large B-cell lymphoma [SHI02]. They employed machine learning techniques (support vector machine, k -nearest neighbor) to predict the survival periods of a group of patients. It was shown that the predictive accuracy based on the expression profiles is higher than that based on simple clinical parameters.

Despite the undisputable credentials of microarray technology, transcriptional profiling alone is insufficient to explain the whole spectrum of alterations involved in cancer genesis. Combining gene expression data with proteomic data, cytogenetic data (e.g., FISH data), and clinical patient data might be a promising approach for developing new prognostic tools [BUT00, SLO02, OCH03]. Particularly in the context of cancer outcome prediction, the integration of heterogeneous data sources is considered to be a promising new approach.

The question whether decision support systems based on machine learning approaches and microarray data will find their way into clinical practice is still open, and many other problems remain unresolved. One of the main bioinformatics challenges is the integration of heterogeneous data sources and the development of methods and tools for analyzing high-dimensional microarray data. To adequately organize, manage, analyze, and interpret the

deluge of information, the adaptation of existing and the development of new computational methodologies and tools are required. The task of integrating heterogeneous data usually requires sophisticated data warehousing and database technologies. Many existing and newly evolving machine learning techniques promise to successfully complement classical statistical methods in addressing these challenges. However, the machine learning methods that are currently being used to analyze biological data have not been developed to address the specific requirements of life science applications such as microarrays. First, the analysis of gene expression microarrays is hampered by the high dimensionality of the feature space that often exceeds the sample space dimensionality by a factor of 1,000 or more. Second, the fact that gene expression data are very noisy represents another challenge. The majority of methods used in standard data mining applications are very sensitive to noise. Third, most of the existing methods operate in weak-theory domains, and thus fail to incorporate the enormous body of existing formal background or domain knowledge into the analysis process. This paper is a small step towards tackling these challenges.

In the present study, we have analyzed the Harvard lung cancer data set [BHA01] and the Michigan lung cancer data set [BEE02]. Both data sets were generated on an Affymetrix platform. The data sets comprise a different number of genes and clinical parameters for the patients, but there exist a subset of genes that is contained in both data sets. The Harvard Data Set comprises expression data from 12,600 transcript sequences for 186 patients, including 139 adenocarcinomas. The Michigan Data Set contains 86 primary lung adenocarcinomas as well as 10 non-neoplastic lung samples. Furthermore, various clinical data are provided, such as tumor stage, and anamnestic data (e.g., smoking habits, sex, age).

We are interested in the question whether the combination of both data sets in conjunction with the clinical data can be used to predict the 5-year survival chance of patients. We decided to consider the 5-year survival prediction task because this question is clinically motivated and of practical relevance. The question is: Can we predict the survival risk group of the patients (survival < 5 years or survival ≥ 5 years)? The following tasks are tackled in this study:

- (1) Integration of transcriptional and clinical data;
- (2) Data preprocessing (selection, cleansing, and enrichment);
- (3) Identification of survival risk groups;
- (4) Classification of patients into survival risk groups based on
 - a. clinical data only (*Patient Data Set*);
 - b. gene expression data only (*Expression Data Set*);
 - c. both clinical and gene expression data (*Patient + Expression Data Set*);
 - d. selection of lung cancer specific marker genes only (*Marker Gene Set*);
- (5) Analysis of the classification results;
- (6) Interpretation of the biological relevance.

2. DATA INTEGRATION

The main obstacle to data integration in the scientific domain is the degree of heterogeneity between both data repositories and analysis tools. Conceptually, this heterogeneity can be divided into:

- (1) *Syntactic heterogeneity*, resources are found on different hardware platforms, use different storage paradigms, and utilize different application programming interfaces (APIs);
- (2) *Semantic heterogeneity*, resources use different conceptualizations to model their data.

Within this study the main concern is that of semantic heterogeneity associated with tasks (1) and (2). We developed a relational database to facilitate the integration and preprocessing of the heterogeneous data. The database provided the means to automate these tasks through the use of *SQL*. *SQL* is an example of a *transform-orientated language* and allows the user to:

- (1) Create a database and relation structures;
- (2) Perform basic data management tasks, i.e. insertion, modification, and deletion of data;
- (3) Perform both simple and complex queries to transform the raw data.

We described the data sets using the entity-relationship model [CHE76]. The gene expression data for both sets was integrated using the common attribute *gene name*. Once this had been established we undertook the process of data preprocessing. This step entailed the removal of data where no gene name existed. In total, 3,588 genes are in common in both the Harvard and the Michigan Data Set. We selected these genes only for further analysis. The gene expression data were normalized using *median centering*. Furthermore, we selected the following clinical and anamnestic parameters: age (e.g., 78), sex (e.g., Male), smoking history (e.g., 23 pack-year), TNM classification (e.g., T1N0Mx), tumor stage (e.g., IA), survival time in months (e.g., 58), and censor index (either 1 or 0). For some patients in the Harvard Data Set, no survival information was given. We excluded these patients from further analysis, so that the data set for analysis contained a total of 211 patients.

(Complete description of the database schema, implementation details, and problems with the data integration will be discussed in the final paper.)

We performed a Kaplan-Meier survival analysis on the set of the remaining 211 patients. We excluded all patients that are censored and whose survival time is shorter than 5 years (75 patients). Figure 1 depicts the survival curve for all 211 patients as well as the survival curve for the remaining 136 patients. These are the patients that we included in our study.

