Finding and Interpreting Local Models in Analysis of Epidemiological Data

Dmitriy Fradkin

Rutgers University

Joint work with Dona Schneider and Ilya Muchnik

Plan

- Introduction
- Our Approach
- Example: Lung Cancer Survival Data
  - Data Preparation
  - Analysis of Global and Local Models

Machine Learning Background

- A set W, $|W| = N$, of points $x_j \in R^d$,$j = 1, \ldots, N$, with labels $y_j \in L_1, \ldots, L_K$.
- Classification is the task of constructing a classifier (rule) R that, given a point x, predicts/assigns it to a particular class. A classifier is usually constructed from a set W of labeled training examples.
- Clustering is the task of grouping a set of unlabeled points into k clusters S_i, $i = 1, \ldots, k$.

Classification Models for Description

- A classifier can be used as an automated classification/prediction tool; or as a description of the phenomena or data. These uses are different and their relative importance is largely application dependent.
- We propose using a simple combination of cluster analysis with classification for finding local models that may be more appropriate than global ones, while simultaneously finding clusters of related instances.
- The comparison between selected local models and the global model, and the identification of the characteristics of the clusters may allow domain experts to obtain new insights and directions for further work.

Interesting Subsets

Intuitively, an “interesting” subset of data is a subset where a general model of the data fails to capture the target relation or does so in an overly complicated way. Such subsets are of particular interest when they have (relatively) simple descriptions.

