Using cluster analysis to determine the influence of demographics on medical status of lung cancer patients

Dmitriy Fradkin

Ask.com
dfradkin@gmail.com

Joint work with Dona Schneider and Ilya Muchnik

Overview

- Epidemiology is interested in disease patterns in populations.
- One specific approach is to look for effects of demographic features on the medical characteristics.
- We describe how cluster analysis can be used to discover such effects/relations.
- This is illustrated with an example analysis of Lung Cancer Data from SEER.

Plan

- Our Approach
- Example: Lung Cancer Survival Data
  - Data Preparation
  - Cluster Analysis
  - Analysis of Results

Our Method

1. Identify a set of features (F_1) to be used in cluster analysis, and keep the remaining features (F_2) for analysis.
2. Perform cluster analysis in the space F_1, obtaining a partition into k clusters.
3. Compute statistics (mean, st. dev.) on the distribution of both F_1 and F_2 in the clusters.
4. Examine the distributions of the features in F_2 for significant differences or similarities across the clusters, and for interactions with features in F_1.

Potential Things of Interest

Focusing on F_2, look for:

- Features whose distribution is different across (almost) all clusters.
- Features whose distribution is the same in (almost) all clusters.
- Clusters that are not separated by any features.
- Clusters that are separable by (almost) all features.
- Clusters that are separable from (almost) all others by a subset of features.

Lung Cancer Data Analysis

- Pre-processing
  - Extracting raw data
  - Constructing features
  - Handling missing data
- Applying cluster analysis
- Analysis of the Results

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.

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 (23 demographic and 22 medical) and 217,558 cases are left. Small changes the processing have little effect.

Applying Cluster Analysis

- We used K-Means [Forgy 1965,McQueen 1967] - minimizes the sum of intra-cluster variances (weighted by cluster sizes).
- Specified $k = 20$.
- Applied K-Means 10 times with different initial conditions and selected the best result.
- The average cluster size is 10,878. The largest and smallest clusters (2 and 6) have 18,110 and 4,234 points respectively.
- Computed mean values for all features (both F_1 and F_2) in each cluster.

Performing Comparisons

- We can make pairwise comparisons between the clusters for each medical feature and look for statistically significant differences using t-test.
- There are $\frac{20*19}{2} = 190$ comparisons for each feature, and 4,180 comparisons altogether.
- At significance level $\alpha = 0.1$ we would expect 418 significant results by chance - we have 1733.

Focusing on Clusters

- For each pairs of clusters, there is at least 1 medical feature whose distribution is significantly different in these clusters:
  - Clusters 9 and 14 differ in: Histology code 807*
  - Clusters 14 and 20 differ in: Site specific surgery (code 10 or higher)
- Each pairs of clusters differs in distribution of at most 17 (out of 22) medical features.
- No feature has different distributions in all clusters.

Focusing on Features

- Some features have the same distribution in all clusters:
  - Laterality of the tumor
  - Extentions code 40-59
- Several features, such as Histology and Surgery Performed have different distributions in a lot of clusters:
  - Site specific surgery (code 10 or higher): 135 pairs are different
  - Surgery recommended: 128 pairs are different
  - Surgery performed: 127 pairs are different
  - Histology code 807*: 124 pairs are different

Interesting Clusters

- "Interesting clusters" - those that have a subset of features with distributions different from almost all other clusters

Using $\alpha = 0.05$ in order to be more selective:
Cluster 3:

- The cluster can be characterized as: all less than 65 years old, Detroit registry
- Compared to other clusters: high surgery and radiation rates separately but not together

Interesting Clusters (cont'd)

Cluster 6:

- The cluster can be characterized as: old (over 75) black people
- Compared to other clusters: low rates of surgery/radiation

Cluster 13:

- The cluster can be characterized as: females, younger than 65, almost all white, US born
- Compared to other clusters: low rates of 807*, high surgery/radiation values

Cluster 16:

- The cluster can be characterized as: Detroit registry, almost all white, older than 65, largely Born in East North Central
- Compared to other clusters: low rates of site specific surgery (code 10 or higher)

Interesting Clusters (cont'd)

		Clusters			
		3	6	13	16
3	Registry: Detroit	1.000	0.378	0.000	0.935
13	Place of birth: US	0.941	0.989	0.878	0.979
14	Race: White	0.991	0.000	0.875	0.997
15	Race: Black	0.000	0.999	0.000	0.000
17	Sex: Male	0.560	0.609	0.000	0.557
22	Age 55 or greater	0.671	1.000	0.615	1.000
23	Age 65 or greater	0.000	1.000	0.000	1.000
24	Age 75 or greater	0.000	1.000	0.000	0.280
28	Born in East North Central	0.602	0.041	0.083	0.772
70	Site specific surgery (code 10 or higher)	0.517	0.384	0.667	0.458
73	Surgery was performed	0.345	0.123	0.361	0.263
75	Radiation therapy	0.508	0.372	0.494	0.402
76	Radiation therapy with surgery	0.859	0.970	0.894	0.924
84	Histology code 807*	0.205	0.292	0.125	0.255

Summary

- We proposed and illustrated a method that uses cluster analysis in a feature subspace to find relations between features, which are validated with statistical tests.
- Such approach can be beneficial to epidemiology in finding relations between different types of features, such as demographic and medical ones, and in formulating focused studies.

Directions for Future Work

- An in-depth epidemiological study of one of the "interesting" clusters.
- Experimental work on other epidemiological datasets (other sources of data, other diseases).
- Experiments on synthetic data with various relations between features.