Evaluating classifiers to determine user-preferred stops in a personalized recommender system

A.A. Borodinov, V.V. Myasnikov

САМАРСКИЙ УНИВЕРСИТЕТ
SAMARA UNIVERSITY

Recommender systems major problems

- The cold start.
- The receiving information method from the user is not formalized.
- Individual characteristics such as personal income, age, gender, family size, access to public transport influence the choice of the route even for the same purpose of the trip.
- User preferences change over time. In addition, context influences user selection.
- Typical existing solutions mainly use the Bayesian approach with a sequential parameter recalculation scheme.
- It is possible to use transfer learning to improve recommendations.
- The problem of determining traffic flow on the vehicle route.

Data

For the mobile service "Pribyvalka-63" data for analysis are presented as follows:

- Public transport stop information (identifiers and coordinates).
- Public transport route information (identifiers and stop identifier list).
- Information about the vehicle (identifiers), location coordinates (the vehicle transmits its coordinates two times per minute), route identifier.
- Coordinates of users and request parameters.

The data set contains information about the request time and GPS coordinates of the mobile device during the request. We further divided the timestamp on the time of day in seconds and on the day of the week, which allowed to determine the weekends and weekdays. Requests were recorded for four months. The mobile application recorded 18441744 requests from 116524 users to 1479 stops.

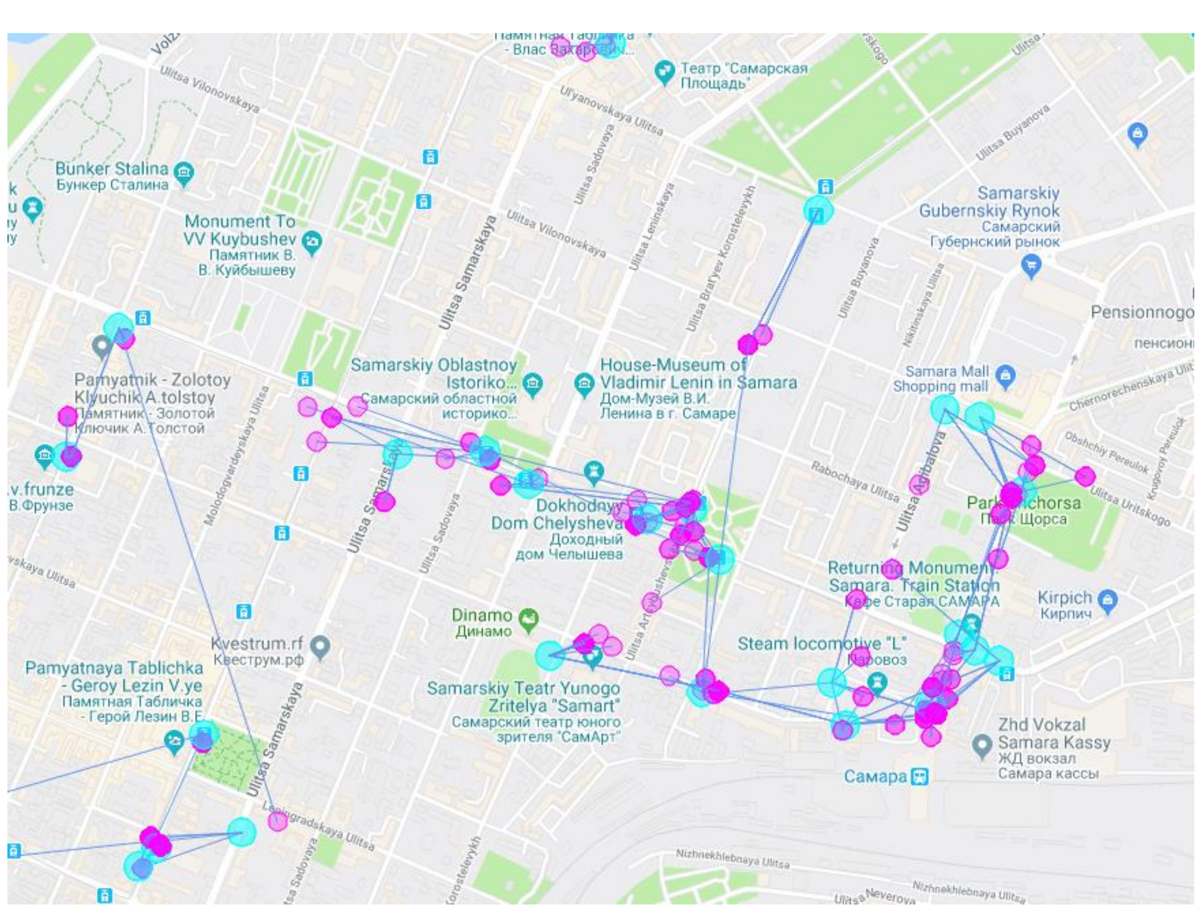

Figure 1. Visualization of user requests on google maps.