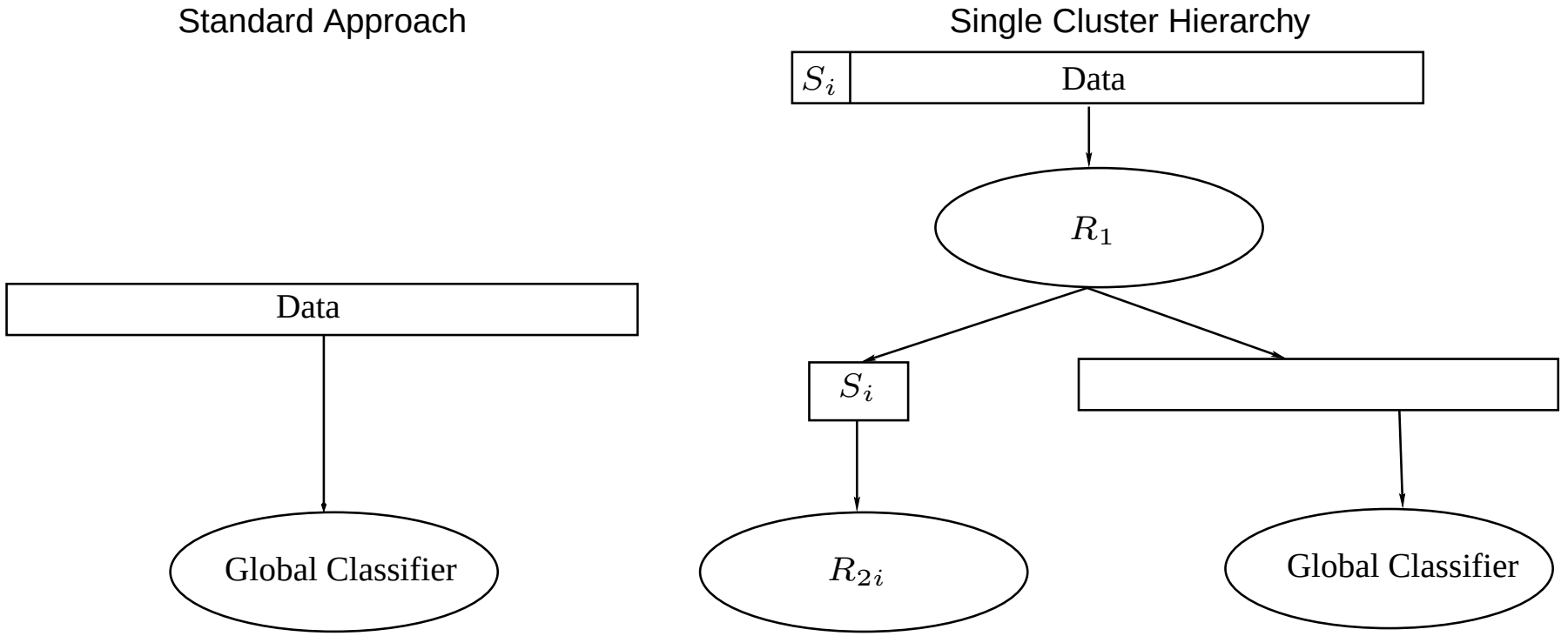


Figure 1: Global Classifier approach and Single Cluster Hierarchy

Single Cluster Hierarchy

Require: A set W, cluster $S_i \in W$. {Training stage}

1: Train a global classifier R_0 to distinguish between classes of W.
2: Train classifier R_{1i} to separate S_i (class 1) and W/S_i (class 0).
3: Train classifier R_{2i} on points in S_i.
4: Return R_0,R_{1i} and R_{2i}.

Require: Classifiers R_0,R_{1i} and R_{2i}; a point x. {Test stage}

1: Let $c = R_{1i}(x)$.
2: **if** $c = 1$ **then**
3: Return $R_{2i}(x)$.
4: **else**
5: Return $R_0(x)$.
6: **end if**

This is repeated in turn for all clusters in the data.

Types of Features

Estimating “feature significance” in a model is an well-known problem. The purpose is to assess the effect of feature values on predictions.

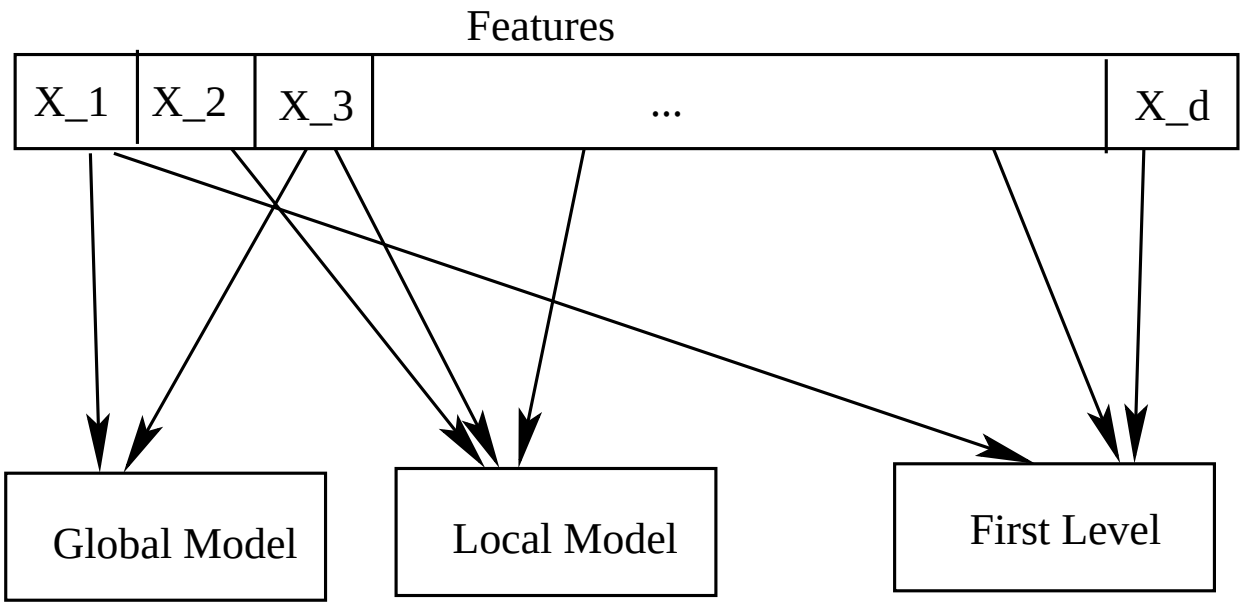


Figure 2: Our approach creates 3 (possibly overlapping) groups of significant features.

