CSCI620.1 Group2 Data Mining Proposal

Shuo Yao
Rochester Institute of Technology
sy9663@rit.edu

Yi Lan
Rochester Institute of Technology
yl8217@rit.edu

Haoran Sheng
Rochester Institute of Technology
hs1911@rit.edu

Jack Smith
Rochester Institute of Technology
jss7268@rit.edu

We have researched through available resources and selected two data sets that could be used for our data mining component: `Yelp Data Set` from our data management component and `Census Income Data Set` from UCI Machine Learning Data Set Repository.

1. YELP DATA SET

The `Yelp Data Set` is a subset of `Yelp's businesses`, `reviews`, and `user` data. Detailed information won't be presented here since we have worked on this data set for our data management component. There are several data mining directions on this data set, these are:

1. Restaurant's star rating might be affect by its attributes such as: `Restaurants Reservations`, `Restaurants Table Service`, `Outdoor Seating`, `Has TV`, `WiFi`, `Alcohol`, `Restaurants Delivery`, `Good For Kids`, `Business Accepts Credit Cards`, `Business Parking` and `Number of Reviews`. Those attributes could be used as features that predict a restaurant's star rating. One possible model that could be used for this problem is a multi-valued and multi-labeled decision tree.

2. Sentiment analysis on user's review content also sounds interesting. Users' attitude (positive and negative) might be quantified as the star rating. This gives us the possibility of doing sentiment analysis on the review content (star rating can converted to 'positive' and 'negative' through a threshold, makes the sentiment analysis a supervised learning). Possible classifiers for this problem could be Naive Bayes and SVM. In addition, it could also be modeled as a clustering problem such as k-means.

3. Prediction of a user's review rating of a business. Based on either similar users opinions of the business or the user's reviews of similar business. This would require quantifying similarity between users and businesses. Possible attributes for quantifying a business include the specific attributes of the business, the type of business, check-in frequency, and rating). Possible attributes for quantifying similar users would be frequency of review for certain business types, review rating based on the attributes of a business, and businesses/chains in common with similar review levels. Being able to determine similarity between businesses and between users could allow for accurate prediction of a user's rating by selecting their rating for a similar business, or the rating by a similar user.

2. CENSUS INCOME DATA SET

The `Census Income Data Set` from UCI Machine Learning Data Set Repository contains 48,842 instances of `Adult` information containing:

1.`Income` information: >50K, <=50K.

2.`age` information: continuous.

3.`workclass` information: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

4.`fnlwgt` (final weight, the weights on the Current Population Survey (CPS)) information: continuous.

5.`education` information: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

6.`marital-status` information: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

7.`occupation` information: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners,

8.`Machine-op-inspct` information, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

9.`relationship` information: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

10.`race` information: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

11.`sex` information: Female, Male.

12.`capital-gain` information for money gained from investigation: continuous.

13.`capital-loss` information for money lost from investigation: continuous.

14.`hours-per-week` information for each adult's average weekly work time: continuous.

15.`native-country` information for each adult's native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, etc.

One possible data mining direction on this data set is: an adult's income may be affected by his/her `age`, `workclass`, `fnlwgt`, `education`, `marital-status`, `occupation`, `Machine-op-inspct`, `relationship`, `race`, `sex`, `capital-gain`, `capital-loss`, `hours-per-week`, `native-country`. Those could be used as features in the classification model. If we ran continuous features through certain thresholds, those features will become discrete. One possible classification model could be used for this problem is a multi-valued and multi-labeled decision tree.

Another possible data mining direction would be to investigate whether a person is being compensated fairly. Taking into account age, education, occupation, race and sex and comparing income could determine if biases against certain races or sexes in different occupations. With this data it could be possible to estimate the disparity that certain groups have on average as well as specifically for their positions.