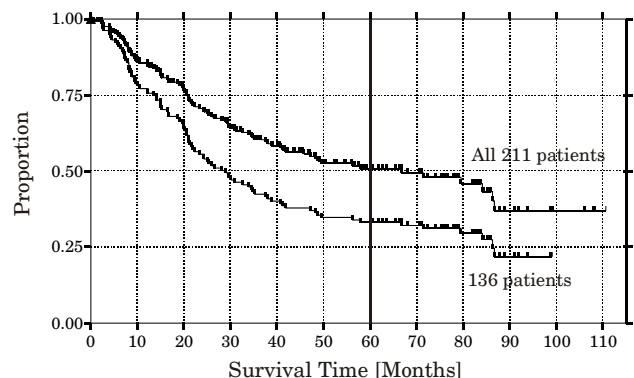


Figure 1. Kaplan-Meier analysis of the patients from the Harvard and the Michigan Data Set.

In the next step, we discretized the continuous values of the patients' survival data into two classes, HIGH RISK and LOW RISK. Patients in the group LOW RISK died before the 5 years mark, whereas patients in the group HIGH RISK survived at least 5 years after diagnosis.

3. DATA ANALYSIS

3.1 Material

We split the entire data set of 136 patients and 3,588 genes randomly into a learning set of 96 patients (70%) and a test set of 40 patients (30%). The learning set comprises 64 (67%) patients of class HIGH RISK and 32 (33%) patients of class LOW RISK. The test set comprises 25 (62.5%) patients of class HIGH RISK and 15 (37.5%) patients of class LOW RISK. The learning set comprises 7 M1-patients, the test set comprises 4 M1-patients (patients with metastasis).

We generated three pairs of learning and test sets: pair #1 comprises the patient data only (age, sex, smoking history, TNM, tumor stage, and the discrete survival class); pair #2 comprises the gene expression data only (gene #1 to gene #3,588, and the discrete survival class); pair #3 comprises both patient data and expression data. Finally, we used a set of eight marker genes whose expression profile highly correlates with the survival group. We refer to this data set as *Marker Gene Set*. These marker genes were identified by the decision tree C5.0 [QUI93], a classifier that is able to detect the most informative genes using an entropy-based criterion.

3.2 Methods

We investigated the classification performance of six state-of-the-art machine learning methods: *decision trees*, ensembles of decision trees using *boosting*, *support vector machines* (SVMs), *probabilistic neural networks* (PNNs), *k-nearest neighbor classifier* (k-NN), and *artificial neural networks* (*multilayer perceptrons*, MLPs). Recent studies have reported successful application of these methods to classification of microarray data, e.g. decision trees in [ZHA01], boosted decision trees in [DUD00, BER01], SVMs in [BRO00, SHI02], PNNs in [BER03a], and k-NN in [SHI02], and MLPs in [KHA01].

We trained the SVM, PNN, and k-NN models with the *leave-one-out cross-validation* procedure on the learning set: each model is trained on all but one sample (*hold-out* case), and then the model is used to predict the class of the hold-out case. This is repeated until each case was used as hold-out case. We trained the decision tree and the MLPs by randomly subsampling the learning set, i.e. by splitting the learning set into a training and a validation set.

We assessed the models' performance on the basis of two criteria. First, based on a quantitative criterion, the classification accuracy with respect to the survival risk groups. Second, on the basis of a qualitative criterion, the output interpretability, i.e. we are interested in the question: "How intelligible is the model's output to humans?"

In the following, we briefly describe the classifiers (We will provide a more detailed description of these methods in the final paper.)

3.2.1 Decision Tree C5.0

Decision tree learning follows a kind of top-down, divide-and-conquer strategy. The basic algorithm for decision tree learning can be described as follows [BER97, ZHA01]:

(1) Select (based on some measure of "purity" such as entropy, information gain, or diversity) an attribute to place at the root of the tree and branch for each possible value of the tree. This splits up the underlying case set into subsets, one for every value of the considered attribute.

(2) Recursively repeat this process for each branch, using only those cases that actually reach that branch. If at any time all instances at a node have the same classification, stop developing that part of the tree.

The measure of purity for the C5.0 decision tree is the information gain. This measure is based on the notion of *entropy*.

Definition 1: Entropy and Information Gain

Given a set of predefined classes, $C = \{c_1, c_2, \dots, c_n\}$, and the indices $1, 2, \dots, i, \dots, n \in I$, and the (training) set X with m observations, $X = \{x_1, x_2, \dots, x_m\}$, with each object, $x_j \in X$, described by k attributes and one class $c_j \in C$, such that $x_j = (x_{1j}, x_{2j}, \dots, x_{kj}, c_j)$. Then the entropy, $entropy(X)$, of the set X relative to this n -wise classification is defined as

$$entropy(X) = \sum_{i=0}^n -p_i \log_2 p_i \quad (1)$$

where p_i is the proportion of X belonging to class $c_i \in C$.

The measure called *information gain*, $gain(X, A)$, is simply the expected *reduction in entropy* caused by partitioning the set of observations, X , based on an attribute A :

$$gain(X, A) = entropy(X) - \sum_{v \in values(A)} \frac{|X_v|}{|X|} entropy(X_v) \quad (2)$$

where $values(A)$ is the set of all possible (discrete) values of attribute A , and X_v is the subset of X for which attribute A has the attribute value v , i.e., $X_v = \{x \in X \mid A(x) = v\}$.

