Figini, Silvia; Giudici, Paolo

Working Paper

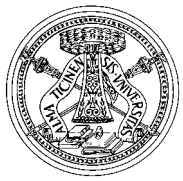
Building predictive models for feature selection in genomic mining

Quaderni di Dipartimento - EPMQ, No. 184

Provided in Cooperation with:
University of Pavia, Department of Economics and Quantitative Methods (EPMQ)

Suggested Citation: Figini, Silvia; Giudici, Paolo (2006) : Building predictive models for feature selection in genomic mining, Quaderni di Dipartimento - EPMQ, No. 184, Università degli Studi di Pavia, Dipartimento di Economia Politica e Metodi Quantitativi (EPMQ), Pavia

This Version is available at:
https://hdl.handle.net/10419/87117



Quaderni di Dipartimento

Building predictive models for feature selection in genomic mining

Silvia Figini
(University of Pavia)

Paolo Giudici
(University of Pavia)

184 (03-06)

Dipartimento di economia politica
e metodi quantitativi
Università degli studi di Pavia
Via San Felice, 5
I-27100 Pavia

Marzo 2006

Building predictive models for feature selection in genomic mining

Silvia Figini *
University of Pavia, Italy

Paolo Giudici
University of Pavia, Italy

Abstract

Building predictive models for genomic mining requires feature selection, as an essential preliminary step to reduce the large number of variable available. Feature selection is a process to select a subset of features which is the most essential for the intended tasks such as classification, clustering or regression analysis. In gene expression microarray data, being able to select a few genes not only makes data analysis efficient but also helps their biological interpretation. Microarray data has typically several thousands of genes (features) but only tens of samples. Problems which can occur due to the small sample size have not been addressed well in the literature. Our aim is to discuss some issues on feature selection in microarray data in order to select the most predictive genes. We compare classical approaches based on statistical tests with a new approach based on marker selection. Finally, we compare the best predictive model with a model derived from a boosting method.

*Address for correspondence: Silvia Figini, Data mining laboratory, department of economics and quantitative methods E-mail: silvia.figini@eco.unipv.it

Keywords: Association models, Boosting, Chi-square selection, Feature selection, Gene expression, Marker Selection, Model Assessment, Predictive models.

1 Introduction

Machine learning techniques have been used for many pattern recognition problems. A large variety of learning techniques have been studied and proved to be useful. For example, Decision Trees, Nearest-Neighbor, Support Vector Machines, and Neural Networks are all widely used methods in many different fields. These methods often have applications in a variety of areas such as bioinformatics, robotics, and vision. The general problem, however, is similar regardless of the learning technique and area of application. Given a data set with a number of samples, each of which with a corresponding set of feature values and a classification, we want to find a rule or model to classify each sample according to its feature values. The existing learning techniques work well for most instances of this problem. However, when the number of samples or the number of features in the data is large, the performance of the learning methods degrades. The samples may become noisy and unclassifiable, or the features may become irrelevant to the classifications. Many authors, see e.g. [1] discuss the problem of selecting relevant features, and the problem of selecting relevant samples on data sets containing large amounts of irrelevant information. For large data sets, we can usually choose only a few of the most relevant features to build a model to classify the data. The resulting model will be at least as good as the one built from all the features. Hence it is often useful to select a subset of features of a data set to describe the data. Many feature selection algorithms have been devised for this task, see e.g. [2]. In this paper, we focus on data sets with many features and a few samples.

We discuss the different approaches for this problem, and review some of the work on dealing with data sets with many features. In Section 2 we present a review of statistical methods for feature selection. In Section 3 we describe our proposed method for feature selection and in Section 4 our proposed predictive models. Finally in Section 5 we present the application of our methods to the available data.

2 Feature selection: a review

The basic feature selection problem is an optimization problem, with a performance measure for each subset of features to measure its ability to classify the samples. The problem is to search through the space of feature subsets to identify the optimal or near-optimal one with respect to the performance measure. Feature selection is a fundamental process within many classification algorithms. A large dictionary of potential, flexible features often exists, from which it is necessary to select a relevant subset.

Feature selection is generally an empirical process that is performed prior to, or jointly with, the parameter estimation process. Many successful feature selection algorithms have been devised. Yang and Honavar [3] classify many existing approaches into three groups: exhaustive search, heuristic search, and randomized search.