Lung Cancer Survival Analysis

- Pre-processing
  - Extracting raw data
  - Constructing features
  - Creating labels and selecting cases
  - Handling missing data
- Constructing a global model
- Applying the our approach and analyzing local models

About SEER Data

- The Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute (http://seer.cancer.gov/about/) is an authoritative source of information about cancer incidence and survival in the United States.
- Records are stored 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. (Sometimes there are multiple records for a patient).
- The SEER database has evolved over time and therefore certain kinds of information available in recent years are not present in older records. The 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.
- Information for each patient can be partitioned into two sets: demographic and medical.

Constructing Features

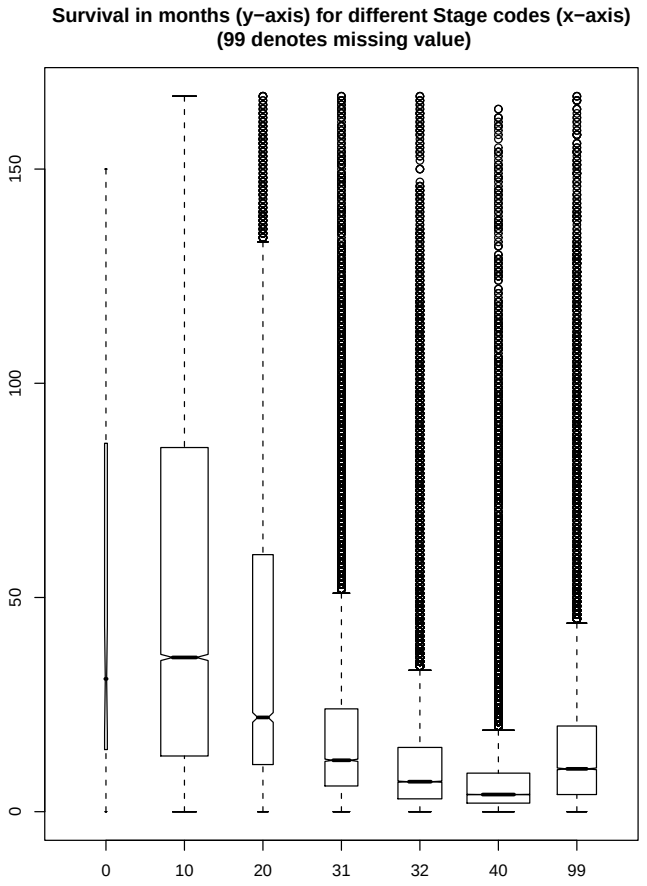
The fields in a SEER record can be grouped into 3 types:

- categorical: m possible values can be represented by m binary variables where x_i has value 1 only if the i-th category occurred in the field.
- ordinal: the values in these fields can be ordered but there is no distance function defined. An ordinal variable v taking values $\{1, \ldots, m\}$ can be represented by an m-tuple of binary variables v_i,$i = 1, \ldots, m$:

