DIMACS Technical Report 2005-35
February 2006

Machine Learning Methods in the Analysis of Lung Cancer Survival Data

by

Dmitriy Fradkin
Dept. of Computer Science
Rutgers University
New Brunswick, New Jersey 08854

Dona Schneider
Edward J. Bloustein School of Planning and Public Policy
Rutgers University
New Brunswick, New Jersey 08854

Ilya Muchnik [1]
DIMACS
Rutgers University
New Brunswick, New Jersey 08854

[1]Ilya Muchnik was supported by the NSF grant CCR 0325398

ABSTRACT

Support Vector Machines (SVM) and penalized logistic regression are well known to the machine learning community but are yet to be actively used in an epidemiological application. We apply them to the task of constructing a predictive model for the survival of patients diagnosed with lung cancer and analyzing the importance of features based on model parameters.

The methods produce distinct and complementary models, making it advantageous to consider both whenever possible to gain different perspectives into large datasets. After applying the methods to Surveillance, Epidemiology and End Results (SEER) data, we also compute several measures of feature importance in the final models, showing these measures to be strongly correlated.

1 Introduction

A number of modern machine learning methods have the potential to be used in analysis and modeling of medical and epidemiological data, yet their use, with the exception of logistic regression, is not widespread.

In this work we discuss two modern machine learning methods, Support Vector Machines (SVM) and penalized logistic regression, that are well known to the machine learning community but are yet to be actively used in an epidemiological application. These methods are applied to the task of constructing a predictive model for survival of patients diagnosed with lung cancer and analyzing the importance of features of the models based on the model parameters.

Our work involves preparing the data and constructing and analyzing a high-quality model for lung cancer data obtained from the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute [1]. Section 3 discusses the SEER data preprocessing, record and feature selection. Section 4 describes the experiments for constructing a single classification model for the selected data. Section 5 is concerned with the analysis of feature significance in the final model for the purpose of understanding their role.

2 Related Work

Some machine learning methods have previously been applied to medical data.

Street [19] built neural networks for predicting recurrence or survival time. Friedman and Meulman [6] used additive regression trees to distinguish between different types of cervical dysplasia based on the features of cells extracted with biopsy. They also suggested methods for determining the importance of the individual features in making a diagnosis.

Inductive Logic Programming (ILP) [17] was used by Han et. al. [10] for predicting sarcoma survival. ILP extracts first-order logical rules that cover the cases in the training data and generalize to new cases. Logical Analysis of Data (LAD) [2], another approach for finding logical rules discriminating between the classes, has also been applied to medical diagnosis.

Delen et. al. [4] empirically compared neural networks, decision trees and logistic regression for the task of predicting 60 months breast cancer survival. The conclusion was that decision trees were best, followed by neural networks. Logistic regression was the worst of the three. The evaluation was performed using 10-fold cross-validation. The data was extracted from SEER. The original data consisted of 433,272 records and 72 SEER fields. After data cleaning and preparation, there were 202,932 cases with 16 predictor variables (11 categorical and 5 numeric), including race, age, marital status and some medical features. Additionally, sensitivity analysis was conducted on neural networks to determine importance of individual variables.

3 SEER Data Preparation

3.1 SEER Data Format

The Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute is an authoritative source of information about cancer incidence and survival in the United States.

In our work we used SEER data for the years 1973-2001, released in April 2004. The data were collected from 12 population-based cancer registries. All the following text, in particular when describing the fields and data format, refers to that version.

Data are stored in SEER in rows of fixed width (166 characters), containing 77 fields of fixed length. Each patient is uniquely identified by the combination of "SEER registry" and "case number" fields. A patient may have several different entries. These are distinguished by "record numbers."

Information for each patient can be partitioned into two sets: demographic and medical. Demographic information includes fields such as age, sex, race/ethnicity, place of birth, etc. Medical information includes location of the disease, its type (morphology, histology) and extent, and also the types of treatment (radiation, surgery) and cause of death (COD) where applicable.

The SEER database has evolved over time and therefore certain kinds of information available in recent years are not present in older records. Year 1988 seems particularly significant, with the introduction of several new fields (such as extent of the disease) and of detailed schemes for several other fields.

3.2 Feature Conversion and Preparation

The fields in a SEER record can be grouped into 3 types: categorical, ordinal and numeric.

A categorical field, such as race, with m possible values can be represented by m binary variables (or fewer, depending on which categories are of interest) where x_i has value 1 only if the i-th category occurred in the field.

Many of the fields in a SEER record are ordinal, i.e. the values in these fields can be ordered but they don't have a defined distance function. One example is field 18, "Grade". Possible values are: Grade I, Grade II, Grade III, Grade IV and some others. While there is a clear ordering to the possible values, it is not known how much worse Grade II is than Grade I, and how this relates to the difference between Grade III and Grade II.

Consider an ordinal variable v taking values $\{1, \ldots, m\}$. It can be represented by an m-tuple of binary variables v_i,$i = 1, \ldots, m$ in the following way:

$$v_i = 1 \iff v \geq i \tag{1}$$

Such representations allow a model, such as logistic regression model, to automatically determine the difference between the levels by assigning different coefficients to different variables v_i.

Certain fields have integer values, such as age (in years) or size of tumor (in millimeters). Since the exact values are not likely to be significant (i.e. it probably does not matter whether a person is 31 or 35), these can be considered ordinal variables with values in m specific intervals, and can then be converted into m binary variables using equation (1).

The process of constructing variables requires consultation with an epidemiologist who can point out features of interest based on medical significance, reporting practices and coding methods. For example, similar information on stage of the disease can be reported by different sources using somewhat different coding. Some of these sources may be more reliable or provide more detailed information. An outsider would not normally be aware of such distinctions.

The variable construction processes also may require multiple iterations. We give one example of a situation where necessity of recoding became clear only after the completion of initial experiments. Some SEER fields, such as extension, appear to be coded as ordinal variables (i.e. the larger codes appear to correspond to more serious conditions). However we found that this is not the case, and that the effect is not monotonic. For this reason we coded extension SEER field using 6 separate binary variables, rather than the 6-bit scale as described above.

In some cases, a field may potentially have many values (i.e. there are many codes specified) but in practice only a few of them occur in the data. In such cases only the values that occur are of interest. This can only be determined by looking at the distribution of the values for a particular field.

The long tables in Section B describe which specific features were used in this analysis and how they were coded. Our representation consists of 98 binary variables, either demographic or medical. Demographic features are related to the individual's social and demographic description and are independent of the individual's medical condition. The other set of features describes the patients' medical status.

All the features are binary and are described below as having values 1/0. However, they can be coded as 1/0 or as 1/-1. We experimentally analyze the difference between the two coding schemes in Section 4.2.

3.3 Data Processing

From the files "RESPIR.TXT" in directories "yr1973_2001.seer9" and "yr1992_2001.sj_la_ak" we extracted all entries with a diagnosis of cancer of lung or bronchus (ICD Codes C34.0-C34.9). There were 436,022 such records.