Exhaustive search is a brute force approach where every possible subset is tested with the performance measure, and the best one is chosen. It guarantees the optimal subset as a result. However, if the number of features is large, this approach is intractable.

Heuristic search is where certain heuristics are used to greedily but intelligently search through the subset space to identify a subset with a reasonable performance measure. Forward Selection and Backward Elimination [4] are two examples of heuristic search. Forward Selection starts with the empty

set of features. It evaluates all the one-feature subsets, and selects the one with the best performance measure. It then evaluates all the two-feature subsets that include the feature already selected from the first step, and selects the best one. This process continues until extending the size of the current subset leads to a lower performance measure. Backward Elimination starts with the complete set of features. It evaluates all the subsets that are one less than the complete set, and selects the one with the best performance measure. It then evaluates all the subsets with one less feature than the best subset from the previous step, and selects the best one. The process stops when decreasing the size of the current best subset leads to a lower performance. There are many variants of Forward Selection and Backward Elimination.

Randomized search uses randomized or probabilistic methods to search through the subset space. Genetic Algorithms [3] and Scatter Search algorithms [5] are examples of this approach. Neither heuristic search nor randomized search techniques guarantee optimal results.

Another common way to classify feature selection algorithms is determined by how the learning method is integrated into the algorithm. A filter approach is where the selection of features is independent of the learning algorithm. On the other hand, if the features are generated and directly evaluated by a learning algorithm, the method is known as a wrapper approach. Kira and Rendell's Relief algorithm [6] is an example of a filter approach. It uses a procedure that is independent of a learning algorithm to assign weights to each feature. Then the features are selected based on whether or not it is above a pre-specified threshold value. An example of a wrapper approach is the Genetic Algorithm. Each subset of features is evaluated using a learning algorithm in order to progressively generate new and better subsets.

The feature selection algorithms discussed above have been tested to work

well for many problem domains. In the field of high-dimensional gene expression data Xing, Jordan, and Karp [7] report on their use of a hybrid of filter and wrapper approaches to a data set with 72 samples and 7129 features. Golub et al. [8] describe a method for classification and prediction for the same data set. Both groups achieved reasonable results for their method's ability to classify new samples. Xing et al. [7] specifically argue that the reduced set of features perform significantly better than the full set when used in classifying the samples. Califano, Stolovitzky, and Tu [9] report on their use of a supervised learning algorithm to identify patterns in gene expression data. Their data has 6817 genes or features, and their method give reasonable results for classifying the samples. Breiman [10] describes the use of Random Forests, a forest of decision trees built using a randomized process, to classify the samples in a data set.
In this paper we present a new method in feature selection based on marker selection, see e.g. [12] and we compare our approach with classical approaches based on chi-square selection.

3 Feature selection for genomic data

DNA microarrays have been used by biologists to monitor the level of gene expression of thousands of genes in different biological tissues. This technology measures the expression of a gene in a cell by measuring the amount of mRNA present for that gene, mRNA which is extracted from samples of human tissues. A target sample and a reference sample are hybridized with DNA. The log intensities of mRNA hybridizing is measured for a few thousand genes. These numbers, that normally range beetwen -6 and +6, represent the expression level of each gene in the target, relative to the reference sample, so that, for example, positive values indicate higher expression in the target versus the reference.

Microarray technologies produce gene expression patterns that provide dynamic informations about cell functions. These informations can be used to investigate complex interaction within the cell. In this contest, data mining methods can be used to determine co-regulated genes and suggest biomarkers for specific diseases, or to ascertain and summarize the set of genes responding to a certain level of stress in an organism.

A typical question in genomic mining, see e.g. Speed [11], is in fact "which gene is the most similar to which" in terms of gene expression. Another important aspect is the correlation between gene expressions and malignant samples. Gene expression data being typically high-dimensional, they need appropriate statistical features to discern possible patterns and to identify mechanisms that govern the activation of genes in a organism.

Our approach was first to treat the problem of the correlation between gene expression and malignant samples as a predictive problem with a categorical predictor variables - genes - and a response binary variable being the sample's status '1' (malignant) or '0' (normal). Then we have used association rules to analyse genes resulting from these predictive methods. Mining of association rules, in fact, had already been successfully applied on microarray data by using the A priori algorithm. Associations rules can be used to express associations between cell environmental effects and gene expressions, to diagnose a profiled cancer sample, or to analyse drug treatment effects.