Boosting is a method for combining classifiers. This approach creates several different models and combines their predictions using a weighted voting scheme (e.g., plurality voting). Boosting is an effective method for obtaining classifiers for microarray data with high accuracy [DUD00]. The complete methodology is described in [DUD00, DUB01]. We used SPSS' Clementine implementation of C5.0 in this study [SPS03].

3.2.2 Support Vector Machines

Support vector machines (SVMs) belong to the family of binary statistical classifiers [VAP97, BUR98, SCH99]. The basic principle of a SVM consists of finding the optimal hyperplane between two distinct classes. Multiple hyperplanes exist that separate the cases of the two classes. The best hyperplane can be found by placing a margin around each data point, and by increasing gradually this margin. The SVM finds that hyperplane that results from the largest possible margin. The larger the margin, the better the ability to generalize. Figure 3 illustrates a linear SVM.

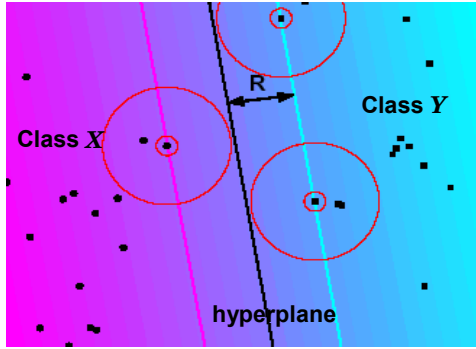


Figure 2: A linear support vector machine.

The SVM in Figure 2 discriminates class X from class Y . The encircled points are called “support vectors”. Only the support vectors define the optimal hyperplane; the margin around the remaining points do not contribute to finding the hyperplane. If the two classes are overlapping, thus not linearly separable, then the SVM projects the data into a higher-dimensional space using a kernel function, and solves the classification problem in the higher dimension. Commonly used kernel functions are the *linear kernel*, the *radial basis kernel*, and the *polynomial kernel*. In this study, we employed all three kernels to build different models of SVMs. We employed a Matlab implementation of SVM for the analysis [CAW00].

3.2.3 Probabilistic Neural Networks

A *probabilistic neural network* (PNN) is the parallel implementation of the *Bayes-Parzen* classifier [SPE90]. The PNN used in this study implements the following decision rule for classification: Let the prior probability that a sample \mathbf{x} belongs to population k be denoted as h_k . The costs associated with a misclassification of a sample belonging to population k is denoted as c_k . The estimated conditional probability that a specific sample belongs to class k , $\hat{p}(k|\mathbf{x})$, is given by the probability density function $\hat{f}_k(\mathbf{x})$. Then an unknown sample \mathbf{x} is classified into class i if

$$h_i \cdot c_i \cdot \hat{f}_i(\mathbf{x}) > h_j \cdot c_j \cdot \hat{f}_j(\mathbf{x}) \quad (3)$$

for all classes $j \neq i$ (*Bayes' decision criterion*).

3.2.4 k -Nearest Neighbor Classifier

The k -NN classifier is based on an *instance-based learning* concept, which is also referred to as *lazy-learning*. *Lazy methods* need to access all learning data at the time when a new case is to be classified. In its simplest implementation, k -NN would compute the similarity between the new case and all learning cases, and the new case is classified as a member of the same class as the most similar case. In this study, we implement a weighted k -NN that takes into account the similarity of the nearest neighbors for classifying a new case. This similarity is translated into a measure of confidence for the classification result. (We will discuss more details of our k -NN implementation in the final paper.)

3.2.5 Multilayer Perceptrons

Multilayer perceptrons (MLPs) belong to the family of artificial neural networks. These networks consist of basic processing units, called *neurons*, that are modeled on biological neurons. The

learning capability of a neural network is defined by the processing function (called *activation function*) of the neurons and the way these elements are connected together, i.e., the network's *topology*. Once a network has been trained on a learning data set, the learned “knowledge” is represented in the network by the *weights* connecting the neurons. In the present study, we use SPSS Clementine's implementation of MLPs and train the networks using the backpropagation algorithm.

3.3 Curse of Dimensionality

Microarray data are characterized by very many variables (genes) with respect to very few observations (samples). This fact is known as *curse of dimensionality* and involves the following problem: All analytical methods rely on a definition of a measure of distance or similarity between objects. However, the definition of similarity/distance between objects in high-dimensional space is not trivial and has not yet received the adequate attention in the context of microarray data. Xing et al. discussed the problem of distance metrics in the context of clustering algorithms and point out that many supervised learning techniques ultimately rely on the choice of the metric [XIN02]. Particularly in high-dimensional data space, the notion of similarity or distance between data points becomes ill-defined. Aggarwal et al. investigated different distance metrics in high-dimensional space [AGG01]. They conclude that the meaningfulness of the distance between objects in high-dimensional space – measured by the commonly used L_k norm (e.g., L_1 = Manhattan distance, L_2 = Euclidean distance) – depends on the chosen value of k . Loosely speaking, the higher the dimensionality, the smaller should be the value for k . For example, the Manhattan distance seems to be consistently more preferable to the Euclidean distance in high-dimensional data spaces such as microarray data. Aggarwal et al. formulate the *fractal distance*, a new concept for measuring the similarity of object in high-dimensional space. The fractal distance is defined as follows (adapted from [AGG01]):

Definition 2. Fractal Distance

Given two vectors \mathbf{x} and \mathbf{y} , both of dimension p , the fractal distance $d_{fract}(\mathbf{x}, \mathbf{y})$ is defined as:

$$d_{fract}(\mathbf{x}, \mathbf{y}) = \left[\sum_{k=1}^p (\mathbf{x}_k - \mathbf{y}_k)^{fract} \right]^{1/fract}, \quad fract \in]0; 1[\quad (4)$$

In the present study, we implement the following distance metrics for the PNN and the k -NN: Euclidean distance, Manhattan distance, Canberra distance, and fractal distance. The distance metric is a parameter of the specific implementations of PNN and k -NN in the present study; these models consider the distance metric an optimization parameter. In the learning phase, the PNN and the k -NN find that distance metric that is most appropriate for the classification problem at hand.

