TB Treatment-Seeking Behavior

Ruixue Guo and Sze Suen
CS229: Machine Learning, Dec 2012

Background

For this project, we looked at a dataset containing tuberculosis (TB) symptoms, symptom onset time, and time until treatment was sought for 1691 non-urban households in Bihar, India. All households had at least one household member who was being treated for active TB disease. We are interested in identifying the relationship between an individual's characteristics (such as disease symptoms an individual experienced) and the time it took for the individual to seek treatment. This could potentially guide health education programs that support timely identification of disease cases to prevent transmission.

We first cleaned and prepared the data, explored some basic statistics to get an intuition for the dataset, then used several algorithms to see if we could accurately predict the time of seeking treatment, and, more importantly, use the algorithm outputs to identify reasonable characteristics of patients that sought care quickly. The methods we used were logistic regression, PCA, K-means clustering, and SVM. These results are presented below.

Data Exploration

The dataset provides care-seeking and disease symptom information on TB patients in clinics in the state of Bihar, India. We formatted and cleaned the data – variables with missing observations or observations with missing variables were dropped. Each observation also had a sample weight attached, due to the sampling design of the survey the data comes from, and we "unweighted" the data by repeating observations according to its weight (for instance, those observations with weights of 2 appear twice in the unweighted dataset, those with weights of 3 appear 3 times, etc.). We created the variables of interest from the raw data – our x variables will be the duration of the symptom prior to seeking care, for the seven symptoms we have information on (cough, cough with blood, fever, weight loss, night sweats, constant tiredness, loss of appetite). We also included a flag for each symptom indicating whether the symptom was present when the patient sought care. This allows differentiation between a patient who had 10 days with cough and did not seek care until much later, indicating the cough did not motivate him to seek treatment, and a patient who only had 10 days with cough prior to treatment because he started treatment due to becoming alarmed about the cough. For our y variable we used the time from the first symptom to the time of seeking treatment. This variable is bimodal (see Figure 1) with means around 13 days and 62 days. We use 43 days as the cutoff between "fast" care-seekers and "slow" care-seekers for use in categorical learning algorithms.

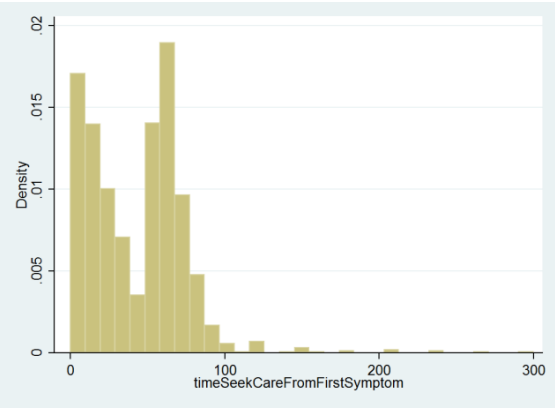

Figure 1

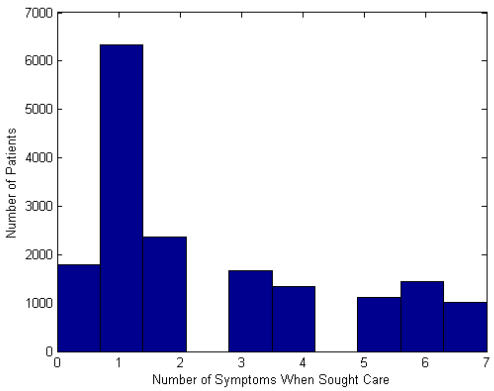

Figure 2

We first explored basic statistics about our dataset. We saw that the number of symptoms patients had experienced when they sought care varied from zero to seven out of the seven symptoms we have information about in the survey (see Figure 2). It does not seem that any one symptom had always been experienced by the time the patient sought care (see Figure 3), although from these graphs we can see that the majority of people had cough at the time they sought treatment (which is not the case for any other symptom).