In order to find the best predictive models, we have reduced the number of input variables with feature selection. We have compared chi-square selection and a new approach on variable selection, based on marker selection.

The chi-square selection criterion is available for binary targets. This criterion provides a fast preliminary variable assessment and facilitates the rapid development of predictive models with large volumes of data. Variable selection, based on chi-square, is performed using binary variable splits for

maximizing the chi-square values of a contingency table. Each level of the ordinal or nominal variables is decomposed into a binary variable. In high dimensional data sets, identifying irrelevant inputs is more difficult than identifying redundant inputs. A good strategy is to first reduce redundancy and then tackle irrelevancy in a lower dimension space.

Marker selection approach, see e.g. R. Mott [12], is based on the structure of the genes. Consider the general representation of the frequency distribution of a qualitative variable with K levels. Null heterogeneity, holds when all the observations assume the same level. That is if $p_i = 1$ for a certain i, and $p_i = 0$ for the other *k-1* levels. Maximum heterogeneity, holds when the observations are uniformly distributed amongst the k levels, that is $p_i = 1/k$ for all $i = 1, ..., k$. Heterogeneity measures can be extended and applied to gene expressions. As a measure of genes diversity, the entropy (E) can be calculated using:

$$E = -\sum_{k=1}^{m} p_i log p_i,$$

where p_i is the probability of gene i being activated, and K the number of genes. Wanting to obtain a 'normalised' index, which assumes values in the interval [0,1], one can rescale E by its maximum value, obtaining the following relative index of heterogeneity:

$$E' = \frac{E}{log(K)},$$

In order to select the most predictive genes, genes are sequentially subdivided in groups (as in a divisive cluster analysis algorithm). The previous entropy is calculated for each chosen subset. It will continuously increase starting from 0 up to a maximum of E. Grouping and, hence, gene marker selection is stopped when a suitable threshold is reached (e.g. 0.95).

If s is a subset of t we have that $E(s) \prec E(t) \prec E$. The difference *E(t) - E(s)* is a good measure of how nested subsets compare in describing the data. A

sequence of marker subsets $s_1 ... s_k$ generates a monotonic sequence of optimal approximations (as measured by their entropy) to the gene structure of the data. The probability of detecting an association between a marker and a diseased phenotype decreases with the distance between the marker and the actual position of the gene responsible for the phenotype. Thus, one can maximize the probability of detecting disease linkage by choosing markers as closely spaced as possible. This procedure is closely related to principal component analysis and can be used as an alternative method for eliminating redundant dimensions.

This type of variable clustering well finds groups of variables that are as correlated as possible among themselves and as uncorrelated as possible with variables in other clusters. If the second eigenvalue for the cluster is greater than a specified threshold, the cluster is split into two different dimensions. The reassignment of variables to clusters occurs in two phases. The first is a nearest component sorting phase, similar in principle to the nearest centroid sorting algorithms described by Anderberg [15]. In each iteration the cluster components are computed and each variable is assigned to the component with which it has the highest squared correlation. The second phase involves a search algorithm in which each variable in turn is tested to see if assigning it to a different cluster increases the amount of variance explained. If a variable is reassigned during the search phase, the components of the two clusters involved are recomputed before the next variable is tested.

In this paper, for marker selection we use the entropy as a measure of genes diversity that attains a maximum if all genes are present in equal quantities. If only a subset s of genes were typed then some of the original might become indistinguishable and hence will be merged. The sequence of genes subset generates a monotonic sequence of optimal approximations (as measured by their entropy) to the structure of the data. This method has been

implemented in a software code using simple recursive search algorithm, which generates and evaluates all possible genes subsets. In this way we select the variables that are most important to explain the patient disease.