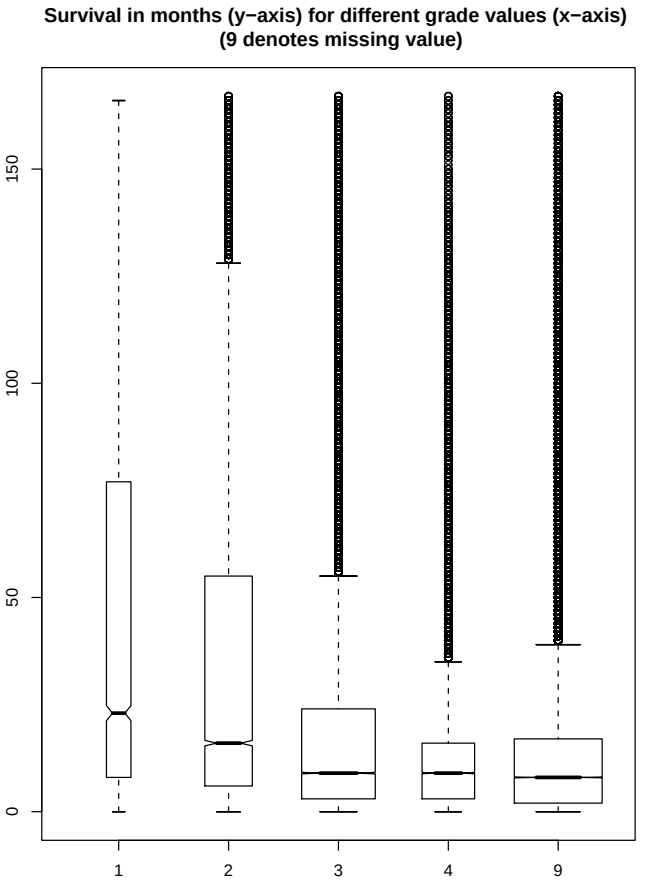
$$v_i = 1 \iff v \geq i \tag{1}$$

- numeric (age): can be partitioned into m intervals and treated as an ordinal variable with m levels.