These entries contain multiple records (at most 4) for a small group of people (6374). In order to keep the analysis simple, we removed all records pertaining to such people, leaving us with 423,078 records.

We chose to use only the data from years 1988 and later. There were several reasons for this. One reason is that SEER contains more information on cases starting in 1988, and there is less missing data. The other reason is that the median survival time (8 month) has not changed since 1985 until now, while the third quartile survival increased slightly

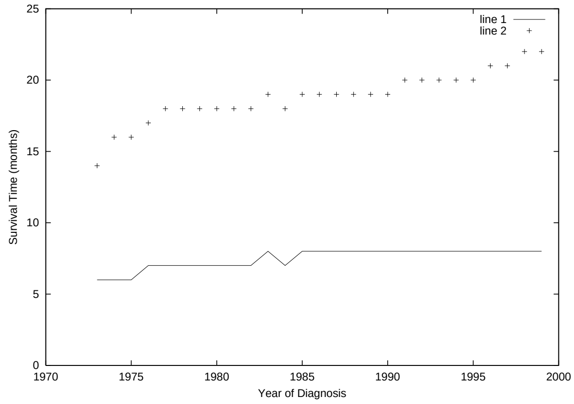

Figure 1: Plot of the median (line) and third quartile, 75%, (points) survival time as function of the date of diagnosis.

from 19 to 22 months. (The plots of median and 75% survival times are given in Figure 1.) Thus there were no drastic changes in the survival time caused by advances in diagnosis or treatment in the longitudinal data.

As a cut-off for short-term survival (class 1) and long-term survival (class 0) we chose the median survival time, 8 month.

The label was determined as follows:

1. If survival time is unknown, or if a person was diagnosed in 2001 and is not dead and survival is less than 8 month, he has to be discarded (since we cannot assign him to either class).

2. If a person is dead, and cause of death was cancer of lung or bronchus, and survival time is less than 8 months, we assign that person to class 1.

3. If a person survived longer than 8 months, then he is assigned to class 0.

4. Otherwise, the person has to be removed since a label cannot be determined (for example if a person died within 8 months of diagnosis but not from lung cancer). Since we are not gathering statistics, but attempting to construct a case-based model, we only want to include cases where we are certain of the relationship between the label and the features.

The selected data from years 1988+ were split approximately evenly into a training set (up to and including 1995) and the test set (1996-2001). The training set consists of 122,613 cases, and the test set consists of 100,292 cases.

3.4 Missing Value Analysis

Before constructing the classifier, we wanted to check that the selected features and cases were appropriate. Therefore, we conducted the following data cleaning:

Table 1: The matrix below represents the frequency of values of some feature on the training set. The missing values will be filled using the most frequent value in the most frequently missing class. For example, if $x > y$, then missing values will be set to 1 if $c > a$ and to 0 otherwise.

	Class 0	Class 1
Value 0:	a	b
Value 1:	c	d
Missing values:	x	y

1. If a feature is missing a value in more than 25% of cases, it is removed.
2. If a feature has the same value in 95% or more of cases where the value is not missing, it is removed (constant feature).
3. The objects that are missing more than 25% of the feature values on the remaining features are removed as well.

The discussion of the effects of changing these thresholds is given below.

All the statistics on the features are computed exclusively on the training set. The remaining missing values are filled in with the most frequent value among the class with the most frequently missing values on the training set. Table 1 explains this in detail.

After processing, 45 features are left. The training set retains 120,318 cases, while the test set now consists of 97,240 cases (in other words 2,295 training examples and 3,052 test examples were removed).

All the experiments discussed below are performed starting with these data. We would like to comment on the stability of the above procedure.

Varying the cut-off for missing values in the interval $[15\%, 38\%]$, while keeping the other parameters fixed, did not change the number of variables selected. In other words, there is a large gap in the frequency of missing values for the variables (see Table 2). Some variables have a lot of missing values, while others have a lot fewer missing values.

Changing the parameter controlling constant features has a greater effect. Setting it to 99%, rather than 95%, leads to 64 variables remaining in the model. Additionally, somewhat fewer examples are skipped (1,770 in training, and 2,080 in test). The sensitivity and specificity of a model built on these data is slightly worse than of that built on the 45 variables. While this result is achieved on a slightly larger data, it is unlikely that these additional variables are significant.

Increasing the threshold for the acceptable number of missing features in the example from 25% to 30% leads only to a minor increase in the number of cases: 1751 train cases and 2055 test cases are skipped, as opposed to 2295 and 3052 respectively for threshold of 25%. This change does not significantly affect the model or the results.

Table 2: Frequency of missing features with a particular missing rate (out of the initial 98). One can easily see the large gap $[15\%, 38\%]$.

Number of Features	Missing Rate (% of all cases)
9	46.34
4	41.18
1	38.54
1	38.34
6	14.70
6	13.77
9	10.63
10	10.00
2	5.23
1	4.92
1	1.98
4	0.41
2	0.23
1	0.18
2	0.08

We believe that the observations and experiments described in this subsection demonstrate that the parameters used in the data cleaning stage are reasonable. While the experimentation and the methodology could have been more extensive, this would take us away from the main subject of this work.

4 Constructing A Global Model

Having selected the features we now turn to the task of building the model. There are several aspects to this task. One is the choice of representation - as we shall see below even in the simple case of a fixed set of binary features there are different ways of representing the features numerically. Second is the possibility of expanding the set of features with functions of several features, in order to better represent possible interactions. Additionally, closely interwoven with these choices is the choice of the method for constructing the model. In general, different methods have different learning biases and may perform better with different representations. Thus we have to consider the choices of representation and of the method together. Below we will consider two methods for building models: linear Support Vector Machine (SVM) [21], implemented in LIBSVM software [3] and penalized logistic regression, implemented in BBR software [8]. These methods are discussed in the Appendix. Here we only mention that both of these methods construct a linear discriminant function that assigns to each observation a score indicative of its estimated class membership. In other words both methods construct a function $f(x|w,b)$:

$$f(x|w,b) = \sum_{i=1}^{d} w_i x_i + b \tag{2}$$

In particular, the SVM score is the distance to the separating hyperplane between the two classes, the logistic regression score (after some transformations) is the estimated probability of a point belonging to the positive class (given the feature values for the point). Both of these methods use some threshold to make an assignment based on the score.

Here we shall evaluate several different representations with BBR software (penalized logistic regression model). We will also experiment with the penalization parameter of BBR. We then compare BBR results with those of linear SVM. In other words, we shall consider a single model and a number of alternatives. The purpose of this is to make sure that we are not missing a simple way of producing a better model.