4 Predictive models for genomic data

The interest in exploratory methods for evaluating patterns of association between many variables has been reinforced by problems involving large-scale gene expression studies. Problems of modeling structure under the "'large p, small n" paradigm challenge modern statistical science both conceptually and computationally. Our approach is based in two modelling steps:

- a descriptive step, based on association and link models;
- a predicitve step based on regression and tree models

For the descriptive step we use gene link analysis methodology, see e.g. [14]. The main output of a link analysis is a graph, formed of nodes and links. In the graph, each gene is represented by a node, and links are placed between such nodes. A link is placed between two nodes if the count of the corresponding sequence of order two is non null. The graph thus informs on which nodes are connected, and which are not. Usually the thickness of a link is directly related with the size of the count. The links are directed; for instance, to orientate an edge between two nodes A,B the two counts of A B and B A in the link dataset are compared; the higher determines the orientation. Both orientations will be present in case of substantial parity. The size of the nodes typically depend on a so-called centrality measure. This concepts comes from ideas in social networks. A first order centrality measure basically means that the importance of a node depends on the number of connections it has. On the other hand, a second order centrality

measure means that the importance of a node depends on the number of connections that the nodes connected to it have. In both cases each connection and link can be weighted according to its count in the link dataset. In our case we have chosen to use, to describe the size of a node, an unweighted first order centrality measure. We remark that the position of a node may also depend on the counts. The counts between each pair of pages are put in a proximity matrix. Multidimensional scaling is then used to reproduce such proximities with a bidimensional Euclidean distance, and, correspondingly, derive two X and Y coordinates. The higher the count, the closer the points corresponding to the coordinates in a cartesian graph. The second step of our modelling is predicitve. Predictive modeling tries to find good rules (models) for guessing (predicting) the values of one or more variables in a data set from the values of other variables in the data set. Once a good rule has been found, it can be applied to new data sets (scoring) that may or may not contain the variable(s) being predicted. In a predictive model, one of the variables is expressed as a function of the others. This permits the value of the response variable to be predicted from given values of the others. In our example we have a target variable Y that is the type of tissue and a set of explanatory variable (activations of genes), $X_1, ..., X_p$.

The record for the *i-th* past tissue can be conveniently represented as *(X(i),Y(i))*. Here *y(i)* is the outcome (good or bad) of the *i-th* tissue, and *x(i)* is the vector $x = (x_1(i), ..., x_p(i))$ of genes. A useful predicitve model for binary target is logistic regression.

Let Y_i, *i=1,2,...,n* be the observed values of a binary response variable which can take only the value 1 or 0(tissue diseased or not diseased). A logistic regression model is defined in terms of fitted values to be interpreted as probabilities that the event occurs in different subpopulations:

$$\pi_i = P(Y_i = 1), i = 1, ..., n,$$

More precisely, a logistic regression model specifies that an appropriate function of the fitted probability of the event is a linear function of the observed values of the available explanatory variables. Here is an example:

$$log\frac{\pi_i}{1-\pi_i} = a + b_1 x_{i1} + ... + b_k x_{ik},$$

The left-hand side defines the logit function of the fitted probability as the logarithm of the odds for the event, namely the natural logarithm of the ratio between the probability of diseased tissue and the probability of not diseased tissue. For each patient the logistic regression gives a probability of diseased tissue as a function of the genes.

A second class of predictive models that can be usefully employed is tree models. While logistic regression methods produce a score and then possibly a classification according to a discriminant rule, tree models begin by producing a classification of observations into group. Tree models can be defined as a recursive procedure, through which a set of n statistical units is progressively divided in groups, according to a divisive rule which aims to maximize a homogeneity or purity measure of the response variable in each of the obtained groups. At each step of the procedure, a divisive rule is specified by: the choice of an explanatory variable to split; the choice of a splitting rule for such variable, which establishes how to partition the observations. The main result of a tree model is a final partition of the observations: to achieve this it is necessary to specify stopping criteria for the divisive process. Indeed different criteria give rise to different tree model algorithms that should be compared in terms of performance (see e.g. [14]).We first split the space into two region, and model the response by the mean of Y (type of tissue) in each region. We choose the variable (gene) and split-point to achieve the best fit. Then one or both of these

regions are split into two more regions, and this process is continued, until some stopping rule is applied. In principle this splitting procedure can be continued until each leaf node contains a single training data point. A common strategy is to build a large tree and then to prune it back. The score function used to measure the quality of different tree structures is a general misclassification loss function (MLF), defined as:

$$MLF = \sum_{i=1}^{n} C(Y(i), \hat{Y}(i)),$$

where $C(Y(i), \hat{Y}(i))$ is the loss incurred (positive) when the class label for the *i-th* data vector, *y(i)*, is predicted by the tree to be $\hat{Y}(i)$. In general, C is specified by an $m \times m$ matrix, where m is the number of classes. In our case classification tree use a cross-validation techniques to estimate the misclassification loss function. Basically, this implies to partition the data into a training and a validation data set, on which to estimate the misclassification rate.