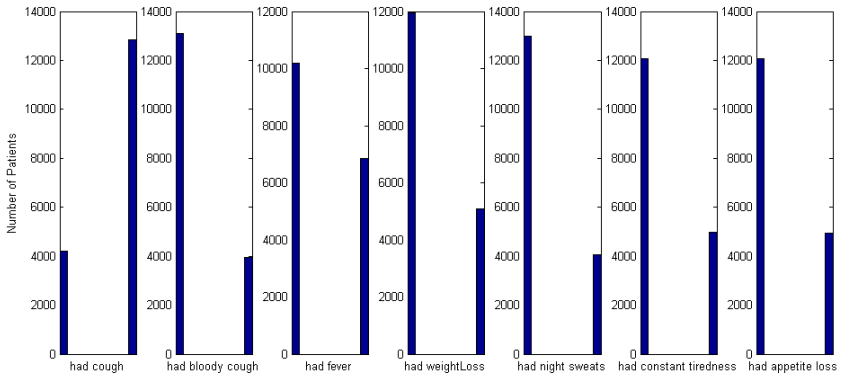

Figure 3

Principal Components Analysis

We used PCA to get a better visualization of our data. In our PCA analysis, we ranked the components along dimensions on which the variance of the data was maximized, allowing us to reduce the dimensions of the data until it could be visualized (see Figure 4). We found that 53% of the variation was captured by reducing to two dimensions, which increased to 62% if we use three dimensions.

In the figures below, "fast" treatment seekers are plotted in green, "slow" treatment seekers in red, and the blue points (X1 through X14) correspond to the variables. The first component is aligned with variation in all variables, whereas the second principal component seems to capture differences between the durations of symptoms (X1-X7) and the presence of the symptom at the time of treatment (X8-X14).

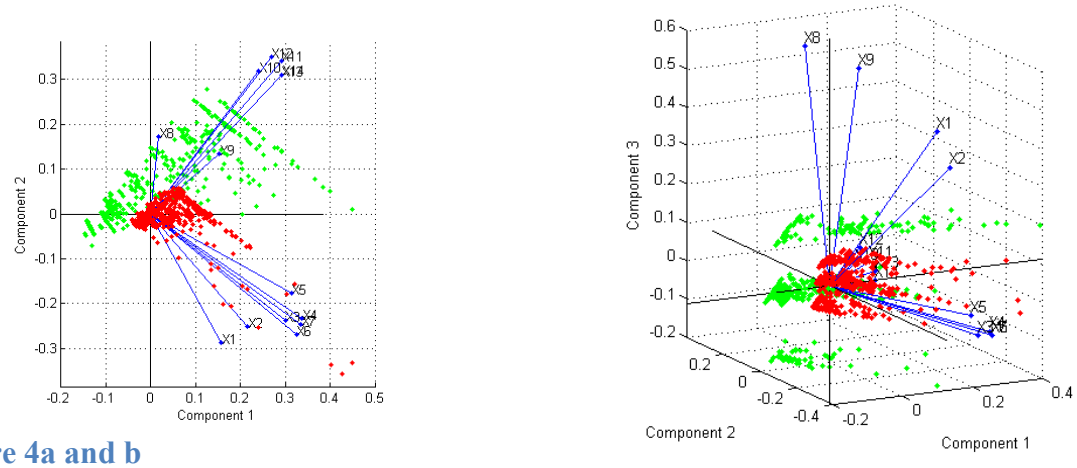

Figure 4a and b

The inclusion of the third principal component on our visualization shows that the third principal component captures mainly variation between variables coding for the presence of the

symptom at time of treatment. The presence of a cough (X8) or a bloody cough (X9) seems to be generating most of the variation in these variables. Fast and slow patients appear to be well separated, indicating that we might be able to use the PCA loadings to separate the types of patients. Figure 5 below shows our dependent variable plotted against such an index generated using the first principal component.