4.1 The Baseline Result

As the baseline result we consider the performance of BBR with Laplace prior (a setting favoring feature selection), using the 45 binary $\{0, 1\}$ variables, as mentioned above. The hyperparameter of BBR (variance of the prior) was selected by internal 10-fold cross-validation (the choices were: 0.0000001,0.000001,0.00001,0.0001,0.001,0.01,0.1,1). The prior variance selected was 1. Despite the use of the sparseness-inducing prior no features were removed. The threshold tuning parameter of BBR was set to minimize the sum of errors on the training set (this is an option available in BBR software). The training (including cross-validation) takes approximately 30 minutes.

The baseline approach gives sensitivity (percentage of correct identification of class 1) of 72.10 and specificity (percentage correct identification of class 0) of 72.50 on the test set. Estimates obtained with 2 runs of 5-fold cross-validation on the training set are 71.50 and 72.87 respectively. This demonstrates that the training set is representative of the test set and that the model built on it should generalize.

In order to demonstrate that this is indeed a reasonable result we compare this result to a number of alternatives. We will make use of the following formula for relative importance of a variable in a model. Suppose that the model is given by a vector of weights w, with weight w_j corresponding to the feature j. Then the relative importance of the variable j is given by:

$$r_j = \frac{|w_j|}{\sum_h |w_h|}. \tag{3}$$

Intuitively, this indicates how much effect compared to the other variables a particular variable has on prediction.

Table 3 shows the top 12 variables according to the measure (3). Together they have 50% of the total weight.

Table 3: Features (from BBR), sorted by importance that add to 50% of the total weight. Histology code "xyz*" means that the first three digits of the code are xyz.

ID j	Relative importance r_j	Coefficient w_j	Meaning
73	0.080	-1.197	Surgery was performed
76	0.063	-0.950	Radiation and surgery
66	0.048	0.717	Extension code 80-85
83	0.040	-0.599	Histology code 804*
64	0.040	0.598	Extension code 71-76
75	0.038	-0.573	Radiation
94	0.036	0.542	Stage code 10 or higher
31	0.034	0.512	Born in East South Central region
97	0.034	0.511	Stage code 32 or higher
74	0.034	-0.501	Surgery recommended
33	0.033	0.501	Born in Mountain region
32	0.026	0.391	Born in South West Cental region

It is interesting that the features with large negative weights (73 - surgery was performed, 76 - radiation and surgery were performed) are associated with treatment procedures. This can be related to the effectiveness of these procedures, or to the fact that these procedures are used when they will extend the patient's life and thus are "proxy" variables for other factors. The features with large positive weights (64, 66, 94, 97) indicate the extension into different organs and the stage of the disease. Histology code "804*" Carcinoma, Not Otherwise Specified (NOS) - variable 83 - may be a relatively less dangerous or more controllable type of cancer. While most of the top features are medical, several features relating to the place of birth (31, 32, 33) also have large weights. This could be a reflection of some cultural factors influencing behavior of the patients.

We experimented with decreasing prior variance in BBR to perform variable selection. Setting prior variance to 0.0001 resulted in 24 variables with non-zero coefficients. The most important are given in Table 4. All of these were present with large weights in Table 3. Demographic variables here are not among the most important ones. The performance of this model is 74.62 and 70.57 - comparable to the baseline model.

Setting prior variance to 10^{-6} resulted in 11 variables with non-zero coefficients. All of them are given in Table 5. The main variables are still the same - related to the treatment procedures (73,74) and to the stage and extension of the cancer (61,66,96). The only demographic variable present (with a small positive weight) is 24 - age 75 or greater - implying that older people are less likely to survive 8 months after the diagnosis. The performance of this model is 74.84 and 68.26 - somewhat worse than that of the baseline or the model obtained with variance 0.0001.

Table 4: Features (from BBR with $\sigma^2 = 0.0001$), sorted by importance, that add to 50% of the total weight. Histology code "xyz*" means that the first three digits of the code are xyz.

ID j	Relative importance r_j	Coefficient w_j	Meaning
73	0.150	-1.164	Surgery was performed
76	0.090	-0.698	Radiation and surgery
66	0.083	0.643	Extension code 80-85
83	0.078	-0.600	Histology code 804*
75	0.065	-0.506	Radiation
74	0.062	-0.479	Surgery recommended

Table 5: All features retained by BBR with $\sigma^2 = 1 * 10^{-6}$, sorted by importance.Histology code "xyz*" means that the first three digits of the code are xyz.

ID j	Relative importance r_j	Coefficient w_j	Meaning
73	0.249	-0.858	Surgery was performed
61	0.182	-0.629	Extension code 10-30
66	0.135	0.467	Extension code 80-85
96	0.099	0.342	Stage code 31 or higher
74	0.093	-0.322	Surgery recommended
75	0.072	-0.247	Radiation
83	0.061	-0.210	Histology code 804*
97	0.044	0.152	Stage code 32 or higher
24	0.035	0.119	Age 75 or greater
70	0.020	-0.068	Site specific surgery performed
76	0.010	-0.036	Radiation and surgery

Table 6: Features (from BBR) on $\{-1, 1\}$ representation that add to 50% of the total weight, sorted by importance. Histology code "xyz*" means that the first three digits of the code are xyz.

ID j	Relative importance r_j	Coefficient w_j	Meaning
73	0.083	-0.641	Surgery was performed
66	0.054	0.417	Extension code 80-85
76	0.051	-0.391	Radiation and surgery
83	0.042	-0.322	Histology code 804*
64	0.040	0.306	Extension code 71-76
31	0.037	0.282	Born in East South Central region
94	0.036	0.282	Stage code 10 or higher
33	0.036	0.281	Born in Mountain region
75	0.036	-0.280	Radiation
32	0.029	0.223	Born in South West Cental region
30	0.028	0.218	Born in New England
28	0.028	0.217	Born in East North Central
26	0.028	0.214	Born in South Atlantic

4.2 Effect of Changing Coding

An alternative representation of the data is as $\{-1, 1\}$, rather than as $\{0, 1\}$. In this representation taking the inner product between two vectors is equivalent to computing the Euclidean distance between them and to the total number of identical entries, while in $\{0, 1\}$ representation it is the number of ones they have in common. However the change of representation did not drastically alter the quality of the result or the variable significance.

The sensitivity and specificity were 71.26 and 73.09 respectively. The features that add to 50% of importance are given in Table 6. Comparing Table 3 and Table 5 one can see that most of the variables are present in both tables, with the same signs and similar relative importance. The variables that disappeared off this list are 97 and 74. New variables are 30, 28 and 26, all of which are place of birth indicators.

4.3 Capturing Variable Interactions

In an attempt to capture feature interactions, we computed frequency information for all triples of features (out of the 45 remaining ones). An example of a triple is "feature 23 = 1 and feature 62 = 0 and feature 95 = 1", which can be interpreted as "a man above 65 years of age, with tumor code 10-30 (i.e. localized) and stage 20 or higher." Such triples can provide a compact description of particular groups of patients, and in the event that some triples are important, they are easy to interpret.