In our approach different logistic regression and tree models will be compared in terms of performance measures. We shall choose measures particularly suited for feature selection. More precisely, the total variable importance of an input x (gene) is equal to the square root of the sum over nodes of a measure of agreement multiplied by the reduction in impurity. The measure of agreement is equal to 1 if x is used to split the node; or the agreement measure of a surrogate, if x is a surrogate; or 0 if x is neither primary or surrogate. For an interval target, the impurity is equal to the sum of the squared errors. For a categorical target, the impurity is equal to the Gini Index. In the next sction we show the application of our methods to the data.

5 Application to the available data

The goal of the application is to create a valid predictive model to diagnose malignant tissues, based on the observation of gene expressions. The available data is a sub-set taken from a large database (GeneExpress) and analysed by S. Young et al [13]. The resulting data set is composed by 112.896 gene expressions ordered into 224 columns and 504 rows. Columns represent a set of 224 genes, rows correspond to 504 samples, covering 8 tissue types - adipose tissue, breast, colon, kidney, liver, lung, ovary and prostate - both normal (249 samples) and malignant (255 samples). Measured values of the gene expression data have been put in bins and markers as being '1' (highly expressed) or '0' (not-highly expressed). We have also applied label '1' to malignant tissues and '0' to normal tissues.
In order to analyse the data we first discover associations between genes expression. Associations rules can be used to express associations between cell environmental effects and gene expressions, to diagnose a profiled cancer sample, or to analyse drug treatment effects. In Figure 1 we show the relationships in diseased tissue that we have obtained, in the form of a link graph.

Fiugre 1 about here

The associations are strong when the line is read (high probability that in a tissue we find a couple of genes), and fair when the line is green. In particular in diseased tissue we observe high association between gene 216 and gene 217, gene 89 and gene 138, gene 138 and gene 116. Link analysis for not diseased data turned out to be different in terms of results from not diseased tissues, as it is possible to see comparing Figure 1 and Figure 2.

Fiugre 2 about here

We now turn to feature selection. In order to find the best predictive models, we have reduced the number of inputs by eliminating input variables that are not related to the target.

Table 1 about here

In Table 1 we have compared chi-square selection and our proposed approach on variable selection, based on marker selection. In Table 1 it is possible to see that the two approaches have a good overlap in selected genes:(gene 10, gene15, gene28, gene110, gene113, gene116, gene 121, gene138, gene217) are in common.

In order to better compare the two selections, we have run two classification tree models with the same settings and compared the resulting disriminant variables. Table 2 show the results for marker selection approach and Table 3 show the results for chi-square selection approach. It is possible to see that marker selection approach is more parsimonious.

Table 2 about here

Table 3 about here

Table 3 show the genes that have more importance to forecast diseased tissue following chi-square selection approach. After the feature selection process, our focus is to build a valid model to predict diseased tissues. In order to reach this objective we compare logistic regression and decision tree in terms of the confusion matrix, different performance indexes (error rate, accuracy, sensitivity, specificity) and some evaluation graphs (lift chart and ROC Curve).

The confusion matrix is used as an indication of the properties of a classification (discriminator) rule. It contains the number of elements that have been correctly or incorrectly classified for each class. On its main diagonal we can see the number of observations that have been correctly classified

for each class while the off-diagonal elements indicate the number of observations that have been incorrectly classified. We classify the observations of a validation dataset in four possible categories: the observations predicted as events and effectively such (in Table 4 with absolute frequency equal to 151); the observations predicted as events and effectively non events (with frequency equal to 27); the observations predicted as non events and effectively events (with frequency equal to 28); the observations predicted as non events and effectively such (with frequency equal to 151). The errors in the previous confusion matrix (Table 4) are 27 and 28 cases. Table 4 compares the performance of logistic regression and tree models.

