**Project 5 – Advanced Data Mining Applications**

**CS548 / BCB503 / CS583 Knowledge Discovery and Data Mining - Fall 2018**

**Prof. Carolina Ruiz and TA Huimin Ren**

**Student: Paritosh Goel**

|  |  |
| --- | --- |
| Description of the particular problem within the selected data mining topic to be addressed in this project | /15 |
| Description of the approach used in this project to tackle the above problem.  *All data mining techniques you use in this project for pre-processing, mining and evaluation must have been covered in class during this semester* | /25 |
| Description of the dataset selected | /15 |
| Appropriateness of the dataset selected with respect to this topic/problem | /10 |
| Guiding questions | /10 |
| Preprocessing | /10 |
| **Experiments:**   * Sufficient & coherent (most experiments must be performed in Python) | /25 |
| * Objectives, Data, Additional Pre/Post-processing | /20 |
| * Presentation of results | /20 |
| * Analysis of results | /30 |
| Overall discussion, comparisons, and conclusions | /20 |
| TOTAL | /200 |

Total Written Report: \_\_\_\_\_\_\_\_\_\_\_\_\_\_/200 = \_\_\_\_\_\_\_\_\_\_\_/100

Class Presentation: \_\_\_\_\_\_\_\_\_\_\_/100

Class participation during project presentation: \_\_\_\_\_\_\_\_\_\_\_/100

*Do not exceed the given page limits for this written report*

**Topic: Anomaly Detection**

1. **Description of the particular problem within the selected data mining topic to be addressed in this project:**

People often wants to explore new unique experiences to dine. With a list of restaurants in a city and their features, I try to find out unique restaurants which can work as recommendation for people.

1. **Description of the approach used in this project to tackle the above problem:**

I used Anomaly/Outlier Detection technique to find out unique restaurants in a city. The anomaly detection is done on the basis of features of restaurants. The more unique/different the features will be, the restaurant will be considered as unique.

1. **Dataset Name: Entree Chicago Recommendation Data**
2. **Where found:** http://kdd.ics.uci.edu/databases/entree/entree.html
3. **Dataset Description:**

This data contains a record of user interactions with the [Entree Chicago](http://infolab.ils.nwu.edu/entree/) restaurant recommendation system. The dataset comprises list of restaurant with their features such as cuisines, music, on\_the\_beach, cafe, bar etc.

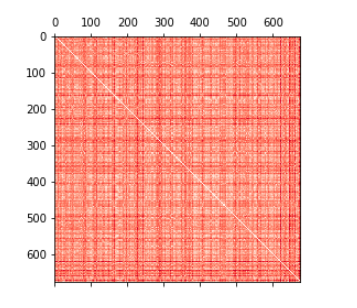
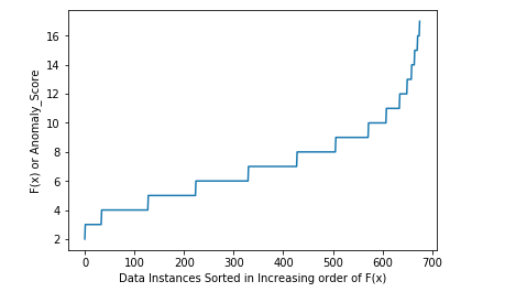
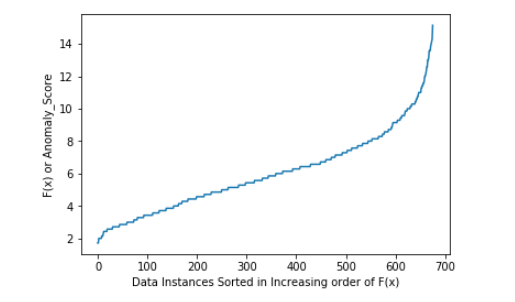
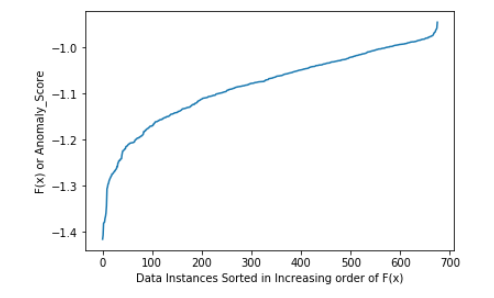
1. **Initial data preprocessing, if any:**

The dataset had the restaurants list with a list of features in a file. There is a separate file representing the meaning of the features. The initial preprocessing done is joining this two files and create a single source of data with features added as columns with restaurant. Finally I got 257 columns with 1st column as restaurant name and rest as features.