(In the final paper, we will discuss the issue of distance/similarity high-dimensional microarray data in more detail.)

3.4 Results

Table 1 summarizes the classification results for the learning sets and the test sets. We evaluated the classification results based on the performance on the test sets only. The overall best classification accuracy was obtained by the ensemble of boosted decision trees (77.5% on the test set). Enriching the *Patient Data Set* with gene expression data did not result in a classification

improvement for the boosted ensemble. The next best classifier was the SVM that classified correctly 70% of the test cases in the *Patient Data Set*, using a *radial kernel*. For the data set comprising expression profiles only and for the data set comprising both clinical and expression data (*Patient+Expression Data Set*), the SVM classified 75% of the test cases correctly and was the highest-scoring classifier in this study for high-dimensional data. These findings are in accordance with the results of recent studies in which SVMs have been found to be particularly suitable for high-dimensional data [RAM01]. The next best classifier is the PNN, whose classification accuracy on the *Patient Data Set* is the same as that of the SVM. Notice, that the PNN is sensitive to the distance metric chosen. For the *Patient Data Set*, the best results are obtained for the Canberra distance. The next best classifier is the decision tree C5.0, which performed slightly weaker than the PNN on the test set of the expression data. The next best classifier is the *k*-NN. Notice that for all but the *Patient Data Set*, the best distance metric is the *fractal distance*.

To assess the significance of the classification results, we performed a random permutation test. In this test, we randomly permuted the class label (i.e., the risk group) of each patient in the learning set, and train the classifier again. This procedure was repeated 1,000 times to obtain the distribution of the correct classifications under the null hypothesis of random gene expression profiles. This approach is described in [KHA01]. The

distribution of correct classification under the null hypothesis is shown in Figure 3. The number of correct classifications for the unpermuted class labels is 75 (78.1%) (black solid line in Figure 5); this result is statistically significant.

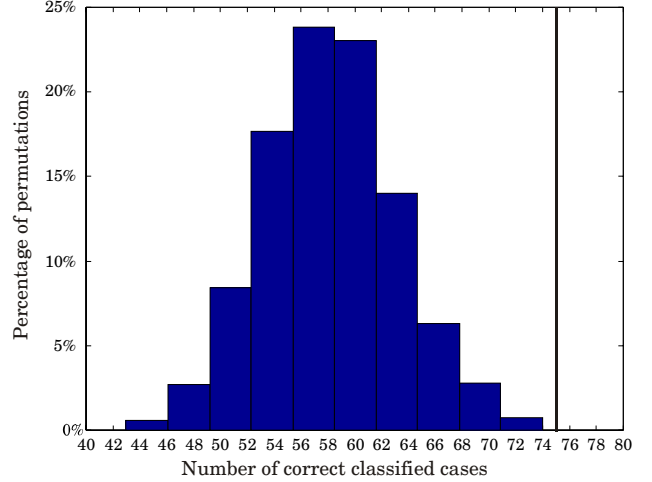


Figure 3: Validation results for randomly permuted class labels of the *Marker Gene Set*. The classifier being used is the support vector machine with radial kernel.

Table 1: Classification results of the six machine learners for the four pairs of learning (L) and test sets (T). Also shown are the parameters of the models that led to the classification results (σ indicates the width of the Gaussian kernel; H1 refers to the first hidden layer, H2 refers to the second hidden layer, H1:5,H2:3 means that the network contains 5 neurons in the first hidden layer and 3 neurons in the second hidden layer; *k* indicates the number of nearest neighbors, *fract* indicates the exponential used in the fractal distance).