Feature Analysis I

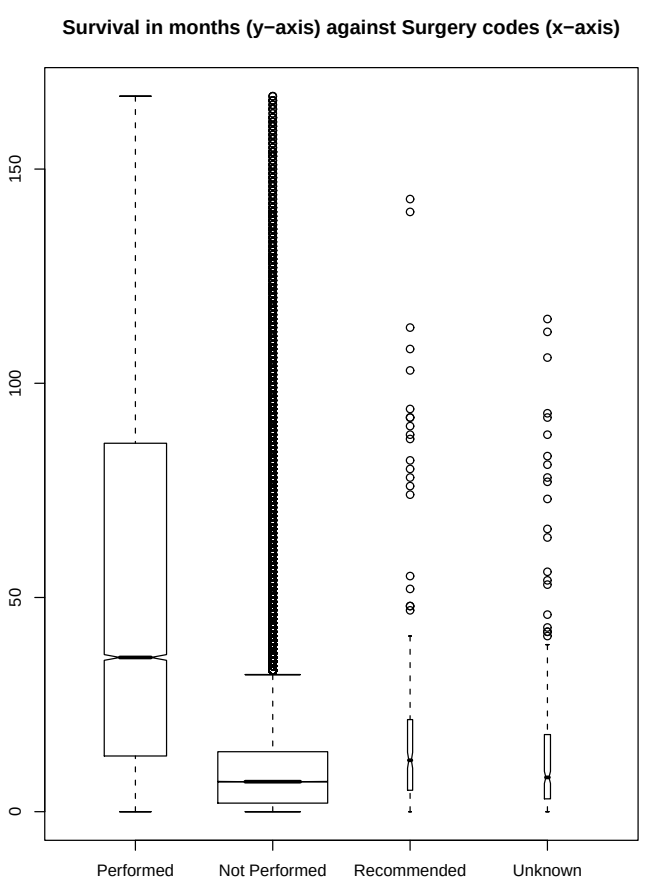


Stage

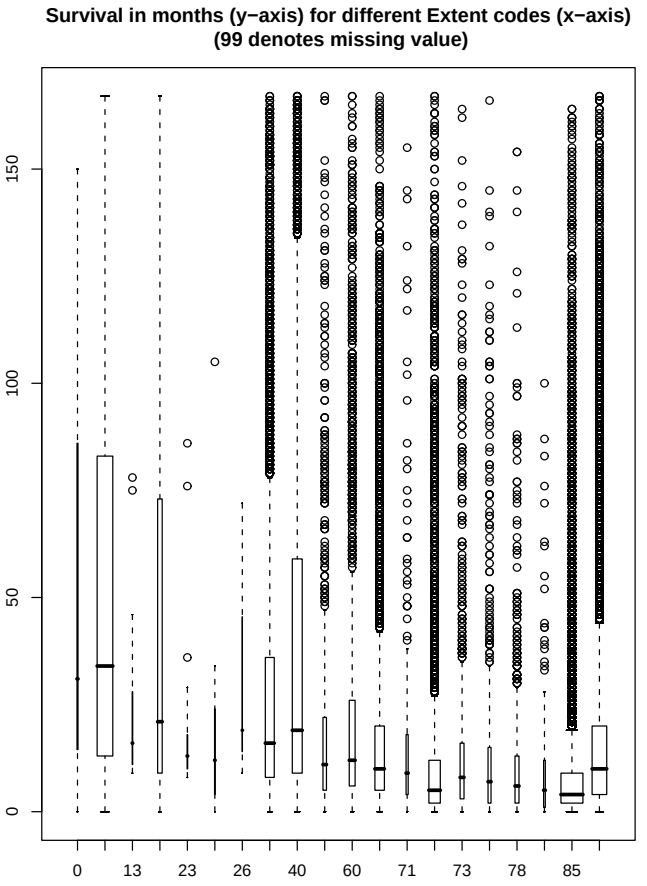


Grade

Feature Analysis II

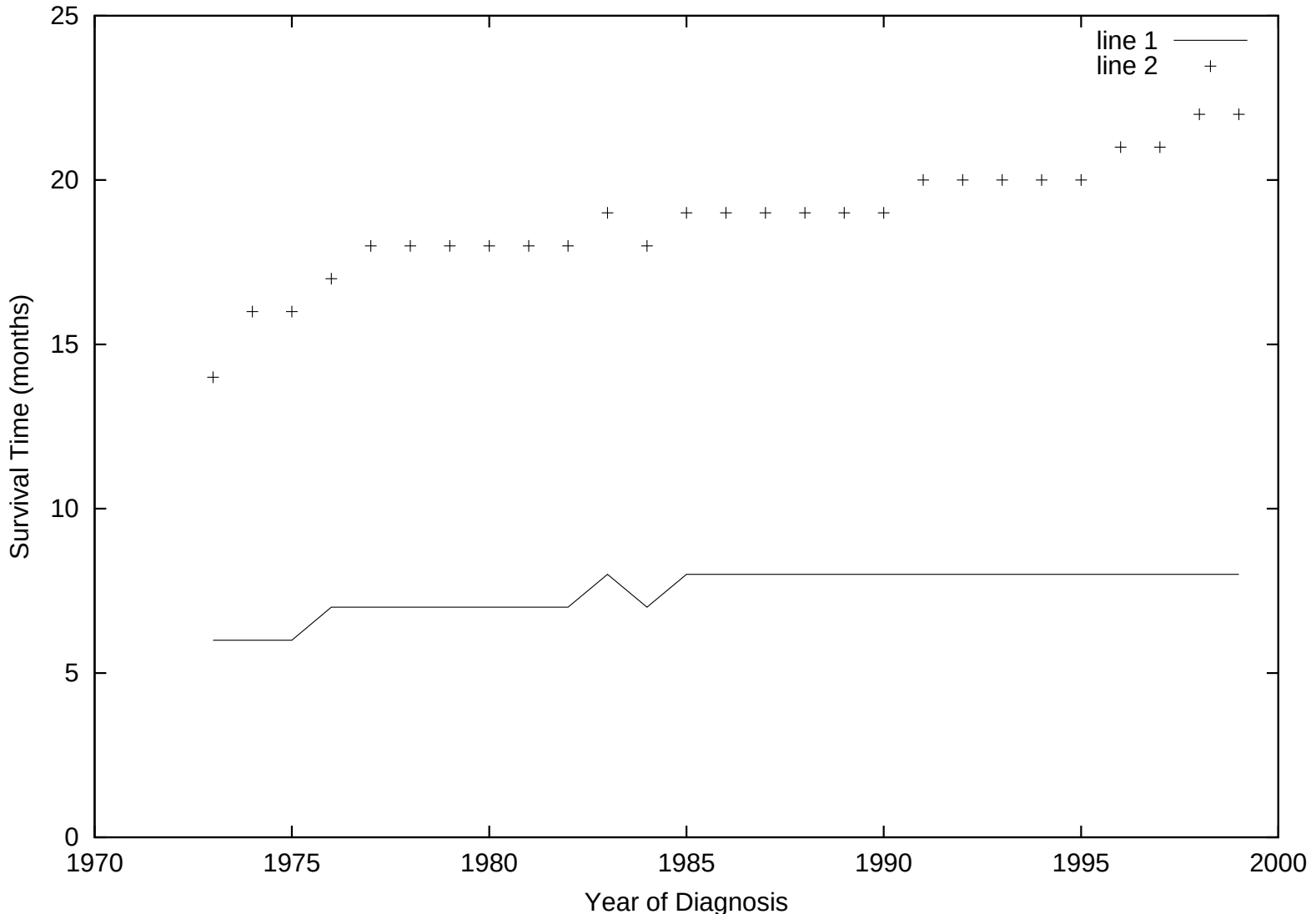


Surgery



Tumor Extent

Survival


line 1
line 2
Survival Time (months)
0
5
10
15
20
25
1970
1975
1980
1985
1990
1995
2000
Year of Diagnosis

Class Labels

Only the data from years 1988+ was used. The class cut-off was chosen as 8 months (median survival time):

- A person is labeled as class 1 if (s)he died of cancer of lungs or bronchus, and survival time was less than 8 months.
- If a person survived longer than 8 months, then (s)he is assigned to class 0.
- Cases where the cause of death was not lung cancer, or where class label could not be determined were discarded.

The selected data were split approximately evenly into a training set (1988-1995) and a test set (1996-2001).

Missing Value Analysis

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

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

Baseline Result

BBR - penalized logistic regression:

- The prior variance selected was 1.
- The threshold tuning parameter of BBR was set to minimize the sum of errors on the training set.
- The training (including cross-validation for parameter selection) takes approximately 30 minutes for the whole training set.
