Predicting Destination Of Taxi Rides

CS4099 Project

End Semester Evaluation

Arpit Augustine, GVS Rahul, Harsh Parsuram Puria, Vineel Patnana
Guided By: Dr.K A Abdul Nazeer, Mr. Ibrahim Abdul Majeed

May 10, 2016

Outline

Introduction

- On-demand public transport solutions.
- Human mobility behavior.
- Shift to unicast based messages.
- Improves efficiency of Taxi dispatch systems.

Problem Statement

- To build a predictive framework that is able to infer the final destination of taxi rides based on their initial partial trajectories :
  - Our approach was to build a probabilistic model by learning the behavior of taxi.
  - Before building the model, we have to find the support points using a proper clustering algorithm.
  - Destination depends on various factors like working day/holidays, Time, passenger etc.
  - So given a sequence of GPS coordinates, we need to predict the destination GPS coordinate.

Literature survey

Name and Year	Clustering	Model Used	Results
J.A Alvarez-Gracia et al and 2010	Not mentioned	Hidden Markov Model	-
Wesley Mathew et al and 2012	Hierarchical Triangular Mesh	Hidden Markov Model	13.85%
Sbastien Gambs et al and 2012	DJ Clustering	Extended Mobility Markov Model	70% to 95%

- Vikas Thada, Dr.Vivek Jaglan, *Comparison of Jaccard, Dice, Cosine Similarity Coefficient.*

Statistics

Total number of trips	17,10,670
Number of trips with NULL Polyline	43,904
Number of co-ordinates	7,83,63,691
Resultant number of trips	16,66,766
Number of taxi stands	64
Number of different passengers	57,105
Number of trips with no missing values	10

Table: Statistics

Call_Type	Number of trips
A	3,64,770
B	8,17,881
C	5,28,019

Table: Number of trips corresponding to each Call_Type

Day_Type	Number of trips
A	17,10,670
B	0
C	0

Table: Number of trips corresponding to each Day_Type

Design

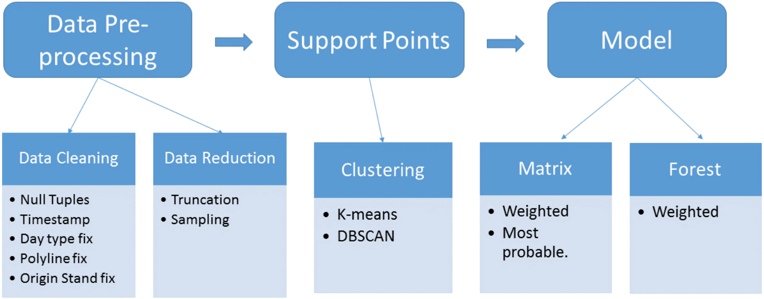

Figure: Our Design

Data Pre-processing

Data Cleaning

- Exclusion of tuples.
- Timestamp conversion.
- Daytype fixing.

Day_Type	Number of trips
A	11,02,229
B	41,704
C	41,336
D	4,81,497

Table: Number of trips corresponding to Day_Type after fixing

- Polyline fixing
- Origin stand fixing

Data Pre-processing

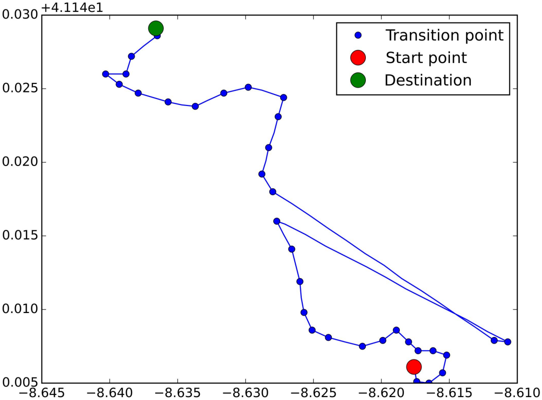

Figure: Noise in polyline

Data Pre-processing

Data Reduction

- Truncation
- Sampling

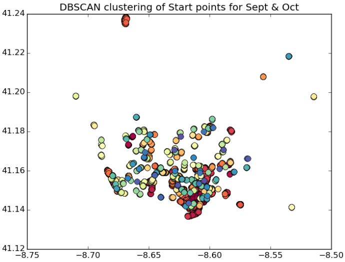

Figure: DBSCAN clustering of starting points for September and October

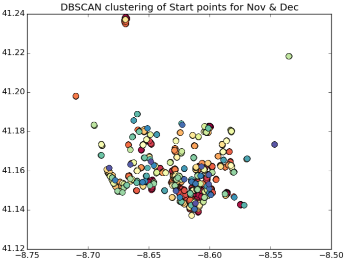

Figure: DBSCAN clustering of starting points for November and December

Support Points

Clustering

- k-means
- DBSCAN

Testing and Evaluation

Data of 60 days trips are trained and tested on next 7 days. This is repeated 10 times by moving the range by 7 days.

Evaluation Metric : Mean Haversine Distance

$$a = \sin^2(\frac{\theta_2 - \theta_1}{2}) + \cos(\theta_1)\cos(\theta_2)\sin^2(\frac{\phi_2 - \phi_1}{2}) \tag{1}$$

$$d = 2.r.a.\tan(\sqrt{\frac{a}{1-a}}) \tag{2}$$

θ is the latitude, Φ is the longitude, d is the distance between two points, r is the Earth's radius.