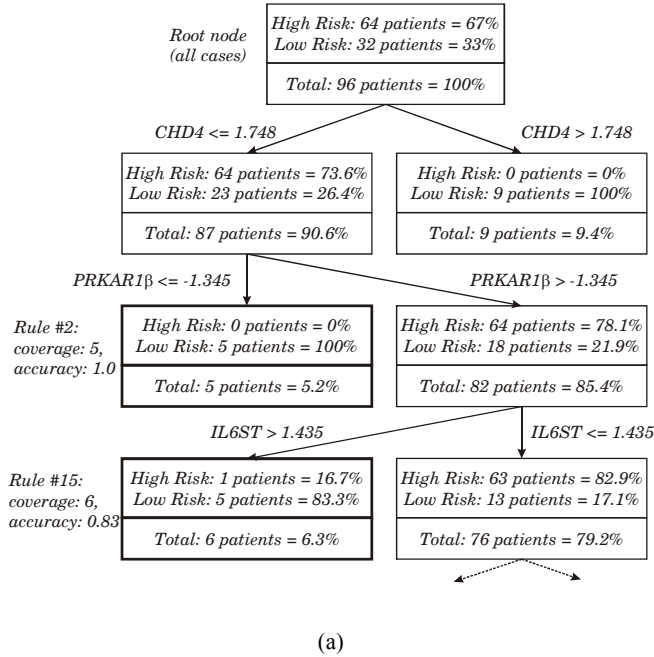
Model	Data Set							
	Patient		Expression		Patient + Expression		Marker Genes	
	L	T	L	T	L	T	L	T
C5.0 (20 fold boosted)	98.0%	77.5%	100%	67.5%	100%	67.5%	100%	67.5%
SVM	74.0%	70.0%	69.8%	75.0%	81.3%	75.0%	78.1%	62.5%
	(C = 10, $\sigma = 2.5$)		(C = 50)		(C = 30)		(C = 10, $\sigma = 0.1$)	
PNN	78.1%	70.0%	72.9%	65.0%	69.8%	65.0%	62.5%	62.5%
	(Canberra, $\sigma = 0.275$)		(Manhattan, $\sigma = 4.3$)		(Manhattan, $\sigma = 2.86$)		(Euclidean, $\sigma = 0.05$)	
C5.0	87.5%	70.0%	95.8%	62.5%	95.8%	62.5%	95.8%	62.5%
	(87.5%)	(67.5%)						
<i>k</i> -NN	66.7%	62.5%	66.7%	67.5%	66.7%	67.5%	80.2%	67.5%
	(Euclidean, <i>k</i> = 14)		(<i>fract</i> = 0.5, <i>k</i> = 10)		(<i>fract</i> = 0.5, <i>k</i> = 10)		(<i>fract</i> = 0.4, <i>k</i> = 5)	
MLP	26.0%	62.5%	33.3%	37.5%	32.3%	37.5%	93.8%	67.5%
	(H1:5;H2:3)		(H1:5;H2:3)		(H1:5;H2:3)		(H1:10)	

The weakest classifier in this study is the MLP. However, given the relatively low accuracy on the learning sets, it can be argued that the chosen network topologies (5 neurons in the first hidden layer and 3 neurons in the second hidden layer for all but the *Marker Gene Set*) is inappropriate for the classification problem at hand. However, we investigated various other topologies, and

their performance is comparable. We cannot exclude the possibility that an appropriate topology exists, for which the MLP would outperform all other classifiers in this study. However, this analysis demonstrates that MLPs are difficult to fine-tune for classifying microarray data, and finding the “right” parameter settings for MLPs is a non-trivial task.

These considerations lead to the qualitative criteria that we should also take into account when we compare different machine learners. In [BER03b], we described a catalogue of features that a classifier for microarray data should have. One of these features is *output interpretability*. Taking this qualitative criterion into account, we believe that the decision tree is the most suitable model in the present study. Decision trees generate classification rules that are easy to understand for humans. Furthermore, decision trees have intrinsic mechanisms of performing feature selection, for example, using *information gain ranking*, i.e. genes are ranked based on their discriminatory power with respect to the survival classes. (This ranking reflects in the hierarchy of the classification rules of the decision tree; cf. Figure 4a). Therefore, decision trees do not require explicit feature selection procedures. However, it is known that other classifiers, such as *k*-NN, benefit from such pre-processing steps.

The results of this comparative study suggest that the choice of the distance metric has a crucial impact of the classification accuracy. For high-dimensional data sets, our models of PNN and *k*-NN seemed to prefer a lower L_k norm.



The learning set comprises 7 M1 patients (with metastasis), and the test set comprises 4 M1-patients. These patients with metastasis have poor clinical outcome. Using the decision tree rules in Figure 4a, we could correctly classify all M1-patients in the test set. An interesting question is whether the expression profiles could be used to correctly predict metastasis, which is an interesting risk outcome.

In the following section, we discuss the output of the decision tree because of the aforementioned properties. The following figure depicts the classification rules of the decision tree C5.0. Figure 4a shows the root node and the first intermediate and leaf nodes of the decision tree; Figure 4b depicts the complete rule set. Also shown are the *coverage* and the *accuracy* of the classification rules. The coverage of a rule is the number of cases in which it is applicable (i.e. in which the antecedent – the *if*-clause – of the rule holds). The accuracy is the number of cases that the rule predicts correctly, expressed as a proportion of all cases it applies to (i.e. the number of cases in which the rule is correct relative to the number of cases in which it is applicable).

- (1) if $CHD4 \leq 1.748$ (total cases: 87), then...
- (2) if $PRKAR1\beta \leq -1.345$ (coverage: 5, accuracy: 1.0), then LOW RISK
- (3) if $PRKAR1\beta > -1.345$ (total cases: 82), then...
- (4) if $IL6ST \leq 1.435$ (total cases: 76), then...
- (5) if $PSG7 \leq 1.73$ (total cases: 72), then...
- (6) if $KIAA0057 \leq 1.959$ (total cases: 68), then...
- (7) if $ITGAE \leq -0.388$ (coverage: 5, accuracy: 0.8), then LOW RISK
- (8) if $ITGAE > -0.388$ (total cases: 63), then...
- (9) if $YWHAE \leq -0.216$ (total cases: 6), then...
- (10) if $ZNF174 \leq -0.41$ (coverage: 3, accuracy: 1.0), then HIGH RISK
- (11) if $ZNF174 > -0.41$ (coverage: 3, accuracy: 1.0) then LOW RISK
- (12) if $YWHAE > -0.216$ (coverage: 57, accuracy: 1.0) then HIGH RISK
- (13) if $KIAA0057 > 1.959$ (coverage: 4, accuracy: 0.75) then LOW RISK
- (14) if $PSG7 > 1.73$ (coverage: 4, accuracy: 0.75) then LOW RISK
- (15) if $IL6ST > 1.435$ (coverage: 6, accuracy: 0.83) then LOW RISK
- (16) if $CHD4 > 1.748$ (coverage: 9, accuracy: 1.0) then LOW RISK

Figure 4: (a) Decision tree generated on the learning set of the expression data (tree not entirely shown); (b) the generated rule set comprising 16 rules sufficient for the correct classification of 95.8% of the learning set cases.