The triples were selected based on the two quantities: support (the total number of times that a particular triple occurs) and "strength" (the ratio of frequencies of the triple between the two classes). In other words, if a triple t occurs in $n^-(t)$ cases of the negative class and in $n^+(t)$ of the positive class, then support of t is:

$$u(t) = n^-(t) + n^+(t), \tag{4}$$

and strength of t is:

$$s(t) = max(\frac{n^-(t)}{n^+(t)}, \frac{n^+(t)}{n^-(t)}). \tag{5}$$

There are 112,179 triples that occur at least once in the training set. We aimed to select a small number of important interactions. We discovered that the "strength" of triples from the negative class was greater. Setting support at 25,000 and strength at 10, we obtained 4 triples. For the positive class (class 1), setting strength to 3 and support to 12,000 produced 5 triples.

It is interesting to note the support of these triples (combined) on the training set. The 4 triples covered 23,056 cases of the negative class and only 2,227 cases of the positive class. On the other hand, the 5 triples covered 23,807 of the negative class but 9,198 of the positive class. Thus these triples are useful indicators of the class, despite low combined support.

By combining these 9 triples with the 45 original variables we can expand our space to 54 variables. Building a classifier with BBR in this space resulted in 72.42 sensitivity and 72.43 specificity. This is essentially the same result as without the triples. Furthermore, the triples did not have high weights in the classifier. Decreasing the prior variance led to few triples remaining in the model, all with low coefficients. Thus we have to conclude that introducing interactions of up to 3rd order did not lead to improved classification or more interesting models.

4.4 Comparison with SVM

For SVM, we tried linear kernel with hyperparameter set to 1. SVM turned out to be very slow, taking hours to train (possibly because the particular software we used did not provide special technics for linear SVM). Therefore, we only evaluated SVM performance in several cases.

The sensitivity and specificity obtained were 72.45 and 68.69 respectively. This is slightly worse than the baseline BBR result. However the variable coefficients are rather different from BBR model. Two variables ("Extension code 71-76" and "Stage code 31 or higher") accounted for more than 99% of relative importance. They are given in Table 7. These are the only variables with high weights - all other weights are insignificant in comparison. If we ignore these two weights, the remaining variables have more similar coefficients. The top several are displayed in Table 8. All of these are features that have high weights in BBR models. However, in BBR models their weights are comparable to the largest coefficients.

Similar behavior - a few variables have 99% of the relative weight - is observed when SVM is applied to the data projected onto subsets of variables, or when the coding is changed from $\{0, 1\}$ to $\{-1, +1\}$. The results with $\{-1, +1\}$ encoding are comparable to the baseline, while with decreased dimensionality the quality of the models decreases.

Table 7: Top 5 Features from SVM, sorted by increasing relative importance.

ID j	Relative importance r_j	Coefficient w_j	Meaning
64	0.499	1.999	Extension code 71-76
97	0.499	1.998	Stage code 32 or higher

Table 8: Relative importance of top 5 features from SVM after removing features 64 and 97.

ID j	Relative importance r_j	Coefficient w_j	Meaning
66	0.127	0.0009	Extension code 80-85
63	0.058	-0.0004	Extension code 60-70
74	0.056	-0.0004	Surgery recommended
96	0.049	0.0004	Stage code 31 or higher
61	0.043	-0.0003	Extension code 10-30

The results of this section suggest that SVM may produce a very different (from logistic regression) model for the data, and thus may provide an important perspective on the phenomenon in question. While we will not focus on SVM in the remainder of this work, whenever possible it is interesting to use this method in addition to logistic regression.

4.5 Summary of Classifier Effectiveness Comparisons

Here we summarize the results mentioned in the preceeding Sections (Table 9). These results suggest that the binary $\{0, 1\}$ representation using 45 input variables performs as well or better than alternatives examined. Inclusion of some 3rd degree interactions did not improve the performance or reveal interesting relations. SVM performed somewhat worse than BBR and required much more computational effort, making it impractical for studies of very large datasets.

Therefore, in the analysis of feature significance below we will focus on the BBR model built using all 45 variables.

5 Analysis of Variable Importance in the Global Classifier Model

Once a model is constructed, it has to be analyzed. One important criterion is the predictive accuracy of the model. A model with poor predictive performance is of no practical interest. Even a model that perfectly describes the training data should be discarded if it has poor performance on the test set or in cross-validation. However, model analysis extends further than just the accuracy of a model. The next step, one of great importance for epidemiology,

Table 9: Predictive performance (sensitivity, specificity) using cross-validation and the test set.

Representation	BBR test	SVM test
45 variables (BBR prior variance 1)	(72.10, 72.50)	(72.45, 68.69)
24 variables (BBR prior variance 10^-4)	(74.62, 70.57)	(72.45, 68.69)
11 variables (BBR prior variance 10^-6)	(74.84, 68.26)	(75.44, 63.27)
45 variable as $\{-1, +1\}$	(71.26, 73.09)	(72.45, 68.69)
45 variables with triples	(72.42, 72.43)	—

is to analyze the significance of individual features in terms of their effect on survival.[1] Linear models (such as the ones produced by SVM and logistic regression) are traditionally popular precisely because their parameters can be interpreted as indicators of feature significance. There are however different ways to interpret them. We consider several below.

The problem of determining significance of variables is well known in applied statistics and machine learning, and many different measures have been developed for this task. One approach, relative importance in the model, based on the size of the coefficient, was mentioned above.

Friedman and Popescu [7] suggested the following measure of variable relevance:

$$I_j = |w_j|\sigma_j. \tag{6}$$

One way to interpret this measure is that between the two features with the same weight, the one with larger standard deviation is more useful since it would tend to be more informative. Another interpretation comes from regression analogy - in a linear regression model I_j would correspond to the coefficient of variable j if it was normalized as a preprocessing step. In other words, if all variables are scaled to have the same variance, then their coefficients would be ordered in the same way as values I_j for respective j.

A different approach for estimating feature significance based on the effect that a feature has on prediction was suggested in [13]. Suppose we apply our model to an independent test set. Then, we can estimate the accuracy of our model conditioned on a particular value of the feature. Let $Q_1^+(j) = \{x_i | x_{ij} = 1 \ \wedge \ y_i = 1\}$, and $Q_0^+(j) = \{x_i | x_{ij} = 0 \ \wedge \ y_i = 1\}$ - sets of cases of class 1 where j-th variable has a particular value. We similarly define sets $Q_1^-(j)$ and $Q_0^-(j)$ for cases of the negative class. We can now define sensitivity and specificity of our model with respect to each variable value. So for variable j, let:

$$q_1^+(j) = \{x_i | x_i \in Q_1^+(j) \ \wedge \ f(x) = 1\} \tag{7}$$

$$q_1^-(j) = \{x_i | x_i \in Q_1^-(j) \ \wedge \ f(x) = 0\} \tag{8}$$

[1] In epidemiology the purpose of building a model is not to construct a predictor for individual cases, but to provide a numerical description of the relation between the variables and the outcome.