- Results on the test set: 72.10 sensitivity and 72.50 specificity; 72.32 accuracy.

ID j	Relative importance r_j	Coefficient w_j	Description
73	0.080	-1.197	Surgery was performed
76	0.063	-0.950	No radiation sequence with 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

Table 1: Features (from BBR), sorted by importance that add to 50% of the total weight.

Local Models: Test Set

i	$\alpha(SCH\|W,*)$	$\alpha(R_{2i}\|S_i,*)$	$\alpha(R_0\|S_i,*)$	Classes in S_i	Classes in W/S_i
4	$(71.76, 72.80)$	$(58.65, 70.68)$	$(62.31, 66.97)$	$(4126, 4287)$	$(40265, 48562)$
5	$(72.16, 72.45)$	$(2.36, 99.83)$	$(0.10, 100.00)$	$(1016, 14350)$	$(43375, 38499)$
6	$(73.09, 71.75)$	$(41.23, 89.94)$	$(14.60, 97.38)$	$(1637, 5310)$	$(42754, 47539)$
15	$(72.07, 72.58)$	$(70.88, 60.17)$	$(71.48, 58.56)$	$(2840, 2792)$	$(41551, 50057)$
16	$(71.90, 72.72)$	$(75.21, 53.69)$	$(81.99, 43.70)$	$(1327, 1151)$	$(43064, 51698)$

Table 2: Predictive performance (sensitivity, specificity) of the local and global classifiers inside the cluster on the test data, together with cluster size (on the test set). The overall accuracy of the global classifier is 72.32; sensitivity and specificity are $(72.10, 72.50)$.

