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

Dmitriy Fradkin

Ask.com
dmitriy.fradkin@ask.com

Joint work with Dona Schneider (Bloustein School of Planning and Public Policy, Rutgers) and Ilya Muchnik (DIMACS)

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.
- Previously, we have looked for significant variables in this data by analyzing survival prediction models (DIMACS Technical Report 2005-35). Here we look for groups of demographic variables that are indicative of certain values for medical variables.

Plan

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

Our Method

1. Identify a set of demographic features to be used in cluster analysis; keep the medical features for analysis.
2. Perform cluster analysis in the space demographic features, obtaining a partition into k clusters.
3. Compute statistics (mean, st. dev.) on the distribution of both demographic and medical features in the clusters.
4. Examine the distributions of the medical features for significant differences or similarities across the clusters, and for interactions with demographic features.
5. Validate the observed relations using a hold-out set.

Potential Things of Interest

Focusing on medical features, 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
- Validation

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. Partitioned the data into a training set (120,318) and a validation set (97,240), based on the year of diagnosis (1988-1996, and 1996-2001).

Applying Cluster Analysis

- We used K-Means [Forgy 1965,McQueen 1967] - minimizes the sum of intra-cluster variances (weighted by cluster sizes).
- Specified $k = 20$.
- The average cluster size is 6,016. The largest and smallest clusters (11 and 16) have 10,735 and 3,086 points respectively.
- Computed mean values for all features (both demographic and medical) 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. (This is a rough calculation, ignoring dependencies between features, and in comparisons based on the same features) We have 1141 significant results.

Focusing on Clusters

- Only two pairs of clusters do not differ in any medical features: 2 and 8; 12 and 5.
- For all other pairs of clusters, there is at least 1 medical feature whose distribution is significantly different in these clusters.
- The largest number of differences in medical feature distributions between clusters is 16 (out of 22) - clusters 11 and 19.
  - Cluster 11: Detroit, White, age > 65, born in US, 61% male
  - Cluster 19: Mixed registries, White, born in US, 80% under 55, 65% male

Focusing on Features

- No feature has different distributions in all clusters.
- Some features have the same distribution in all clusters:
  - Laterality of the tumor
  - Extentions code 40-59
- Several features, such as Histology Code 807* and Surgery Performed have different distributions in a lot of clusters:
  - Site specific surgery (code 10 or higher): 122 pairs are different
  - Histology code 807*: 95 pairs are different
  - Surgery performed: 80 pairs are different
  - Radiation Therapy: 79 pairs are different

Interesting Clusters

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

Cluster 11:

- The cluster can be characterized as: Detroit, White, age > 65, born in US, 61% male
- Compared to other clusters: lowest rate of Site-specific surgery (code 10 or higher).

Interesting Clusters (cont'd)

Cluster 13:

- The cluster can be characterized as: cases from smaller registries, all above 65 years old, born in US, White, female.
- Compared to other clusters: high rate of Surgery performed; relatively low rates of Stage codes 10 or higher, and 20 or higher.

Cluster 14:

- The cluster can be characterized as: under 75, White US born, Iowa, mostly male; Born in North Central region.
- Compared to other clusters: highest rate of Stage Code 32 or higher; highest rate of Site-specific surgery (code 10 or higher); highest rate of Extention code 80-85.

Cluster 18:

- The cluster can be characterized as: Detroit registry, almost all white, older than 65, largely Born in East North Central.
- Compared to other clusters: relatively low rate of site specific surgery; high rates of surgery and radiation therapy after surgery.

Validation of Results

- Group together points of the validation set based on the nearest cluster center for the training data.
- Perform significance analysis on the validation set and compare results.
- Observations:
  - Average cluster size is 4,862. Cluster 16 is still the smallest (1,919), but 11 is no longer the largest.
  - Features such as Laterality and Extention code 40-59 still have the same distributions across clusters.
  - Clusters 2 and 8; 5 and 12; are still not distinguished by any features.
  - The largest number of different features is still 16 (again, clusters 11 and 19).

Validation of Results (cont')

- Cluster 11: there are few significant differences with other clusters now.
- Cluster 13: still significant differences in frequencies of Stage code 10 or higher, and 20 or higher; Surgery performed.
- Cluster 14: The largest rate of Stage code 32 or higher (significantly different from many clusters); but not significantly different (from other clusters) rate of Site-specific surgery.
- Cluster 18: still high rate of Surgery performed, but other characteristics are not significantly different.

Validation of Results (cont')

		Training			Validation		
		11	13	14	11	13	14
3	Registry: Detroit	0.959	0.002	0.000	0.957	0.001	0.000
5	Registry: Iowa	0.000	0.149	0.958	0.000	0.151	0.961
17	Sex: Male	0.610	0.000	0.836	0.566	0.000	0.828
23	Age 65 or greater	1.000	0.673	0.485	1.000	0.695	0.485
24	Age 75 or greater	0.392	0.067	0.000	0.447	0.070	0.000
66	Extention code 80-85	0.313	0.304	0.415	0.310	0.310	0.439
73	Surgery was performed	0.235	0.321	0.242	0.272	0.353	0.308
94	Stage code 10 or higher	0.755	0.710	0.838	0.750	0.717	0.848
96	Stage code 31 or higher	0.721	0.671	0.793	0.720	0.677	0.815
97	Stage code 32 or higher	0.326	0.321	0.434	0.349	0.344	0.473

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 and hold-out data.
- Such approach can be beneficial to epidemiology in finding relations between different types of features, such as demographic and medical ones, studying changes in such relations, and in formulating focused studies.

Directions for Future Work

- Consider temporal effects on feature relations: which relations changed (appeared/disappeared), and whether they were "real".
- 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.