Novelty Detection on location prediction problems using gradient boosting framework

Abstract:

The enormous use of mobile applications and the wide spread of use location based services, as Gowalla, Foursquare, google maps, make it a must to discover these extracted data, which is the check-ins of the users, in order to maintain a better service and offer new services, as predicting the new places to be visited, predicting the traffic jammed areas, so we conducted this experiment to classify the check-ins to be visited either Novel or regular points of interest (POI), by extracting the main features of the current check-ins, using the features used by Defu et al.[1] and others, we can predict the novelty or the next Point of Interest (POI) up to 82% overall accuracy, thanks to the added features and the using of gradient boosting framework (XGBoost).

Introduction:

The emerging use of mobile technology with location features, has been widely occurred and location services have attracted millions of users, such as Gowalla, Foursquare, and adding it to other common social platforms like Facebook and twitter, these companies needed to enhance their services with applied analysis over these enriched big data, the researchers used such datasets to solve punch of problems, like analyzing human mobility patterns, travel patterns, social ties and its influence to POI, traffic predictions, and next check-in prediction and its enhancement to advertising and POI recommendations, as we go through the analysis of we found the novelty prediction is one more type of analysis, and it can play an important role in advertising and POI recommendations, it can be applied if a user is expected to visit a novel POI the system should be able to list a number of recommendations based on the users friends or the user’s historical interests, but this is the limit of our analysis, and in future work adding the POI semantic predictions would enhance the results and the application more importantly, but this feature is out of scope as the used dataset (Gowalla) has no semantics of the check-ins, for these experiments we focused on extracting features from the data to be able to predict the novelty of the next POI, using the temporal, spatial, and historical features, then we applied a classifier to classify novelty or regularity of the next POI based on these features, the problem turned to be a binary classification problem, this algorithm is a gradient boosting framework, which has a good fame at Kaggle’s competitions, called extreme gradient boosting (XGBoost), it uses decision trees, to solve these problems with high accuracy and speed. The next sections are ordered as related work, model features, proposed model, results and conclusion.

Related work:

Location based predictions is becoming at the focus of many research projects, as it has great impact on advertising, traffic congestion control and much more fields, these analyses consist mainly of three types. First, predictions using GPS trajectories for the places where the user spent more time, figured out by checking the time GPS points are close to each other for longer times. Then applying sequencing models to predict the next place to be visited, Ashbrook [2002], applied Markov model to get the next location, based on the frequency of the locations found at the users history, Gambs et al. (2012) used Markov chains to figure out the next POI using Geolife dataset, also Mathew et al. (2012) proposed a hidden Markov model for prediction based on the observed time intervals to predict the next locations, these models at the level of single user, can’t predict new POI, furthermore, at the level of a group of users, the model can predict other private POI like home or work of one user to another user with the frequency of the first user home. Second, predictions using location-based social networks (Foursquare, Gowalla), check-ins to predict the next check-in the user expected to check-in which the user can check-in at the place where he/she visited, these datasets are sparse, as the user not always using the application with all the visited places, commonly the user check-in at the important POI, so the predictions can have much more error margin, these predictions as well as the GPS based datasets, researches often use Markov models to tackle these type of analysis, as the data can be processed to be sequences of states the user moving between them, Y.-J. Kim (2013), [Sung-Bae Cho](http://www.sciencedirect.com/science/article/pii/S092523121500569X) (2016), these techniques are used for user specific predictions, otherwise when researches need to solve the problem for all users often then use collaborative recommendation models, Zheng et al. (2010), Ye et al. (2011), these models can recommend novel places for the users based on the social connections between the users and also the spatial and temporal features, but these models failed to personalize the analysis, so they can fail to the problem of predicting private POI of one user to another (home, work), another way to solve the problem can be called a hybrid analysis, which separates the predictions of the novel places from the predictions of the user’s regular places, Defu Lian(2015), check if the next place is novel if so the inputs passed to a recommendation engine, otherwise, if the next location is regular the input data passed to a hidden Markov model , they exploits the power of HMMs to discover the regularities, and the collaborative recommendations to offer new novel POIs, then they combine the result based on the novelty of the POIs , Yingzi Wang (2015) also separated the problem into two small ones, one for regular and the other for novel, then combine the results, we tracked these technique to solve the problem but we focused to discover the novelty of the next location,

Proposed architecture:

Datasets:

We conduct the experiments on two types of datasets:

* Geolife which is a GPS dataset, a dense data, consists of users’ trajectories, these trajectories are quantum of the users’ GPS data related to the user’s trip of the day, it has 37,000,000 GPS points collected from nov. 2007 to jan. 2011 at the city of Beijing, we applied a preprocessing clustering to combine every place the user spent time more than 10 minutes within 50 meters to be able to extract the Points of Interests.
* Gowalla dataset, it is a social network consists of users’ check-ins, which we consider as the POIs, so it did not need preprocessing as Geolife dataset, though we processed these data to get the proposed features.

Preprocessing:

For the Geolife dataset ,We needed to aggregate the GPS points concentrated at one place to be a stay point, we exploited Yu Zheng (2011) algorithm to detect these stay points, briefly, the algorithm checks if the GPS points are less than a spatial threshold and more than another temporal threshold, we set these thresholds, spatial threshold and temporal threshold minutes, after having these stay points, we needed these points which the user stayed at, to be clustered into Points of interest, which they are the aggregation of the stay points located within the same spatial region ,so we used Density-based spatial clustering of applications with noise (DBSCAN), it is a density-based clustering algorithm, we configured it to cluster based on the distance of 50 meters to be considered as the same cluster, with minimum points of two , because if we used a minimum points of one, this means that the cluster with one stay point visited only once throughout the entire dataset, and this can be considered as noise as it can’t be predicted.