Correlations between Models

j	Correlations of coefficients		
	R_0 and R_{2i}	R_0 and R_{1i}	R_{2i} and R_{1i}
1	0.234	0.081	-0.175
2	0.610	-0.094	0.242
3	0.395	0.132	-0.142
4	0.351	0.091	0.022
5	0.083	0.128	0.200
6	-0.245	0.454	-0.245
7	0.654	-0.144	-0.147
8	0.643	-0.062	-0.151
9	0.147	-0.103	-0.390
10	0.333	-0.133	0.019
11	0.531	-0.024	0.036
12	0.670	-0.142	0.004
13	0.502	0.068	-0.188
14	0.656	-0.014	-0.106
15	0.600	-0.046	-0.159
16	0.324	-0.202	-0.057

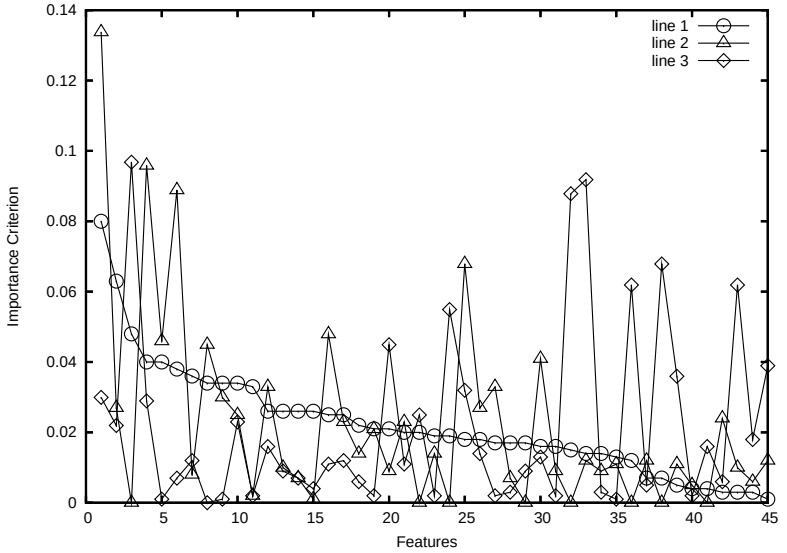
Table 3: Correlations between coefficient weights of the global classifier and those of local and first level classifiers.

Grouping Clusters by Features

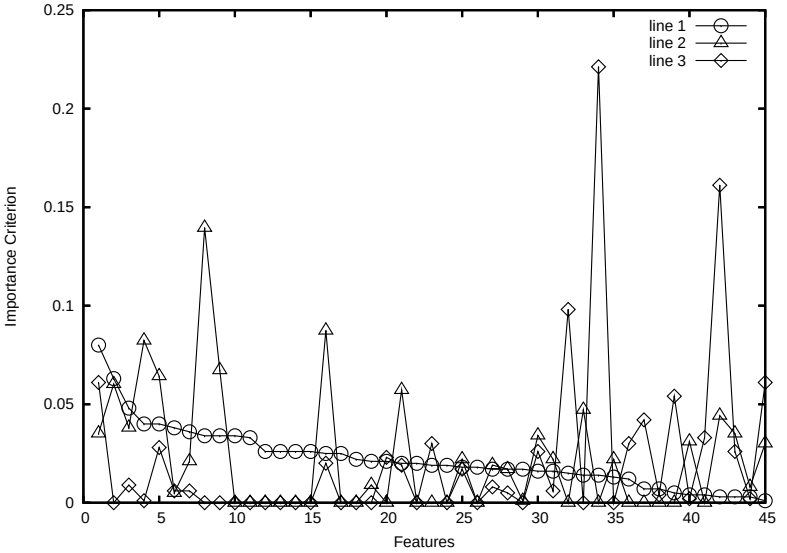
It is possible to group clusters based on the locally important features. Consider histology (82-85) and stage (94-97) features. Features 83 (histology code 804*), 94 (stage code 10 or higher) and 97 (stage code 32 or higher) are important in the global classifier.

- In clusters 2, 7, 8, 12, 14, 15 out of all the histology and stage features only feature 83 is important.
- In clusters 4, 10, 13 only some of the stage features are important.
- In clusters 3, 5, 9, 11, 16 both stage and histology features are important.
- In clusters 1 and 6 none of these two groups of features are important.

