James Abundis
jlabundi@calpoly.edu
Thomas Fahrner
tlfahrner@calpoly.edu
5/14/18

Project Proposal - Turkey Political Opinion

Introduction:

One of the main topics we have covered so far in CSC 466 is the applied concept of clustering based on a calculated similarity between groups. This clustering can be used to "discover" groups that we may able to identify as a result of clustering, or can be used to confirm/disconfirm our previous beliefs about what observations should be grouped together. One of the recent datasets on Kaggle gives us survey data for citizens of Turkey along with useful demographic data that could help us identify different political groups. We find this to be a good dataset for our KDD project it sets up for a discovery project to which we would cluster survey participants by their metadata and use this to classify their political affiliation.

Dataset: https://www.kaggle.com/yemregundogmus/turkey-political-opinions

Metadata Features:

Features :

- Cinsiyet : Sex Feature
- Yas : Age Feature
- Bolge : Areas inhabited in Turkey
- Egitim : Education Level
- (Soru = Question) (Questions include Turkey)
- Soru1/Question1: Do you think our Economic Status is good?
- Soru2/Question2: Need Reform in Education?
- Soru3/Question3: Resolve Privatization Are You?
- Soru4/Question4: Should the state use a penalty like death penalty for certain crimes?
- Soru5/Question5: Do you find our journalists neutral enough?
- Soru6/Question6: From 22:00 am Then Are You Supporting the Prohibition to Buy Drinks?
- Soru7/Question7: Do You Want to Live in a Secular State?
- Soru8/Question8: Are you supporting the abortion ban?
- Soru9/Question9: Do you think that the extraordinary state (Ohal) restricts Freedoms?
- Soru10/Question10: Would you like a new part of the parliament to enter?
- Parti : Political View

Models Proposed:

The first model we would like to make is a C4.5 decision tree that would be used to classify an individual's party affiliation. This C4.5 tree implementation will need to support both categorical inputs (survey data) and numerical data (demographic metadata).

The second model that we would explore is using different clustering algorithms such as agglomerate hierarchical clustering, density based clustering, and k-means clustering. One of the challenges we expect to face is that a lot of the survey data consists of "Yes/No" binary data. In order to cluster, we may have to use just the demographic data alone or find some way to convert the survey data to a numeric scale that make sense.

Validation of Models:

Validation of the models above can be achieved by splitting our data set into train and test sets with methods learned from class: all but one, randomly selecting hold out sets, and cross validation. This will allow us to calculate metrics for each model that shows how well they identify/classify correctly political affiliation and then compare accuracy metrics to conclude which model did better and possibly explain why.

Posting our Findings:

Our final report will include visualizations for all models used. For the C 4.5 classifaction model, we will include pictures of the tree created from the data that was used to classify political affiliation/group. For the clustering algorithms we will include labeled group scatter plots to see what political groups we identified. Furthermore, it would be useful to see for the clustering algorithms, how well the clusters line up with the number of major political parties in Turkey. This will require some further research into Turkey's political parties, but it is important that we put our findings in context.

Key Questions To Explore:

1. Do the clustering used algorithms correctly identify political affiliations in Turkey? If not how do they differ?
2. Looking at the clusters produced, what does it tell us about the poltical spectrum of Turkish citizens? For example: A cluster for very conservative, very liberal, moderate, etc.
3. If we can identify political affiliations/parties accurately, can we label data points with classes and use this to classify other Turkish citizens in our test set with our C4.5 algorithm?