Table 4 about here

The observation predicted as non events and effectively events are 28 and the observations predicted as events and effectively non events are 22.
The ROC (Receiver Operating Characteristic) curve is a graph that also measures predictive accuracy of a model, see e.g. [14]. It is based on the confusion matrix. Given an observed table, and a cut-off point, the ROC curve is calculated on the basis of the resulting joint frequencies of predicted and observed events (successes) and non events (failures). More precisely, it is based on the following conditional probabilities:

- sensitivity: proportion of events, predicted as such;
- specificity: proportion of non events, predicted as such;

The ROC curve is obtained representing, for any fixed cut-off value, a point in the Cartesian plane having as x-value the false positive value (1-specificity) and as y-value the sensitivity value. Each point in the curve corresponds therefore to a particular cut-off . The ROC curve can thus also be used to select a cut-off point, trading-off sensitivity and specificity. In terms of model comparison, the best curve is the one that is leftmost, the

ideal one coinciding with the y-axis.
Figure 3 compares ROC Curve for logistic regression and tree models and is possible to derive that when we apply chi-square selection, the power in prediction turns out to be similar between logistic regression and decision tree.

Figure 3 about here

If we use marker selection approach, the best model is logistic regression, as can be seen in Figure 4.

Figure 4 about here

We remark also that marker selection approach is more parsimonious in terms of gene selected.
Finally, we compare the best model with a model derived from boosting method. Boosting method is based on re-weighted re-sampling developed from a weak learning algorithm, with the weights in the re-sampling are increased for those observations most often misclassified in the previous models. Table 5 shows the comparison of the misclassification errors; we can see that in our application it is possible to improve the previous results in term of misclassification rate (logistic regression and tree-based model) using a boosting method .

Table 5 about here

References

1 Blum, A.L., and Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97:245-271.

2 www.ics.uci.edu/ mlearn/MLRepository.html

3 Yang, J., and Honavar, V. (1997). Feature subset selection using a genetic algorithm. Proceedings of the Genetic Programming Conference, pages 380-385. Stanford, CA.

4 Aha, D.W., and Bankert, R.L. (1996). A comparative evaluation of sequential feature selection algorithms. In D. Fisher and J.-H. Lenx (Eds.), Artificial Intelligence and statistics V. New York: Springer-Verlag.

5 Glover, F., Laguna, M., and Marti, R. Scatter search. Theory and Applications of Evolutionary Computation: Recent Trends. A. Ghosh and S. Tsutsui (Eds.), Springer-Verlag.

6 Kira, K., and Rendell, L. (1992). A practical approach to feature selection. Proceedings of the Ninth International Conference on Machine Learning (pp.249-256). Aberdeen, Scotland: Morgan Kaufmann.

7 Xing, E., Jordan, M., and Karp, R. (2001). Feature selection for high-dimensional genomic microarray data. International Conference on Machine Learning.

8 Slonim, D.K., Tamayo, P., Mesirov, J., Golub, T., and Lander, E. (2000). Class prediction and discovery using gene expression data. Proceedings of the 4th Annual International Conference on Computational Molecular Biology, pp. 263-272. Tokyo, Japan: Universal Academy Press.

9 Califano, A., Stolovitzky, G., and Tu, Y. (2000). Analysis of gene expression microarrays for phenotype classification. Proceedings of the Annual Intelligent Systems in Molecular Biology, 8:75-85.

10 Breiman, L. (1999). Random forests. Machine Learning, 45:5-32. Kluwer Academic Publishers, Boston.

11 Speed T. (2003), Statistical analysis of gene expression microarray data, New York, Chapman e Hall

12 M Mott R. (2003), Marker selection, University of Oxford.

13 L. Liu, D. M. Hawkins, S. Ghosh, S. S. Young (2000), Robust Singular Value Decomposition Analysis of Microarray Data.

14 P.Giudici (2003) Applied data mining, Wiley

15 Anderberg, M.R. (1973) Cluster analysis for applications, New York Academic Press

Chi-square	Marker selection
GENE 10	GENE 8
GENE 15	GENE 10
GENE 28	GENE 15
GENE 39	GENE 28
GENE 92	GENE 64
GENE 97	GENE 72
GENE 98	GENE 75
GENE 110	GENE 76
GENE 113	GENE 80
GENE 116	GENE 102
GENE 119	GENE 110
GENE 121	GENE 113
GENE 138	GENE 116
GENE 164	GENE 121
GENE 173	GENE 126
GENE 175	GENE 134
GENE 195	GENE 138
GENE 217	GENE 217

Table 1: A comparison between feature selection

Genes Marker selection
GENE 15
GENE 28
GENE 217
GENE138
GENE126

Table 2: Feature selection with marker selection tree

Genes Chi-square selection
GENE 15
GENE 217
GENE28
GENE138
GENE119
GENE121
GENE110
GENE97