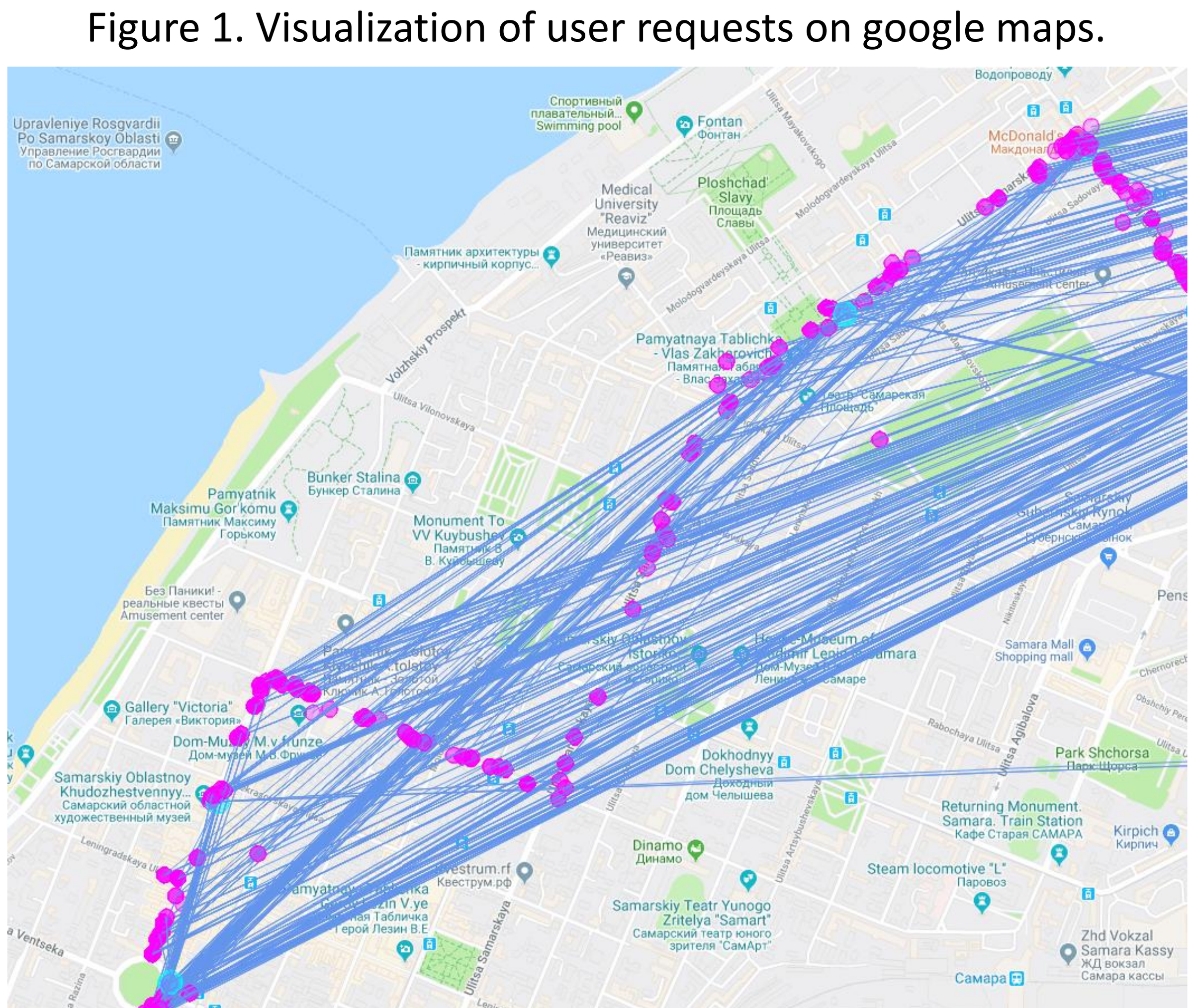

Figure 2. An example of non-standard user behavior.

Results

For the experiments, we randomly selected data from 300 users with the number of requests from 3083 to 63. About half of the selected users had an average number of requests and amounted to about 160 requests. Then we divided the data on the requests of each user for the training set and test set in the ratio 4:1. To obtain a reliable estimate of model performance, we use a five-fold cross-validation approach.

Table 1. Evaluation of the accuracy and performance of various machine learning methods.

№ Algorithm \ User		1	2	3	4	Avg. of all users
Estimation algorithm	Accuracy	56.191	14.287	40.000	38.489	20.947
	F1 score	0.441	0.048	0.232	0.211	0.115
Min distance	Accuracy	10.219	4.762	10.000	16.547	7.847
	F1 score	0.031	0.017	0.033	0.0064	0.024
SVM	Accuracy	83.741	21.161	58.000	35.437	50.175
	F1 score	0.518	0.121	0.289	0.078	0.273
Decision Tree	Accuracy	89.496	36.645	64.000	46.312	60.128
	F1 score	0.484	0.213	0.385	0.256	0.347
Random Forest	Accuracy	71.429	**59.311**	**84.000**	**67.376**	**74.002**
	F1 score	0.687	**0.579**	**0.817**	**0.671**	**0.524**
AdaBoost	Accuracy	**93.237**	43.226	78.000	**69.125**	**73.953**
	F1 score	**0.737**	0.265	0.507	**0.412**	**0.508**
kNN, k=2	Accuracy	67.619	35.172	78.000	27.148	55.868
	F1 score	0.663	0.345	0.763	0.289	0.426
kNN, k=3	Accuracy	68.571	41.379	**84.000**	31.711	55.923
	F1 score	0.672	0.413	**0.814**	0.326	0.428
kNN, k=5	Accuracy	65.714	41.379	70.000	38.479	55.966
	F1 score	0.606	0.863	0.675	0.361	0.429
MLP	Accuracy	50.863	5.419	36.000	42.509	32.591
	F1 score	0.031	0.001	0.112	0.025	0.054

