CS795: Data Mining and Security

Summer 2011

Project

This project analyzes the data provided by the Entrée Recommender System developed by Dr. Robin Burke. The data is made available in the project directory as entrée_data.tar.gz.tar. We are specifically interested in the data provided under the data subdirectory.

It has data from restaurants in
8 large cities:

1. Atlanta
2. Boston
3. Chicago
4. Los Angeles
5. New Orleans
6. New York
7. San Francisco
8. Washington DC

```
Each entry in the data file has the following format:
    Restaurant Id [tab]
    Restaurant Name [tab]
    Restaurant Features (3 digits ids separated by spaces)
```

The mapping of three digit IDs to feature description is provided in the features file. In general, features should must fall into one or more of the following categories: Cuisine (Indian, Ethiopian, European, etc.), Price (below $15, $15-$30, over $50), Style (Authentic, Cajun, Creole, etc.), Atmosphere (Creative, Poor décor, Eclectic, Fabulous views, etc.), and occasion (After hour dining, early dining, parties and occasions, etc.). However, there are many other features which don't exactly seem to fit into any of these (e.g., Parking/Valet, Excellent service, etc.). There could also be some noise in the data. In addition, you are free to reduce the categories (all French to be treated as one even though there are coded with different numbers, etc.). Use your judgment and make any modifications in the categories. You DO NOT NEED to SEEK instructor's permission for every change. Use your own judgment. But you MUST record all changes and report them in the project report.

What is to be done?

(i)   Study the provided features and classify them into one of the standard (cuisine, style, price, atmosphere, and occasion) or into your own created additional categories. Limit the new categories to at most 5. If you think that a feature fits into more than one of your categories, do put them in all the categories that they fit in. This could be more an exception than a role. Typically, there should be a 1-1 mapping of features to categories.

(ii)   Analyze and form rules to characterize the following five types of cuisine: (i) Indian (ii) Mexican (iii) Italian (iv) French (v) American. For each category, derive a set of rules based on data available from all cities.

(iii)   Derive association rules among the given features. In other words, does Creative atmosphere imply a specific category? In particular, experiment with the following associations:
   a. Cuisine and atmosphere
   b. Price and atmosphere
   c. Price and style
   d. Cuisine and occasion.
   e. Décor and Price

(iv)   Let us now concentrate specifically on the quality of the food. This is specified through features 73-78. Assuming that the outcome you are interested is one of the following categories, determine **if there is any relationship between type of cuisine and the quality indicator**. Categories to be considered are: Fair, Good, Excellent. You can combine the given 6 categories into these 3 categories.

(v)   Find an association between a restaurant offering vegetarian (243) to its price and cuisine.

(vi)   Determine the error that would be incurred by categorizing the restaurants based on the continents they represent: Asia, Europe, Africa, North America, and South America. For each continent, form rules to determine the outcome (which continent they come from) based on other attributes such as price, atmosphere, quality of food, etc.

(vii)   Selecting the "Europe" continent, state rules to determine the city based on the given features. (Choose your own features or categories that are most relevant).

(viii) Finally, given that a user likes a specific continent food and provides the feature set (you can limit these if you like), you should have rules to determine which restaurant they should visit for each city.

The outcome of the project is a final report and the code that you used. If you used Weka, say so.

Clearly state all assumptions you have made in answering each question.