Model

Matrix

- Weighted
- Most Probable

Forest

- Weighted

Primitive Model

Matrix

- Destination $=$ arg max F(Start)
- Destination $=$ weighted (Mean) F(Start)
- Destination $=$ arg max F(Start,transition)
- Destination $=$ weighted (Mean) F(Start,transition)

Model

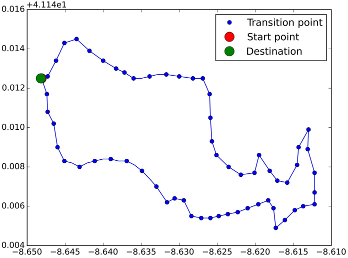

Figure: Round Trip

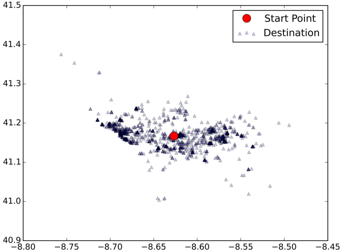

Figure: A Start point and its Destinations

Model

Forest

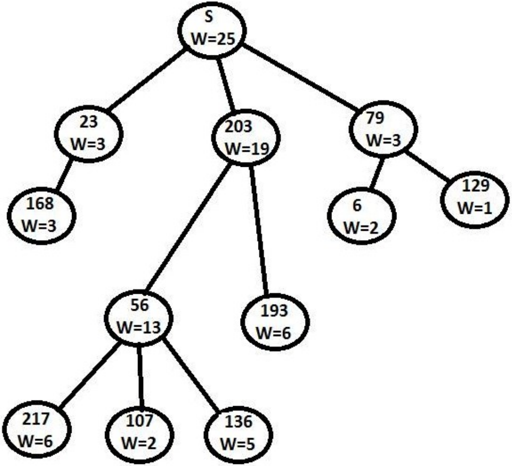

Figure: Model Forest

Improvements

- Call Type
- Day Type
- Time
- Taxi Grouping

Results

Start	Transition	End	Probability Type	Mean Distance (Kms)
k-means	-	k-means	Maximum	4.95
k-means	-	k-means	Weighted	3.14
DBSCAN	-	DBSCAN	Maximum	3.13
DBSCAN	-	DBSCAN	Weighted	3.04
k-means	k-means	k-means	Maximum	2.84
k-means	k-means	k-means	Weighted	2.61
DBSCAN	k-means	DBSCAN	Maximum	3.24
DBSCAN	k-means	DBSCAN	Weighted	3.16

Table: Models

Results

Type	Weighted Probable	GOOD-AVG-BAD (in %)
Call_Type A	1.954	65.7 - 27.3 - 7.0
Call_Type B	1.938	68.3 - 24.4 - 7.3
Call_Type C	2.920	53.4 - 32.7 - 13.9
Call_Type C & Day_Type A	3.142	51.1 - 33.0 - 16. 0
Call_Type C & Day_Type D	2.782	54.1 - 32.5 - 13.4

Table: Results after segregation

Results

Type	Weighted Probable	GOOD-AVG-BAD (in %)
Call_Type A	2.274	64.3 - 21.4 - 14.2
Call_Type B	2.426	64.2 - 20.9 - 14.9
Call_Type C	3.582	52.0 - 24.4 - 23.6
Call_Type C & Day_Type A	3.727	51.6 - 23.0 - 21.3
Call_Type C & Day_Type D	3.358	52.5 - 26.0 - 21.6

Table: Results for Model forest

Results

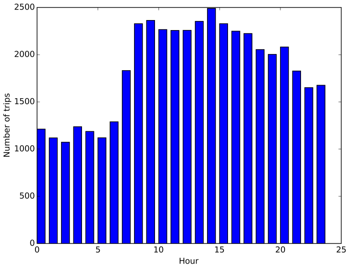

Figure: Hours vs Number of Trips

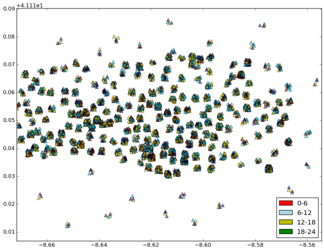

Figure: Clusters based on time range

Conclusion

Improve upon current model using following methods:

- Features like which day the ride was taken and how the taxi was booked by the customer had a major role in the prediction.
- features like individual taxis and individual customers may have a particular pattern.
- Further improvements - time, taxi grouping, city specific

References

[1] Alvarez-Garcia, Juan Antonio, et al. *Trip destination prediction based on past GPS log using a hidden markov model*, Expert Systems with Applications 37.12 (2010): 8166-8171.

[2] Mathew, Wesley, Ruben Raposo, and Bruno Martins. *Predicting future locations with hidden Markov models*, Proceedings of the 2012 ACM conference on ubiquitous computing. ACM, 2012.

[3] Gambs, Sbastien, Marc-Olivier Killijian, and Miguel Nez del Prado Cortez. *Next place prediction using mobility markov chains*, Proceedings of the First Workshop on Measurement, Privacy, and Mobility. ACM, 2012.

[4] Thada, Vikas, and Vivek Jaglan. *Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm*, International Journal of Innovations in Engineering and Technology (2013).