Table 2. The performance of the evaluated machine learning methods, in s.

SVM	Decision Tree	Random Forest	AdaBoost	kNN, k=2	kNN, k=3	kNN, k=5	MLP
0.203	0.001	0.002	0.019	0.001	0.001	0.001	0.001

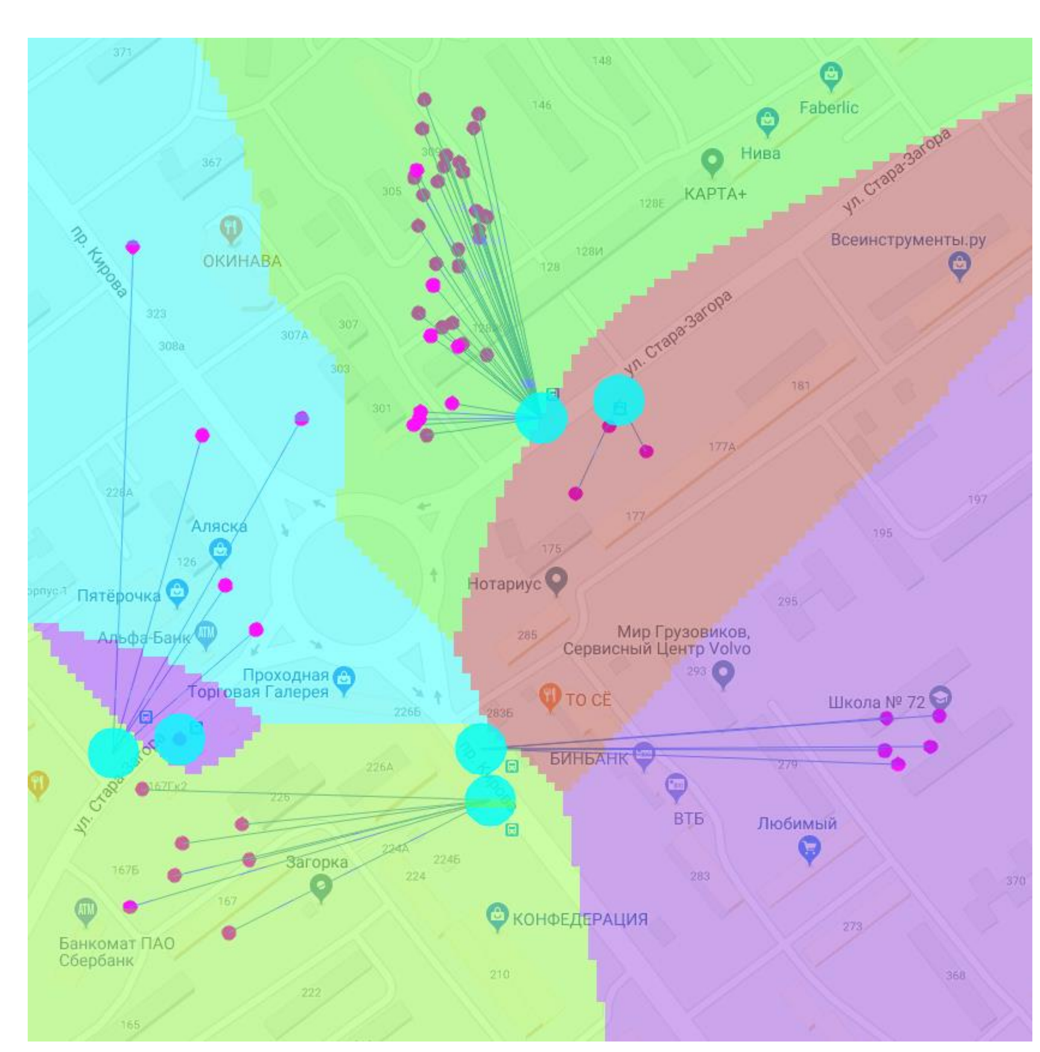

Figure 3. Preferred stops map depending on user location.

Acknowledgments

The work was funded by the Ministry of Science and Higher Education of the Russian Federation (unique project identifier RFMEFI57518X0177).

 ALEKSANDR BORODINOV– aaborodinov@yandex.ru VLADISLAV MYASNIKOV – vmyas@geosamara.ru