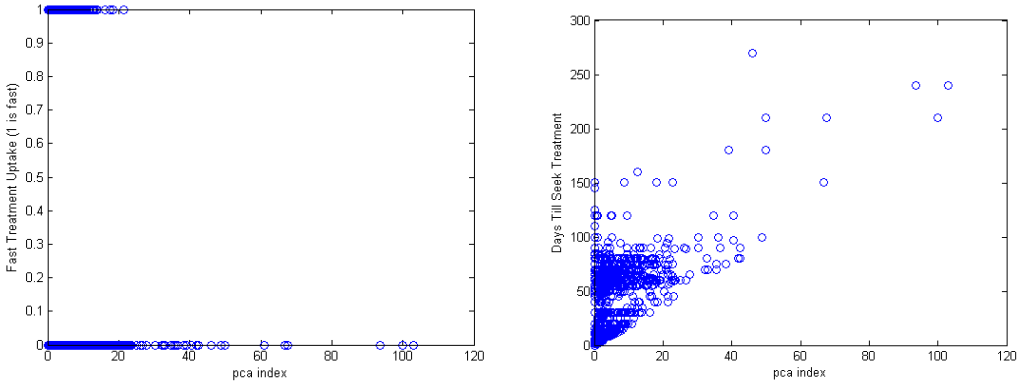

Figure 5a and b

Patients with PCA index larger than 28 are slow treatment-uptake patients. These are not the majority of patients, but we are able to identify them with perfect accuracy. Unfortunately, those with a PCA index smaller than 28 could be either fast or slow treatment-uptake patients and are not possible to separate.

K-means Clustering

From the PCA outputs, however, it seems like we should be able to generate clusters of fast and slow patients. We used K-means to generate "types" of patients. If the clusters the algorithm generated contain predominantly fast or slow treatment seeking patients, we could then make statements about the mean characteristics of these types.

Running two clusters until convergence generated two groups – a smaller group with 37% of the patients, and a larger group. The smaller group is almost completely slow-uptake type patients (95% slow-uptake). This group has, on average, longer durations for every symptom and has more symptoms at the time of seeking treatment for all symptoms. If we assign the majority label to all patients in the cluster, we are able to achieve 81% accuracy with these two clusters.

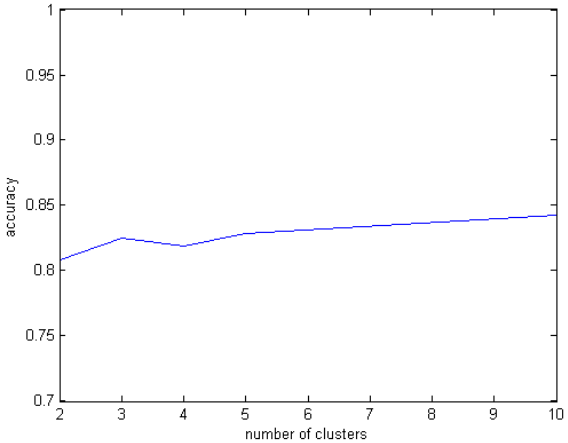

Figure 6

Are 2 clusters the ideal number of clusters? We examine the increase in accuracy with increasing the number of clusters (see Figure 6). Increasing the number of clusters does not significantly increase accuracy. Accuracies shown on the plot are "typical" outputs generated from 10 random initializations, so results are noisy.

Support Vector Machine

Training an SVM algorithm on 70% of our data achieved very high test accuracy (90%-93% accuracy with a linear kernel, and 98% accuracy with a quadratic kernel). However, the SVM approach is difficult to interpret, so the results are not clear from a hypothesis testing perspective – if we want to identify symptoms that encourage (or discourage) patients to seek care, this is not useful. A better approach might be logistic regression.