In other words, $q_1^+(j)$ is the set of true positives with $x_{ij} = 1$, while $q_1^-(j)$ is the set of true negatives with $x_{ij} = 1$. We can similarly define sets $q_0^+(j)$ and $q_0^-(j)$ to consist of correctly labeled cases with $x_{ij} = 0$.

Consider the accuracy of a model with respect to a particular class l, given a particular value c of the jth variable as $p_c^l(j)$. For example:

$$p_1^+(j) = \frac{|q_1^+(j)|}{|Q_1^+(j)|}. \tag{9}$$

Values $p_0^+(j)$,$p_1^-(j)$ and $p_0^-(j)$ are defined similarly. These values can be interpreted as estimates of probability of correctly classifying a point of a particular class given that its jth variable has a specific value (1 or 0). For a variable that is not significant we expect these values ($p_1^+(j)$ and $p_0^+(j)$, and $p_1^-(j)$ and $p_0^-(j)$) to be the approximately the same regardless of the value of this variable. On the other hand, if a variable is significant, we expect that knowing its value changes the probability of classifying a point correctly.

The well-known t-test can be used to determine whether the null hypothesis ($p_1^+(j) = p_0^+(j)$ or $p_1^-(j) = p_0^-(j)$) holds:

$$t^+(j) = \frac{|p_1^+(j) - p_0^+(j)|}{\sqrt{[\frac{n_1+n_2}{n_1 n_2}][\frac{(n_1-1)p_1^+(j)(1-p_1^+(j))+(n_2-1)p_0^+(j)(1-p_0^+(j))}{n_1+n_2-2}]}}, \tag{10}$$

where $n_1 = |p_1^+(j)|$ and $n_2 = |p_0^+(j)|$. The test $t^-(j)$ is defined similarly.

If the null hypothesis is rejected (for both $t^-(j)$ and $t^+(j)$) then the variable has a significant effect on the predictive accuracy. The variables for which the null hypothesis was rejected for both sensitivity and specificity at confidence level of 0.999999 are given in Tables 10 and 11.

This approach is efficient in that it makes use only of the predicted labels and requires a single pass through the test data. There are important drawbacks to such an approach. One is that the failure to reject the null hypothesis does not imply that the variable is not significant. For example, if a classifier has 100% accuracy then the above method will not detect any significant variables. On the other hand, a variable that has no or little effect on the predictions of the model (for example because of a small weight) but is correlated with the label may be deemed significant.

A more appropriate solution is to examine the effect of changing variable values in the test set. If a variable is not significant, then changing its value will not have an effect on the prediction, but if it is then the prediction should change. This approach is related to sensitivity analysis in control theory [5] and mathematical economics [16, 18]. We consider the following method for estimating variable significance: for variable j we flip all values x_{ij} in the test data (i.e. substitute 1 for 0 and vice-versa) and apply the classifier to the resulting dataset. Let f_j be the fraction of points on the test sets for which the prediction changes after flipping the jth variable. In our experiment with the baseline BBR the fractions varied from 0.0057 to 0.391. The correlation with the magnitude of the coefficients in the BBR model is large, as expected (see Table 12). The actual coefficients and the number of points

Table 10: Variables for which the null hypothesis was rejected for both sensitivity and specificity at confidence level of 0.999999 (method of [13]).

ID	Boolean Value	n	Sensitivity (%)	Specificity (%)	t^+	t^-
2	1	11603	67.0771	76.1317	8.03497	7.38129
2	0	85637	72.7363	71.9762	8.03497	7.38129
3	1	14662	68.3448	77.0692	7.13329	10.5189
3	0	82578	72.7561	71.6709	7.13329	10.5189
12	1	18514	77.5076	68.1855	13.1246	10.3341
12	0	78726	70.8076	73.4958	13.1246	10.3341
17	1	53995	74.2117	69.4205	11.6643	16.9613
17	0	43245	69.1335	75.9737	11.6643	16.9613
22	1	85742	73.3556	71.5788	16.7786	12.6527
22	0	11498	60.1663	78.2519	16.7786	12.6527
23	1	64965	75.4369	69.6815	23.9696	19.2754
23	0	32275	63.6204	77.2182	23.9696	19.2754
24	1	31135	80.0233	63.1854	30.934	27.391
24	0	66105	67.1123	75.8456	30.934	27.391
33	1	4060	79.9153	60.7549	9.15483	10.8102
33	0	93180	71.7118	72.9451	9.15483	10.8102
44	1	5545	78.4351	56.7234	8.96648	15.8615
44	0	91695	71.6137	73.2331	8.96648	15.8615
45	1	37960	70.3704	74.7738	6.29589	9.68494
45	0	59280	73.1547	70.9728	6.29589	9.68494
61	1	23598	5.86283	98.9841	163.419	150.708
61	0	73642	77.9792	56.397	163.419	150.708
62	1	7714	21.8642	94.9297	51.5734	71.3472
62	0	89526	74.1519	69.6392	51.5734	71.3472
63	1	8665	35.8199	86.8196	43.5234	32.2238
63	0	88575	74.7751	70.7942	43.5234	32.2238

Table 11: Variables for which the null hypothesis was rejected for both sensitivity and specificity at confidence level of 0.999999 (method of [13]), continued.

ID	Boolean Value	n	Sensitivity (%)	Specificity (%)	t^+	t^-
64	1	10964	84.0422	38.048	27.6346	49.1961
64	0	86276	70.0087	75.5766	27.6346	49.1961
66	1	33815	90.8913	20.8934	101.348	147.315
66	0	63425	51.0254	85.0562	101.348	147.315
70	1	38165	61.0833	87.4972	28.1441	83.7102
70	0	59075	75.7368	56.6395	28.1441	83.7102
73	1	26043	11.0525	98.7244	127.858	152.614
73	0	71197	78.4361	53.982	127.858	152.614
74	1	26158	11.2143	98.6702	127.301	152.313
74	0	71082	78.4678	53.896	127.301	152.313
75	1	41001	64.4181	74.5027	29.4149	9.10203
75	0	56239	77.3601	70.958	29.4149	9.10203
76	1	89501	74.6324	69.5978	64.9861	80.1986
76	0	7739	17.3558	96.1246	64.9861	80.1986
78	1	5814	16.4872	96.6959	58.3852	77.1109
78	0	91426	74.031	70.3407	58.3852	77.1109
82	1	22098	81.7162	56.3337	30.7345	36.4342
82	0	75142	68.3842	76.1376	30.7345	36.4342
84	1	19490	63.6929	80.0306	17.1474	22.4419
84	0	77750	73.8733	70.3392	17.1474	22.4419
94	1	77623	76.1077	59.3188	196.491	154.286
94	0	19617	2.40564	99.8024	196.491	154.286
95	1	74322	76.9136	55.9009	196.41	158.157
95	0	22918	2.94525	99.6905	196.41	158.157
96	1	56680	84.1354	41.2186	114.103	147.447
96	0	40560	27.1758	94.231	114.103	147.447
97	1	37940	89.6319	23.6201	100.953	152.387
97	0	59300	48.0605	87.2708	100.953	152.387

Table 12: Correlations between the magnitudes of coefficients of SVM and BBR, the effect of the changing a variable and a relevance measure.