Table 3: Feature selection with chi-square selection tree

Frequency	Pred chisq=0	Pred chisq=1	Pred marker=0	Pred marker=1
Obs chisq=0	147	27	-	-
Obs chisq=1	28	151	-	-
Obs marker=0	-	-	148	22
Obs marker=1	-	-	28	154

Table 4: Confusion Matrix of marker selection and chi-square selection

Model	Training: Misclassification Rate	Validation: Misclassification Rate
Boosting	0	0.165562
Logistic Regression	0.172804	0.225165

Table 5: Goodness of fit

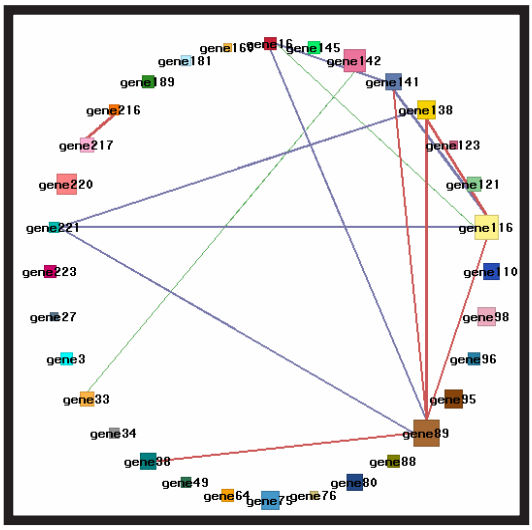


Figure 1: Link analysis for diseased

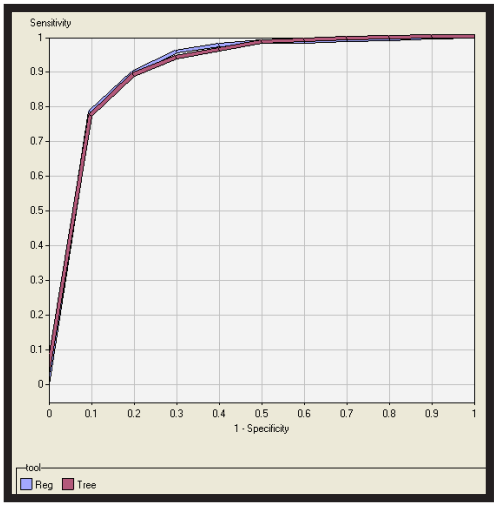


Figure 2: ROC Curve chi-squre selection

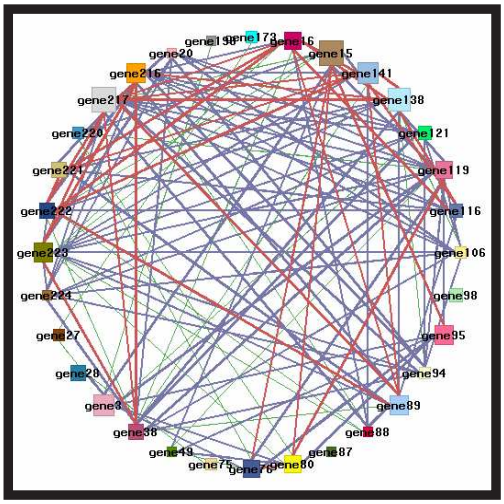


Figure 3: Link analysis for not diseased

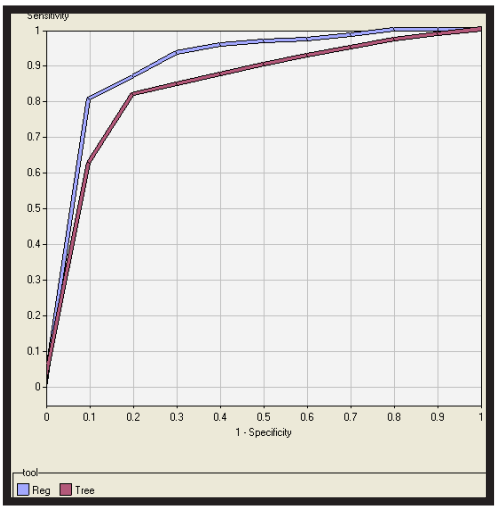


Figure 4: ROC Curve marker selection