Logistic Regression

We randomly divided the 1691 (weighted) examples, i.e., 17040 (unweighted) examples, into 70% training examples and 30% testing examples. If a symptom was present at the time of seeking treatment, then this symptom was a driving factor why the patient went to seek treatment[1]. Since we're interested in which or which combinations of symptoms induce fast treatment-seeking behavior, we find all possible combinations of 2, 3, 4, or 5 symptoms[2], and pick out the ones that are significantly and (relatively) strongly correlated with the categorical y variable – the fast/slow treatment seeking behavior. In addition, having a certain or combination of symptom(s) at the time of seeking treatment, given this person is categorized as slow treatment seeking patient, implies that this or this combination of symptom(s) is in general not alarming, and induces slow treatment seeking. On the other hand, having a certain or combination of symptom(s) at the time of seeking treatment, given this person is categorized as fast treatment seeking patient, implies that this or this combination of symptom(s) is in general alerting and therefore induces fast treatment seeking. Further latent factors should also be taken into consideration when interpreting the results.

We have feature vectors as follows: (a) Duration of each symptom that is present at the time of seeking treatment; (b) Product of combinations of 2, 3, 4, or 5 symptoms' durations, that are present at the time of seeking treatment, given that having this symptom or combination of symptoms is significantly and (relatively) strongly correlated with the fast/slow treatment seeking behavior. Roughly we get 24.5% error. Based on the regression result, we find that[3] (see Table 1), (a) for single symptom, having one more day of bloody cough or appetite loss increases the probability of being a fast treatment seeker while having one more day of cough, fever, night sweat, or constant tiredness increase the probability of being a slow treatment seeker; (b) for combination of 2 symptoms, having one more day of both bloody cough and cough at the same time does not become more alarming than having one more day of bloody cough alone; (c) for combination of 4 symptoms, having one more day of cough, fever, weight loss and constant tiredness increases the probability of being a fast treatment seeker, even though having one more day of each one of these symptoms individually does not increase the probability of being a fast treatment seeker; (d) for combination of 5 symptoms, having one more day of cough, fever, weight loss, constant tiredness, and appetite loss increases the probability of being a slow

[1] We realize that this symptom might not necessarily be the driving factor why a patient went to seek treatment, but in general, this statement should hold.

[2] We didn't look at 6 or all 7 symptoms, because having those many symptoms may offer less insight for us in terms of fast or slow treatment seeking behavior.

[3] We only report highly significant features.

treatment seeker, even those having one more day of a subset of these 5 symptoms increases the probability of being a fast treatment seeker, this is probably because poor patients are reluctant to seek treatment until they have more symptoms.

Symptom(s)	Coefficient	Treatment seeking
1 Symptom		
Cough	-4.1E-2	Slow
Bloody cough	4.1E-2	Fast
Fever	-2.4E-2	Slow
Night sweats	-1.4E-2	Slow
Constant tiredness	-2.6E-2	Slow
Appetite loss	2.5E-2	Fast
Combination of 2 Symptoms		
Cough, Bloody cough	-1.2E-3	Slow
Combination of 4 Symptoms		
Cough, Fever, Weight loss, Constant tiredness	4.0E-7	Fast
Combination of 5 Symptoms		
Cough, Fever, Weight loss, Constant tiredness, Appetite loss	-1.5E-8	Slow

Table 1

Conclusions and Future Work

We are able to isolate several symptoms/combinations of symptoms that encourage or discourage patients from seeking care quickly using logistic regression; in general, it seems that we would recommend health education groups to raise awareness that a cough, fever, night sweats, or tiredness in isolation can be symptoms of TB. Additional financial or transportation support might be necessary for patients who have five or more symptoms, as we suspect that there may be unobserved effects in that group. In addition, it seems that the types of patients can be separated using a hyperplane using some permutation of the feature space (as seen from our results from PCA and SVD), although finding actionable results from this separation is difficult.

In the future we would like to include demographic variables, such as income, family characteristics (such as number of dependents, etc.), or ease of access to treatment facilities in our analysis to better understand the underlying patterns in our data. Presumably these features can help us separate the fast/slow treatment seekers better when they have similar characteristics, i.e., durations of symptoms. In addition, these features would also help target individuals with different socioeconomic background when making TB education outreach.

Acknowledgement: We thank Dr. Goldhaber-Fiebert for providing us the data and Dr. Babiarz for previously cleaning and processing the raw data.