	BBR			SVM		
	$\|w_j\|$	f_j	I_j	$\|w_j\|$	f_j	I_j
BBR $\|w_j\|$	1.00000	0.97499	0.81381	0.20495	0.19396	0.19647
BBR f_j	0.97499	1.00000	0.87078	0.21783	0.21307	0.21797
BBR I_j	0.81381	0.87078	1.00000	0.24009	0.25770	0.27057
SVM $\|w_j\|$	0.20495	0.21783	0.24009	1.00000	0.98247	0.97739
SVM f_j	0.19396	0.21307	0.25770	0.98247	1.00000	0.99948
SVM I_j	0.19647	0.21797	0.27057	0.97739	0.99948	1.00000

for which predictions changed, together with value I_j for each variable are given in Table 13. Notice that while the relationship is strong, there are several cases where a feature with large coefficient (by absolute value) affects fewer predictions than features with smaller coefficients, and the other way around. Two examples of this are features 64 (extension code 71-76) and 74 (surgery recommended).

Table 12 shows several other interesting relationships. While measure I_j is strongly correlated with the coefficients and with f_j, the correlation between $|w_j|$ and f_j is stronger. Also, the measures computed with the SVM model have little correlation with any of the measures based on the BBR weights, but show even stronger correlations between themselves.

6 Conclusions

In this work we have discussed SEER data preparation and construction of epidemiological models using modern machine learning methods (SVM and penalized logistic regression). Our experiments suggest that SVM may provide an interesting and different perspective from the more standard logistic regression approaches. Specialized implementations of linear SVM should make it comparable in computational complexity to the logistic regression.

We also considered several different feature importance measures. They are all strongly correlated within a particular method, but can also lead to different rankings for some variables.

Our future plans involve applying a hierarchical classification scheme with local models to the analysis of epidemiological data. The approach we consider involves partitioning the data into clusters, using K-Means clustering algorithm [14]. After that, for each large cluster we build a classifier to distinguish this cluster from the rest of the data, and we will build a local mode in the cluster. We then compare the local model to the global model in terms of accuracy and the features involved. This method will allow us to find "interesting" clusters in the data, where an interesting cluster is one consisting of a set of points that is distinguishable from the rest of the data and that allows for a different (better) local model

Table 13: BBR model coefficients w_j, corresponding I_j and f_j (multiplied by the total number of points in the test set), ordered by increasing f_j. Histology code "xyz*" means that the first three digits of the code are xyz.

ID	BBR w_j	Nf_j	BBR I_j	Description
46	0.019	552	0.069	Laterality: right
12	-0.040	1105	0.064	Registry: Los Angeles
63	0.046	1223	0.061	extension code 60-70
70	0.049	1332	0.100	Site specific surgery (code 10 or higher)
95	0.060	1579	0.105	Stage code 20 or higher
85	0.064	1707	0.112	Histology code 814*
45	-0.078	1987	0.137	Laterality: left
1	-0.101	2558	0.108	Registry: San-Francisco
84	0.101	2627	0.135	Histology code 807*
7	-0.185	4692	0.145	Registry: Seattle
62	0.202	4991	0.120	Extensions code 40-59
13	-0.206	5274	0.130	Place of birth: US
22	0.213	5310	0.153	Age 55 or greater
78	-0.219	5852	0.115	Radiation after surgery
17	0.236	5998	0.239	Sex: Male
15	0.252	6295	0.153	Race: Black
14	0.247	6551	0.181	Race: White
61	-0.237	6552	0.209	Extensions code 10-30
23	0.249	6562	0.240	Age 65 or greater
82	0.266	6630	0.213	Histology code 801*
27	0.278	6976	0.131	Born in Mid Atlantic Region
5	-0.272	7237	0.170	Registry: Iowa
44	0.308	7617	0.128	Primary Site: Bronchus
3	-0.290	7885	0.206	Registry: Detroit
9	-0.295	7981	0.134	Registry: Atlanta
34	0.327	8197	0.193	Born in Pacific region
2	-0.319	8657	0.193	Registry: Connecticut
96	0.303	8679	0.275	Stage code 31 or higher
26	0.379	9414	0.182	Born in South Atlantic
30	0.385	9506	0.161	Born in New England
28	0.389	9583	0.218	Born in East North Central
32	0.391	9607	0.144	Born in South West Cental region
29	0.384	9669	0.216	Born in West North Central
24	0.371	10182	0.268	Age 75 or greater
33	0.501	12026	0.152	Born in Mountain region
31	0.512	12292	0.152	Born in East South Central region
74	-0.501	15067	0.305	Surgery recommended
97	0.511	15549	0.343	Stage code 32 or higher
64	0.598	15837	0.243	Extension code 71-76
94	0.542	16182	0.297	Stage code 10 or higher
75	-0.573	16423	0.378	Radiation therapy
83	-0.599	16472	0.286	Histology code 804*
76	-0.950	19038	0.256	Radiation therapy with surgery
66	0.717	22367	0.401	Extension code 80-85
73	-1.197	37975	0.471	Surgery was performed

of the phenomena.

A number of other aspects of our current work may be of interest to the epidemiological community and can be developed further. For example, we have spent a great deal of effort preparing the data. While this process probably cannot be automated completely, there are ways of making it much faster and easier. Developing standard methods for data preparation would be of benefit to the medical and epidemiological community.

A Classification Methods

In many real-life situations we want to be able to assign an object to one of several categories based on some of its characteristics. For example, based on the results of several medical tests we want to be able to say whether a patient has a particular disease, should be recommended a specific treatment or is healthy.

In computer science such situations are described as classification problems. A binary (two-class) classification problem can be described as follows: given a set of labeled points (x_i, l_i), $i = \overline{1, N}$, where $x_i \in R^d$ are d-dimensional vectors of features and $l_i \in \{-1, +1\}$ are class labels, construct a rule that correctly assigns a new point x to one of the classes.

The vectors x_i in this formulation correspond to objects, and the dimensions of the space are the features or characteristics of these objects. For example, a vector may represent a person, with individual features corresponding to the measurements given by some medical tests — blood pressure, cholesterol level, white cell count and so on.

A classification method or algorithm is a particular way of constructing a rule, also called a classifier, from the labeled data and applying it to the new data.

A classifier can frequently be represented using a discriminant function $f(x) : R^d \to R$. In a two-class case, a point is assigned to the positive class if $f(x) \geq 0$, and to the negative class otherwise.