1. **Three Guiding Questions about the dataset domain:**
2. What are the most unique types of restaurants in the chicago city which are different from the rest.
3. How many are such restaurants and what are the features of them.
4. Apart from cuisine, what other factors can make a restaurant unique in chicago

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Summary of Experiments.** *At most 2 page.* | | | | | | |
| **Tool** | **Pre-process** | **Mining**  **Technique** | **Results** | **Time**  **taken** | **Evaluation** | **Observations about experiment**  **Observations about visualization**  **Interpretation results** |
| Python | 1. Created a single source of information from two different csv  2.dropped unnecessary columns | Unsupervised  Proximity matrix | Found some data with very low proximities | 70 mins | The matrix with red lines are outliers | In case of multi-D data , proximity matrix can serve as initial way to check for the anomalies. The rows with low proximities can be anomalies.  Note: For proximity, distance considered is euclidean distance between the features of the restaurants |
| Python | Same as above | Unsupervised  Proximity based, K nearest value | Plotted Anomaly Score with sorted data instances, and found an elbow at top 25 restaurants | 1 sec | Found an elbow at top 25, which can be possibly outliers | I did experiments for k-nearest distance with different values of k from 10 to 30. Plotted the data instances with anomaly\_score(which is k nearest distance. )In every case, the elbow was at around top 25 restaurants.  The elbow was not changing for different values of k, top 25 restaurants in the plot could be potentially outliers.  Note: The distance function is same as above |
| Python | Same as above | Unsupervised  Proximity based, Avg of K nearest | Got almost same results as above | 1 sec | Found an elbow at top 25, which can be possibly outliers | In this experiment, I considered, Average of K nearest neighbor and almost got the same result as above.  The results were analyzed after looking at the plot.  Note: Distance function is same as above |
| Python | Same as above | Unsupervised  Density Based Approaches - Local Outlier Factor | Got very low potential outliers (~5) | 1 sec | Found elbow at very last of the data sorted | In this experiment, I used Local outlier Factor as the basis for Anomaly Detection.  The distance function used is default chosen by the sklearn  The Outlier score produced is plotted with sorted data instances, and it is seen that the elbow is found at the very last of the x axis (sorted data instances). |
| Python | Removed the cuisine features from the data set | Same as above | Same as above | 1 sec | Same as above | Very loess outliers were found after removing the cuisine features from LOF Approach |
| Python | Same as above | Used Neigherest Neighbour | Same as above | 1 sec | Same as above | Same as above |

**Analysis of Results: (at most 1 page)** 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the Domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include and explain (some of) the best / most interesting results you obtained in your experiments. 4. Include visualizations.



Local Outlier Factor Average(Knearest) KNearest Proximity Matrix between all data points

The above first 3 diagrams are Plot of Anomaly Score vs Data points and the 4th diagram is the proximity matrix of data points.

It was noticed that when we change the value of K in Knearest the plot of the Anomaly Score vs data points was changing initially but it remained constant while K from 10 to 30.

The dark red horizontal lines in proximity matrix represent the restaurants which are at maximum distance from other restaurants (potential outliers).

From the K nearest and Average of K Nearest, I captured the restaurants above the elbow.

**Guiding Questions** - **What are the most unique types of restaurants in the chicago city which are different from the rest.**

From the results above we can say that top 25 restaurants in the plot are different from the rest. Restaurant with top anomaly score was - **Ed Debevic's Short Orders Deluxe.** **How many are such restaurants and what are the features of them** - From the elow, we can say that around 25 restaurants are potential outliers. When I checked the features of the above mentioned restaurant some of them were - Weekend Brunch, After Hours Dining, Good for Younger Kids, Excellent Decor etc. **Apart from cuisine, what other factors can make a restaurant unique in chicago -** All types of features mentioned above

**Summary of what you learned in this project:**

I learned that we can also plot the proximity matrix for high dimension data to find out potential outliers. Also learned the puthon sk learn package to find outliers. Created my own code to find the knearest also.

When I plot the Anomaly Score with Datasets in my dataset, The average of KNearest Distance gave good results. There was a point when the plot wasn’t changing on changing the value of K (from 10 to 30)

Overall I learned How to use Anomaly Detection techniques in python.