4. DISCUSSION

In the learning set *Expression Data Set* comprising 96 patients, the decision tree C5.0 identified a set of eight genes that are sufficient to achieve 95.8% classification accuracy on the learning set and 62.5% accuracy on the test set comprising 40 samples. Previous studies have associated these genes with cancer, but most of them have not been associated with lung cancer. We discuss these genes in the order of their importance with respect to the classification rules.

CHD4 is chromodomain helicase DNA binding protein 4, also known as Mi2 β , and is located on 12p13. CHD4 is a dermatomyositis-specific autoantigen. Zhang et al. reported that patients with dermatomyositis have a high rate of malignancy, and that chromatin reorganization plays a role in cancer metastasis [ZHA98].

PRKAR1 β (protein kinase, cAMP-dependent, Type I β) has a regulatory function; the encoding gene is located on 7pter-p22. This protein is an essential enzyme in the signaling pathway of the second messenger cAMP. Through phosphorylation of target proteins, PKA controls many biochemical events in the cell

including regulation of metabolism, ion transport, and gene transcription [OMI03]. Leveonson et al. investigated the expression profiles in cell lines of HT1080 fibrosarcoma cells in dependence of drugs inhibiting DNA replication [LEV00]. They found a clearly identifiable group of up-regulated genes in the resistant cell lines. These genes include genes involved in DNA repair and replication and genes for signal transduction proteins, e.g. PRKAR1 β . Interestingly, Bhattacharje et al. also identified a kinase (PRKA) anchor protein 2 as a marker in their study [BHA01]. Beer et al. identified A kinase (PRKA) anchor protein (gravin) 12 as marker gene [BEE02].

IL6ST is a signal transducer and also known as gp130. The gene is located on 5q11. IL6ST is the signal transducer protein for receptors recognizing IL11, leukemia inhibitory factor, ciliary neurotrophic factor, and oncostatin M [OMI03]. Interestingly, this gene is known to play a pivotal role in breast cancer prognosis [GEA03]. Lacroix et al. investigated marker genes for human breast cancer using a low-density microarray, carrying only a few hundreds of capture sequences specific to markers whose importance in breast cancer is generally recognized or suggested by the current medical literature [LAC02]. Lacroix et al. identified IL6ST as a marker gene for human breast cancer.

PSG7 is pregnancy specific beta-1-glycoprotein 7 (locus: 19q13.2). Proteins of this group are molecules that are mainly produced by the placental syncytiotrophoblasts during pregnancy [OMI03]. These proteins comprise a subgroup of the carcinoembryonic antigen family, which belongs to the immunoglobulin superfamily [OMI03]. Ross et al. identified a group of highly expressed markers genes in glioblastomas, including PSG7 [ROS00].

KIAA0057 is a TRAM-like protein of yet unclear function [BAR99]. However, a recent study by Suzuki et al. suggests that KIAA0057 might be a marker gene for juvenile myoclonic epilepsy [SUZ02]. Beer et al. identified a group of KIAA proteins as markers in their study (e.g., KIAA0005, KIAA0020, KIAA0084) [BEE02].

ITGAE is integrin, alpha E (antigen CD103, human mucosal lymphocyte antigen 1; alpha polypeptide). These cell-surface adhesion molecules are known to play a major role in diverse cellular and developmental processes, e.g. morphogenesis, hemostasis, leukocyte activation, cellular adhesion, and homing [OMI03]. This gene has been identified as a marker gene for pancreatic cancer [CRN01]. Clark et al. recently identified ITGAE as a marker gene for breast cancer [CLA02]. They identified ITGAE as one of the top 20 marker genes (ranked by Fisher score). Clark et al. used a support vector machine for classification. The expression profile of the ITGAE gene was one of the most important for the SVM. Interestingly, Bhattacharje et al. identified integrin alpha 3 (CD49C antigen) as a marker gene.

YWHA is Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein (locus: 22q12.3). This protein is a protein kinase-dependent activator of tyrosine and tryptophan hydroxylases (191290, 191060) and an endogenous inhibitor of protein kinase C [OMI03]. Stavridi et al. hypothesis that YWHA enhances the transcriptional activity of p53 [STA01], the well-known tumor suppressor protein that plays a

pivotal role in the DNA damage repair process by activating genes that are responsible for cell cycle arrest and apoptosis [VOG00].

ZNF174 is zinc finger protein 174 (locus: 16p13.3) and is a member of the Cys2-His2 zinc finger family [OMI03]. These proteins are likely to have an impact on the transcriptional repression of the growth factor gene expression [OMI03]. For example, the tumor suppressor gene WT1 encodes a DNA-binding zinc finger protein that downregulates the expression of various growth factor genes (e.g., IGF2 and TGFB1). Li et al. have identified ZNF174 as a marker gene in Burkitt's lymphoma cells [LI03].