A classifier is linear if its discriminant function $f(x)$ can be expressed as:

$$f(x; w, b) =< w, x > +b \tag{11}$$

where w, b are parameters of the function and $<, >$ denotes the inner product of two vectors.

A set S of points (x_i, l_i), $i = \overline{1, N}$, where $l_i \in \{-1, +1\}$ are class labels, is called *linearly separable* if a linear classifier can be found so that $l_i f(x_i) > 0$, $\forall i = 1, \ldots, N$.

A.1 Logistic Regression

Logistic regression is a well-known statistical method, used in many applications. The aim of logistic regression is to model conditional probabilities of the class given the data. For such a task linear regression cannot be used because the outputs have to be restricted to the range $[0, 1]$. When dealing with a two-class problem (i.e. $l \in \{-1, +1\}$), this is achieved by applying the logistic transformation to the linear regression problem, leading to the following formula:

$$p(l|x) = \frac{1}{1 + exp(-l[< w, x > +b])}. \tag{12}$$

Binary logistic regression can be interpreted as modeling the log-odds ratio with a linear function:

$$\ln \frac{p(l = 1|x)}{p(l = -1|x)} =< w, x > +b. \tag{13}$$

The parameter vector w is usually found by solving the following optimization problem:

$$\text{minimize}_w \sum_{i=1}^{N} \ln(1 + exp(-l_i[< w, x_i > +b])$$

This optimization problem may be ill-conditioned (i.e. having multiple solutions). The usual way to address this is by adding a multiple of the identity matrix to the Hessian, which amounts to the formulation below:

$$\text{minimize}_w \ \lambda < w, w > + \sum_{i=1}^{N} \ln(1 + exp(-l_i[< w, x_i > +b]). \tag{14}$$

This formulation is sometimes referred to as "ridge" logistic regression (see [11]).

Another regularization method, called *lasso*, was suggested in [20]. It uses L1 penalty for the coefficients, resulting in the following criterion to be minimized:

$$\text{minimize}_w \lambda ||w||_1 + \sum_{i=1}^{N} \ln(1 + exp(-l_i[< w, x_i > +b]). \tag{15}$$

There is no analytical solution to minimizing (15) and numerical methods have to be used to find the optimal vector of coefficients.

Both lasso and ridge regression lead to feature selection, by shrinking feature weights towards zero. However, lasso sets coefficients of some variables to 0 exactly, thus removing them from the model, while ridge regression shrinks coefficient values but usually doesn't remove them completely. In this sense lasso is better for feature selection [20].

A.1.1 Bayesian Logistic Regression

A Bayesian approach to logistic regression is to specify a prior distribution of the coefficients w_i and then to find *maximum a posteriori* (MAP) estimates of w_i.

Setting the prior distribution for each coefficient independently to be Gaussian with mean 0 and variance σ^2 for all coefficients leads to the "ridge" logistic regression and is equivalent to (14), with $\lambda = \frac{1}{2\sigma^2}$.

It is also possible, in a hierarchical fashion, to specify a prior distribution for the variances σ_i^2. Using an exponential prior with parameter γ for parameter variances σ_i^2 leads to a Laplace prior for the coefficients, and to optimization problem (15) with $\lambda = \sqrt{\frac{2}{\gamma}}$, as shown in [20].

Recently, an efficient implementation of Bayesian logistic regression with Gaussian and Laplace priors was described in [8]. This software, Bayesian Binary Regression (BBR), is available online.

A.2 Support Vector Machine (SVM)

Support Vector Machines (SVM) recently became one of the most popular classification methods. They have been used in a wide variety of applications such as text classification [12], facial expression recognition [15], gene analysis [9] and many others.

Support Vector Machines can be thought of as a method for constructing a special kind of rule, called a linear classifier, in a way that produces classifiers with theoretical guarantees of good predictive performance (the quality of classification on unseen data). The theoretical foundation of this method is given by statistical learning theory [21].

In our work we used LIBSVM package [3].

A.2.1 Maximal Margin Classification

The idea of *margin* has come to play an important role in the theory of statistical learning.

A hyperplane

$$< w^*, x > +b^* = 0,\ ||w^*|| = 1 \tag{16}$$

is called a γ-margin separating hyperplane if

$$l_i[< w^*, x > +b^*] \geq \gamma \tag{17}$$

for all (x_i, l_i) in set S. Here γ (clearly $\gamma > 0$) is the margin.

Any separating hyperplane can be converted into this form. Suppose

$$l_i[< w, x > +b] \geq 1. \tag{18}$$

Then, by setting $w^* = \frac{w}{||w||}$ and $b^* = \frac{b}{||w||}$, we obtain a γ-margin separating hyperplane with $\gamma = \frac{1}{||w||}$.

It has been shown [21] that the probability of making a mistake on the unseen data can be bounded by a quantity that is inversely proportional to the size of the margin. Thus we would like to find a classifier with the largest margin that still correctly separates the training points.

The maximal-margin separating hyperplane can be found by solving the following optimization problem:

$$\text{minimize}_{w,b} < w, w > \tag{19}$$

subject to:

$$l_i[< w, x_i > +b] \geq 1, \ \forall i = 1, \dots, N \tag{20}$$

This problem can be solved using quadratic programming [21].

Similar machinery can be used to handle the non-separable case. The main idea is that those points that lie on the wrong side of the hyperplane are explicitly penalized by introducing *slack* variables, ξ, that control how far on the wrong side of a hyperplane a point lies.

The optimization problem becomes:

$$\text{minimize}_{w,b} \frac{1}{2} < w, w > +C \sum_{i=1}^{N} \xi_i \tag{21}$$

subject to:

$$l_i[< w, x_i > +b] \geq 1 - \xi_i, \ \xi_i \geq 0, \ \forall i = 1, \dots, N \tag{22}$$

where the parameter C, controlling the trade-off between the size of the margin and the training errors, is chosen by the user. This problem can also be solved using quadratic programming [21].

Both in the separable and non-separable cases the solution is given by a linear combination of inner products with some of the training vectors, called "support vectors": $w = \sum_{i=1}^{N} l_i \alpha_i x_i$, where α_i are parameters to be solved for. A training point x_i is a support vectors if corresponding $\alpha_i \neq 0$.

There are no general methods for choosing parameters b in a non-separable case - it is usually set to optimize some performance measure on a training or a validation set. The same approach is often taken for choosing a value of C.