Plots of Feature Significance



Cluster 15



Cluster 16

Figure 4: Plots of coefficients for the global classifier and local and first-level classifiers in each cluster.

Example of Local Analysis: Cluster 15

j	R_0	R_{2i}	R_{1i}	μ_j	Description
1	-	-	1	0.00	Registry: San-Francisco
3	-	-	1	0.00	Registry: Detroit
7	-	-	1	0.00	Registry: Seattle
12	-	-	1	0.86	Registry: Los Angeles
13	-	-	1	1.00	Place of birth: US
78	-	-	1	0.00	Radiation after surgery
66	1	-	1	0.00	Extention code 80-85
32	1	-	-	0.14	Born in South West Central region
33	1	-	-	0.07	Born in Mountain region
74	1	-	-	0.07	Surgery recommended
76	1	-	-	1.00	No radiation sequence with surgery
94	1	-	-	0.98	Stage code 10 or higher
97	1	-	-	0.04	Stage code 32 or higher
31	1	1	-	0.04	Born in East South Central region
64	1	1	-	0.45	Extention code 71-76
73	1	1	-	0.07	Surgery was performed
75	1	1	-	0.43	Radiation therapy
83	1	1	-	0.19	Histology code 804*
5	-	1	-	0.01	Registry: Iowa
24	-	1	-	0.28	Age 75 or greater

Example of Local Analysis: Cluster 15

- All born in the USA
- Almost all from LA
- No cases with extention code 80-85 (which is globally significant)
- Almost no cases with stage 32 or higher
- Age 75 or greater is locally significant, but US region of birth and surgery recommendation are not.

Example of Local Analysis: Cluster 16

j	R_0	R_{2i}	R_{1i}	μ_j	Description
12	-	-	1	0.99	Registry: Los Angeles
13	-	-	1	0.00	Place of birth: US
78	-	-	1	0.00	Radiation after surgery
73	1	-	1	0.06	Surgery was performed
94	1	-	-	0.98	Stage code 10 or higher
75	1	-	-	0.46	Radiation therapy
66	1	-	-	0.44	Extention code 80-85
31	1	-	-	0.00	Born in East South Central region
32	1	-	-	0.00	Born in South West Central region
33	1	-	-	0.00	Born in Mountain region
64	1	1	-	0.22	Extention code 71-76
74	1	1	-	0.06	Surgery recommended
76	1	1	-	1.00	No radiation sequence with surgery
83	1	1	-	0.15	Histology code 804*
97	1	1	-	0.46	Stage code 32 or higher
24	-	1	-	0.29	Age 75 or greater
96	-	1	-	0.74	Stage code 31 or higher

Example of Local Analysis: Cluster 16

- All born outside the USA
- Almost all from LA
- Age 75 or greater, and stage code 31 or greater are locally significant.

Summary

- In Epidemiology (and in many other fields) there is need for interpretable models and ways of describing data.
- Before data mining/analysis can be applied, data has to be appropriately prepared. This is a complicated, time-consuming and domain-specific process.
- Data mining methods can suggests models and hypothesis, but these have to be evaluated by experts in a real world.
- We described an approach (combination of clustering and machine learning) for finding descriptions of interesting subsets of data.
- The description is via three kinds of features (global, local and first-level).

Directions for Future Work

- In order to make methods of data analysis more efficient, they have to be user-friendly and intuitive. One way of achieving this is by providing tools for visualization and other decision support, to aid the user in model and parameter selection and in interpreting results.
- In the analysis of epidemiological data boolean vectors were treated as belonging to a Euclidean space. However, by utilizing the information about the nature of the data it should be possible to apply specific methods and obtain better results. Development of analogous methods for boolean data is a promising research direction.
- The choice of features to be used in clustering clearly has a great effect on the results. We have only considered clustering in the space of all features. However, in practice it may be beneficial to consider selecting a subspace for clustering. One direction for future work is to examine the effect of different methods for feature subset selection.