Clearly, there exists no 1-to-1 mapping from gene expression to the actual amount of protein being synthesized. However, we might consider the following simplified scenario in which the expression of a gene has a direct impact on the protein synthesis. Then, what are the consequences of an overexpression of ZNF174-encoding genes? The consequences are more ZNF174 proteins that repress the expression of growth factor genes. Interestingly, this is in accordance with the results of the decision tree (rule #10 and #11): overexpression of ZNF174 is associated with LOW RISK, underexpression is associated with HIGH RISK. Moreover, the classification results of, for example, the SVM based on the marker gene set are statistically significant (cf. Figure 3), indicating that the expression profiles of the identified eight genes correlate with survival outcome.

5. CONCLUSIONS

The present study revealed that integrating heterogeneous clinical and transcriptional data might lead to an improved prediction of patient outcome and might provide new insights in cancer biology. Six of the identified eight genes in the present study are known to play a role in cancer, but not all of them have been associated with lung cancer.

Most (if not all) classification and clustering techniques rely on the explicit or implicit definition of a distance metric. The effects of using a distance-metric approach in high dimensional spaces are not fully understood and are potentially problematic. The problem of finding the "right" distance or similarity metric for high-dimensional microarray data has not yet received adequate attention. As mentioned before, the majority of machine learning techniques have been developed in weak-theory domains such as business, retail, and marketing. The problems in these domains are usually characterized by a large number of observations and few variables. This scenario is completely different from the scenario in life science applications where we have a small number of observations (e.g., patients, samples) and a large number of variables (e.g., genes). Therefore, clustering and classification techniques should be tailored to the specific requirements of life science applications, for example, to address the question of similarity in high-dimensional space.

6. REFERENCES

- [AGG01] Aggarwal C.C., Hinneburg A., Keim D.A.: On the surprising behavior of distance metrics in high dimensional space. In *Proc. of 8th Inter. Conf. on Database Theory (ICDT)*, pp. 420-434, (2001).

- [ALI00] Alizadeh A.A et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**:503-511, (2000).
- [BAR99] Barz W.P., Walter P.: Two endoplasmic reticulum (ER) membrane proteins that facilitate ER-to-Golgi transport of glycosylphosphatidylinositol-anchored proteins. *Molecular Biology of the Cell* **10**, pp. 1043-1059, (1999).
- [BEE02] Beer D.G., Kardia S.L., Huang C.C., Giordano T.J., Levin A.M., Misek D.E., Lin L., Chen G., Gharib T.G., Thomas D.G., Lizyness M.L., Kuick R., Hayasaka S., Taylor J.M., Iannettoni M.D., Orringer M.B., Hanash S.: Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* **8**(8), pp. 816-24, (2002).
- [BER01] Berrar D., Granzow M., Dubitzky W., Stilgenbauer S., Wilgenbus K.D.H., Lichter P., Eils R.: New Insights in Clinical Impact of Molecular Genetic Data by Knowledge-driven Data Mining. *Proc. 2nd Int., Conf. on Systems Biology*, Omnipress, pp. 275-281, (2001).
- [BER03a] Berrar D., Downes C.S., Dubitzky W.: Multiclass cancer classification using gene expression profiling and probabilistic neural networks. *Proc. Pac. Symp. on Biocomputing*, **8**, World Scientific, New Jersey/London/Singapore/Hong Kong, pp. 5-16, (2003).
- [BER03b] Berrar D., Downes C.S., Dubitzky W.: A probabilistic neural network for gene selection and classification of microarray data, *Proc. of the International Conference on Artificial Intelligence (IC-AI'03)*, Las Vegas, June 23 - 26, Vol. 1, pp. 342-349, (2003).
- [BER97] Berry M.J. and Linoff G.: *Data Mining Techniques For Marketing, Sales and Customer Support*. John Wiley & Sons, Inc., New York, (1997).
- [BHA01] Bhattacharjee A et al.: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **98**(24):13790-13795, (2001).
- [BIT00] Bittner M. et al.: Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**:536-540, (2000).
- [BRO00] Brown M.P.S., Grundy W., Lin D., Cristianini N., Sugnet C., Furey T., Ares M., Jr., Haussler D.: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* **97**(1):263-267, (2000).
- [BUR98] Burges C.: A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**, p. 159, (1998).
- [BUT00] Butte A.J., Tamayo P., Slonim D., Golub T.R., Kohane I.S.: Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* **97**, pp. 12182-12186, (2000).
- [CAW00] Cawley G.: Support Vector Machine Toolbox. University of East Anglia, School of Information Systems, Norwich, Norfolk, U.K. NR4 7TJ, (2000). Available at <http://theoval.sys.uea.ac.uk/~gce/svm/toolbox/>.
- [CHE76] Chen P.: The entity-relationship model: Toward a unified view of data. *ACM Transactions on Database Systems* **1**(1), pp.9-36, (1976).
- [CLA02] Clark J. et al.: Identification of amplified and expressed genes in breast cancer by comparative hybridization onto microarrays of randomly selected cDNA clones. *Genes Chromosom Cancer* **34**, pp. 104-114, (2002). Gene list available at http://www.crcdmf.icr.ac.uk/array/brcr_table_B.html.
- [CRN01] Crnogorac-Jurcevic T., Efthimiou E., Capelli P., Blaveri E., Baron A., Terris B., Jones M., Tyson K., Bassi C., Scarpa A., Lemoine N.R.: Gene expression profiles of pancreatic cancer and stromal desmoplasia. *Oncogene* **20**(50), pp. 7437-46, (2001). Gene list available at <http://sci.cancerresearchuk.org/axp/mphh/supplementary/pan.html>.
- [DHA01] Dhanasekaran S.M., Barrette T.R., Ghosh D., Shah R., Varambally S., Kurachi K., Pienta K.J., Rubin M.A., Chinnaiyan A.M.: Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**(6849), pp. 822-826, (2001).
- [DUB00] Dubitzky W., Granzow M., Berrar D.: Data mining and machine learning methods for microarray analysis. in Lin S. and Johnson K. (eds): *Methods of Microarray Data Analysis*, Kluwer Academic Publishers, ISBN: 0792375645, 2001, pp. 5-22.
- [DUD00] Dudoit S., Fridlyand J., Speed T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. *Technical Report 576*, Department of Statistics, University of California at Berkeley, Berkeley, CA, (2000).
- [GEA03] GEArray Q Series – Human Breast Cancer and Estrogen Receptor Signaling Gene Array. Available at <http://www.cosmobio.co.jp/Topics/SPA/Data/HS-020.pdf>.
- [KHA01] Khan J., Wei J.S., Ringnér M., Saal L.H., Ladanyi M., Westermann F., Berthold F., Schwab M., Antonescu C.R., Peterson C., Meltzer P.S.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* **7**(6):673-679, (2001).
- [LAC02] Lacroix M, Zammattéo N, Rémacle J, Leclercq G.: A low-density DNA microarray for analysis of markers in breast cancer. *Int J Biol Markers*, **17**(1), pp. 5-23, (2002).
- [LEV00] Leveonson V.V., Davidovich I.A., Roninson I.B.: Pleiotropic resistance to DNA-interactive drugs is associated with increased expression of genes involved in DNA replication, repair, and stress response. *Cancer Research* **60**, pp. 5027-5030, September 15, (2000).
- [LI03] Li Z., van Calcar S., Qu C., Cavenee W.K., Zhang M.Q., Ren B.: A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl. Acad. Sci. USA* **100**(14), pp. 8164-8169, (2003).
- [OCH03] Ochs M.F., Godwin A.K.: Microarrays in cancer: research and applications. *BioTechniques* **34**, pp. S4-S15, (2003).
- [OMI03] OMIM, Online Mendelian Inheritance in Man, online database available at <http://www.ncbi.nlm.nih.gov/Omim/>.
- [PER00] Perou et al.: Molecular portraits of human breast tumours. *Nature* **406**(6797):747-752, (2000).
- [QUI93] Quinlan J.R. : C4.5 : Programs for machine learning. Morgan Kaufmann, San Francisco, (1993).
- [RAM01] Ramaswamy S. et al.: Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* **98**(26):15149-15154, (2001).
- [ROS00] Ross D.T. et al.: Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24**(3), pp. 227-235, (2000).
- [SCH00] Scherf U., Ross D., Waltham M., Smith L., Lee J., Tanabe L., Kohn K., Reinhold W., Myers T., Andrews D., Scudiero D., Eisen M., Sausville E., Pommier Y., Botstein D., Brown P., Weinstein J.: A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* **24**(3):236-244,