B Tables of Features

Demographic Features		
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f1 1	SEER registry 2 digit code 1 - San Francisco - Oakland; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f2 1	SEER registry 2 digit code 1 - Connecticut; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f3 1	SEER registry 2 digit code 1 - Detroit; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f4 1	SEER registry 2 digit code 1 - Hawaii; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f5 1	SEER registry 2 digit code 1 - Iowa; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f6 1	SEER registry 2 digit code 1 - New Mexico; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f7 1	SEER registry 2 digit code 1 - Seattle; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f8 1	SEER registry 2 digit code 1 - Utah; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f9 1	SEER registry 2 digit code 1 - Atlanta; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f10 1	SEER registry 2 digit code 1 - Alaska; 0 - otherwise

FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f11 1	SEER registry 2 digit code 1 - San Jose - Montgomery; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f12 1	SEER registry 2 digit code 1 - LA; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f13 5	Place of birth 3 digit 1 - USA; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f14 8	Race: White 2 digit 1 - White; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f15 8	Race: Black 2 digit 1 - Black; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f16 9	Spanish Surname or Origin 1 digit 1 - Spanish; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f17 10	Sex 1 digit 1 - Male; 0 - female
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP. COMPUTER REP.	f18-f25 7	Age at diagnosis 3 digit 8-binary scale (0-25),(25-34),(35-44),(45-54),(55-64),(65-74),(75-84),≥ 85
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f26-f34 5	Region of birth in US 3 digit 9 binary variables: New England, 0 - otherwise Mid Atlantic, 0 - otherwise East North Central, 0 - otherwise West North Central, 0 - otherwise South Atlantic, 0 - otherwise

		East South Central, 0 - otherwise South West Cental, 0 - otherwise Mountain, 0 - otherwise Pacific, 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f35-f43 5	Region of birth outside US (with more than 1000 cases) 3 digit 9 binary variables: Canada, 0 - otherwise Mexico, 0 - otherwise UK, 0 - otherwise Scandinavia, 0 - otherwise Germanic Countries, 0 - otherwise Romance language Countries, 0 - otherwise Slavic Countries, 0 - otherwise Indochina, 0 - otherwise East Asia, 0 - otherwise

Table 14: Demographic Features

Medical Features

FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f44 14	Primary Site 4 character 1 - bronchus; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f45-f47 15	Laterality - left, right or both 1 digit 3 binary: 1 - left organ; 0 - otherwise left organ, 0 otherwise right organ, 0 otherwise both organs, 0 otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f48-f51 18	Grade 1 digit 4-scale binary: 1000 - I 1100 - II 1110 - III 1111 - IV, 5-8
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f52 20	Primary Not Found 12 digit 1 - primary not found or not stated (000 or 999); 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f53-f60 20	Size 12 digit 8-binary scale 1 - 001 or 002 2 - 003 3 - 004-005 4 - 006-010 5 - 011-015 6 - 016-020 7 - 020-990 8 - 998
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f61-f66 20	Extension 12 digit 6 separate binary variables

		10-30,40-59,60-70,71-76,77-79,80-85
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f67 20	Lymph Involvement 12 digit 1 - 1-8; 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f68 20	Lymph Regional 12 digit 1 - (1-6); 0 - 7
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f69-f72 21 (until 1998), 66 (1998+)	Site specific surgery 2 digits 4-binary scale 1000 - 0 1100 - 10,20,30,40,90 1110 - 50 1111 - 60,70 9 or 99 - missing -99
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f73 22	Surgery performed 1 digit 1 - yes; 0 - no
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f74 22	Surgery recommended 1 digit 1 - yes, if was not performed (see prev.); 0 - otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f75 23	Radiation 1 digit 1 - performed (1-6); 0 - none, or refused (0,7);
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f76 25	Radiation with surgery 1 digit 1 - performed (2-9); 0 - not
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f77 25	Radiation prior to surgery 1 digit 1 - yes; 0 - no
FIELD NUMBER/NAME SEER ITEM NUMBER	f78 25	Radiation after surgery

SEER DESCRIPTION COMPUTER REP.		1 digit 1 - yes (2); 0 - no
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f79 25	Radiation before and after surgery 1 digit 1 - yes(3); 0 - no
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f80 28	Behavior (recode of field 17) 1 digit 1 - ’3’ (malignant); 0 - ’2’ (in situ)
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f81-f86 27	Histology (most frequent codes ¿ 10000) 4 digit 6 binary variables (“xyz*” means that the first three digits are xyz): 800*, 0 otherwise 801*, 0 otherwise 804*, 0 otherwise 807*, 0 otherwise 814*, 0 otherwise 825*, 0 otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f87-f92 27	Histology (medium frequency codes: ¿ 1000) 4 digit 6 binary variables: 802*, 0 otherwise 803*, 0 otherwise 824*, 0 otherwise 826*, 0 otherwise 848*, 0 otherwise 856*, 0 otherwise
FIELD NUMBER/NAME SEER ITEM NUMBER SEER DESCRIPTION COMPUTER REP.	f93-f98 34	SEER modified AJCC stage 3rd ed 2 digit 6-binary scale 00,10,20,31,32,40

Table 15: Medical Features

References

[1] Surveillance, epidemiology and end results (seer) program of the national cancer institute. http://seer.cancer.gov/about/.

[2] E. Boros, P. L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik. An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering*, 12(2):292–306, 2000.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] D. Delen, G. Walker, and A. Kadam. Predicting breat cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in medicine*, 34(2):113–127, 2005.

[5] A. L. Dontchev. *Perturbations, approximations, and sensitivity analysis of optimal control systems*. Springer-Verlag, 1983.

[6] J. H. Friedman and J. J. Meulman. Multiple additive regression trees with applications in epidemiology. *Statistics in Medicine*, 22:1365–1381, 2003.

[7] J. H. Friedman and B. E. Popescu. Predictive learning via rule ensembles. Technical report, Stanford University, 2005.

[8] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. Software: http://mms-01.rutgers.edu/ ag/BBR/.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[10] H. Han, I.-Y. Song, X. Hu, A. Prestrud, M. F. Brennan, and A. D. Brooks. Managing and mining clinical outcomes. In Y. L. et al, editor, *DASFAA 2004*, number 2973 in LNCS, pages 405–416. Springer-Verlag, 2004.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer, New York, 2001.

[12] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*. Springer, 1998.

[13] J. Li, I. Muchnik, and D. Schneider. Influences on breast cancer survival via svm classification in the seer database. In J. Abello and G. Cormode, editors, *To appear in DIMACS Volume on Data Mining in Epidemiology*. 2006.

[14] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkley Symposium on Mathematical Sciences and Probability*, pages 281–297, 1967.

[15] P. Michel and R. E. Kaliouby. Real time facial expresion recognition in video using support vector machines. In *Proceedings of ICMI'03*, pages 258–264, 2003.

[16] J. Roberts. Sensitivity of elasticity estimates for oecd health care spending: analysis of a dynamic heterogeneous data field. *Health Economics*, 8(5):459–472, 1999.

[17] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.

[18] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Wiley, 2004.

[19] W. N. Street. A neural network model for prognostic prediction. In *ICML'98*, 1998.

[20] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.Series B (Methodological)*, 58(1):267–288, 1996.

[21] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 2nd edition, 1998.