New York City Taxi Trip Duration

Siyi Yu[1]

[1] Jilin University, China

Introduction

In this competition, Kaggle is challenging us to build a model that predicts the total ride duration of taxi trips in New York City.This is a prediciton problem.The raw datasets mainly two files,whose attributes are shown below.

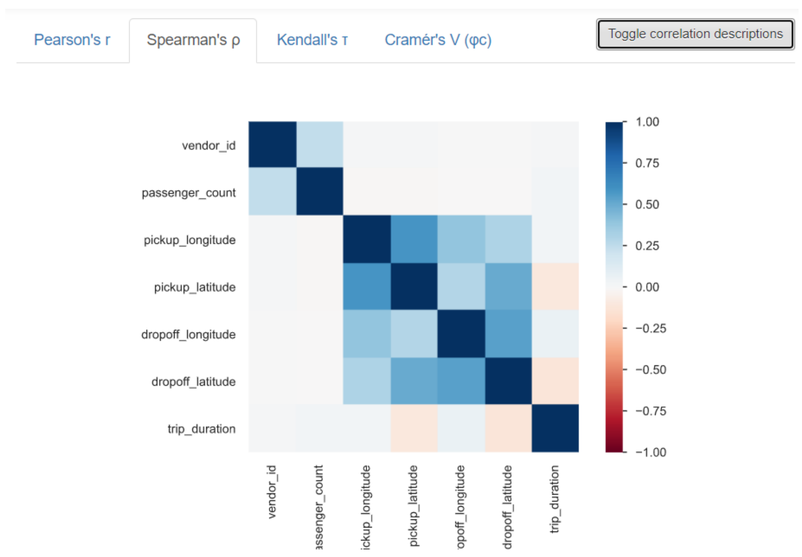
- *train.csv:*
  vendor_id,pickup_datetime,dropoff_datetime,passenger_count
  pickup_longitude,pickup_latitude,dropoff_longitude,dropoff_latitude
  store_and_fwd_flag,trip_duration
- *test.csv:*
  vendor_id,pickup_datetime,passenger_count
  pickup_longitude,pickup_latitude,dropoff_longitude,dropoff_latitude
  store_and_fwd_flag

Data Processing

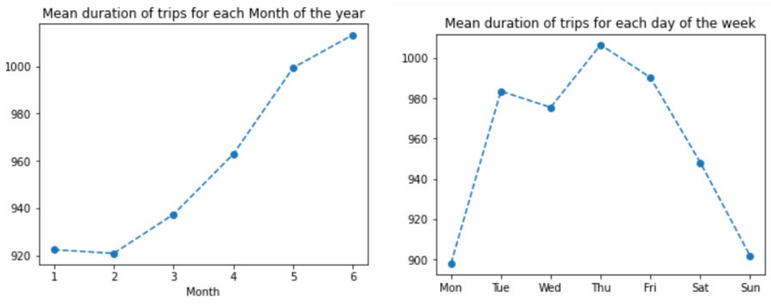
- Remove missing value and NaN value
- Filter outliers and duplicate data
- Process pickup/dropoff_datetime
- Process pick/dropoff_latitude/_longitude
- Process string by one-hot encoding

Data Visualization

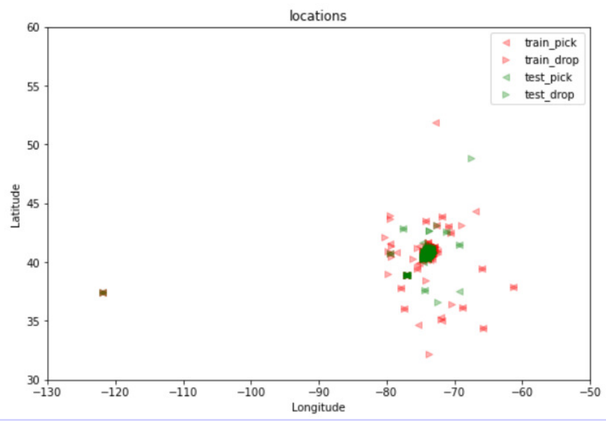
- Spearman Correlation of attributes



- Duration of taxi trips regard to month,weekday



- Pick/Drop locations of taxi trips



GOAM Algorithm

Second, based on the *earth move distance*, we calculate the outlying degree.

Missing figure Testing figcolor

where G_q is the query group, n is the number of compare groups, and h_{k_s} is the histogram representation of G_k in the subspace s.

Outlying Aspects Identification In this step, based on the value of outlying degree we will identify the group outlying aspects. If a feature's outlying degree is greater than a threshold, the more likely the feature is group outlying aspect. When the dimensionality of features is high, we adopt a stage-wise candidate subspace construction strategy to alleviate the exponential explosion.

Experiment

Synthetic Dataset contains 10 groups and 8 features. Each group consists of 10 members, and each member has 8 features.

Method	Truth Outlying Aspects	Identified Aspects	Accuracy
GOAM	$\{F_1\}$, $\{F_2F_4\}$	$\{F_1\}$, $\{F_2F_4\}$	100%
Arithmetic Mean based OAM	$\{F_1\}$, $\{F_2F_4\}$	$\{F_4\}$, $\{F_2\}$	0%
Median based OAM	$\{F_1\}$, $\{F_2F_4\}$	$\{F_2\}$, $\{F_4\}$	0%

It can be observed that the GOAM method can identify the trivial outlying features and non-trivial outlying subspaces correctly and is obvious from the table that the accuracy of GOAM is the best, which is (100%).

NBA Dataset was collected from Yahoo Sports website (`http://sports.yahoo.com.cn/nba`). The data include all teams from the six divisions, and each player in the team has 12 features.

Teams	Trivial Outlying Aspects	NonTrivial Outlying Aspects
Cleveland Cavaliers	{3FA}	{FGA, FT%}, {FGA, FG%}
Orlando Magic	{Stl}	None
Milwaukee Bucks	{To}, {FTA}	{FGA, FTA}, {3FA, FTA}
New Orleans Pelicans	{FT%}, {FTA}	{FTA, Stl}, {FTA, To}

Missing figure Testing figcolor

New Orleans Pelicans on FT%

Missing figure Testing figcolor

New Orleans Pelicans on FTA

`New Orleans Pelicans` has more players with lower {free throw percentage}, {free throws attempted}.

Conclusion

Problem Definition Formalize the problem of Group Outlying Aspects Mining by extending outlying aspects mining.

GOAM algorithm Propose GOAM algorithm to solve the *Group Outlying Aspects Mining* problem.

Strategies Utilize the pruning strategies to reduce time complexity.

Acknowledgement

- International Cooperation Project (Y7Z0511101) of IIE, Chinese Academy of Sciences