- (2000).
- [SHI02] Shipp M.A. et al.: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* **8**:68-74, (2002).
- [SIP03] Available at <http://www.cosmobio.co.jp/Topics/SPA/Data/HS-020.pdf>.
- [SLO02] Slonim D.K.: From patterns to pathways: gene expression data analysis comes of age. The Chipping Forecast II, *Nature Genetics* **32**, Supplement, pp. 502-508, (2002).
- [SPE90] Specht D.F.: Probabilistic Neural Networks. *Neural Networks* **3**:109-118, (1990).
- [SPS03] SPSS Clementine. Available at <http://www.spss.com/clementine>, (2003).
- [STA01] Stavridi E., Chehab N., Malikzay A., Halazonetis T.: Substitutions that compromise the ionizing radiation-induced association of p53 with 14-3-3 proteins also compromise the ability of p53 to induce cell cycle arrest. *Cancer Research*, **61**(19):7030-7033, (2001).
- [SUZ02] Suzuki T. et al.: Identification and mutational analysis of candidate genes for juvenile myoclonic epilepsy on 6p11-p12: LRRCL1, GCLC, KIAA0057 and CLIC5. *Epilepsy Research* **50**(3), pp. 265-75, (2002).
- [VAP97] Vapnik V.: The support vector method. ICANN pp. 263-271, (1997).
- [VEE02] van't Veer L.J., Dai H.Y., van de Vijver M.J., He Y.D.D., Hart A.A.M., Mao M., Peterse H.L., van der Kooy K., Marton M.J., Witteveen A.T., Schreiber G.J., Kerkhoven R.M., Roberts C., Linsley P.S., Bernards R., Friend S.H.: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, pp. 530-536, (2002).
- [VOG00] Vogelstein B., Lane D., Levine A.J.: Surfing the p53 network. *Nature* **408**, pp. 307-310, 2000.
- [XIN02] Xing E.P., Ng A.Y., Jordan M.I., Russell S.: Distance Metric Learning, with application to Clustering with side-information. Proc. of the 16th Ann. Conf. on Neural Information Processing Systems (NIPS), (2002).
- [YEO02] Yeoh E.J. et al.: Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1**, pp. 133-143, (2002).
- [ZHA01] Zhang H., Yu C.H., Singer B., Xiong M.: Recursive partitioning for tumor classification with gene expression microarray data. *Proc. Natl. Acad. Sci. USA* **98**(12), pp. 6730-6735, (2001).
- [ZHA98] Zhang Y. et al.: The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. *Cell* **95**: 279-289, (1998).