For Gowalla dataset, it doesn’t need this type of preprocessing as it already has check-ins (POI) with labels, but it is a more sparse data as the users are the ones who add the check-ins by themselves.

For both datasets we excluded the one time appearance data point, it is not predictable so it may affect the results negatively (trying to remove from the experiment)

Proposed Features:

We extracted the features found at Defu Lian(2015) and then we enhance them by adding new ones, we are revisiting the details of each features to facilitate in case of reproduction of the results, these features are of three types:

* Temporal features:

These features extracted from check-in timestamp.

* + Weekdays: It is a categorical feature between 0 and 6 representing the days of the week
  + Hour of the day: A categorical feature between 0 and 23 representing the hours of the day
  + Hour of the week: a categorical feature between 0 and 167 representing the hours during a week, as a mixture of the previous two features.
  + Time of the day: we classified the day into 3 categories: morning from 5:00 am to 12 am, noon from 12:00 pm to 6:00 pm, and evening 6:00pm to 5:00 am.
  + Check-ins per day: we added a counter feature to count the number of check-ins per day, as the users with many check-ins per a day expected to visit novel next location with low probability.

Visited places per day: it’s a counter to count the number of places visited by the user (u) at certain day (d), this feature showed that the more the user check-in per day the less he/she would check-in at novel POIs.

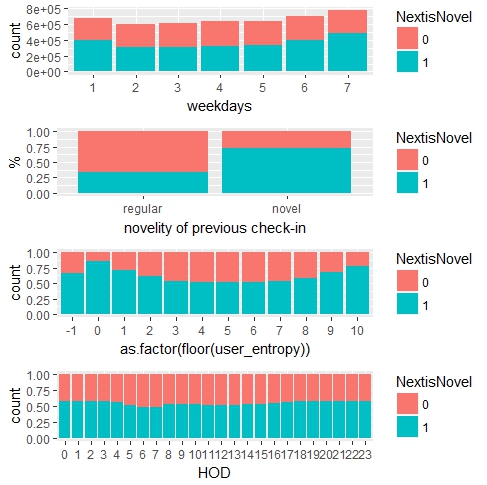
* + Time interval between consecutive check-ins: it’s the difference in time between the current check-in and the previous in minutes
* Spatial Features:

These features are dependent on the location of the check-ins and its surrounding ones.

* + The distance between two consecutive check-ins: using the Haversine distance measure, to calculate the difference in meters between the current check-in and its previous one using the geolocation of the check-ins (latitude and longitude).
  + Location entropy:
  + Visiting ratio: it’s the number of locations visited by the user to the number of all check-ins at this area within geofence (a geographical boundary)
* Historical features:

These features extracted from the users check-in history

* + Distinct number of location the user had visited.
  + User Entropy: As shown by Defu Lian(2015), the users entropy
  + Novelty ratio: which is the ratio between visiting novel POIs to regular POIs for a user (u) up to time (t), the larger value the more the user is expected to visit more novel places. (graph)
  + Novelty of the previous check-in: it’s a binary flag checks if the previous POIs is novel or not for current user. This flag shows the continuity of checking at novel places, and it was shown that it’s related to the probability of the next to be novel.
  + Number of days in user history: it’s an incremental feature which count the number of days in the user’s history up to time (t), it was shown that the more the user uses the service the less novel places will be checked-in.
  + Locations frequency per user: it’s the frequency of visiting a certain location by the user up to time (t).
  + User ID: it’s a unique key for every user, it can offer an insight for the overall behavior of the user.



Algorithms overview

* As we made the data points independent so we divided the data randomly into training, validation and test sets,
* Parameters:
* Grid search hyperparameters: we applied discrete random grid search to find the best tuned hyperparameters (the algorithm parameters such as the number of trees, learning rate, or maximum depth of the trees) for the algorithms we used, this grid search runs the algorithm with different combinations of these parameters, so we can then select the values of parameters with best results.

Ensemble algorithms

* After developing the algorithms we used the predictions from each algorithm against the testing set the Check algorithms correlations, after
* Important variables graph

Experiments:

Results:

Applying the proposed architecture and modified features we were able to enhance the results to 82% table (1) contains the accuracy errors and the effect of every part of the architecture, as shown at the table, by applying the CART algorithm and the features provided from Defu Lian(2015), the results are 74.5% , when adding the new features with the same algorithm we get a value of 79%, also when using the proposed models with the main features the results is 80%, eventually, using the complete architecture we got a results of 82.5%

Conclusion:

The recognition of the Using an ensemble technique and well defined set of features with tuned parameters and hyperparameters we can exceed the threshold of 80% accuracy

References:

Mathew, W., Raposo, R., and Martins, B., 2012. Predicting future locations with hidden markov models. In: Proceedings of the 2012 ACM conference on ubiquitous computing, 5–8 September, Pittsburgh. New York: ACM, 911–918.

S. Gambs, M.O. Killijian, M.N. Prado Cortez. Next Place Prediction using Mobility Markov Chains. Eurosys' [MPM](http://www.cambridgeplus.net/MPM12/) 2012.

Y.-J. Kim and S.-B. Cho. A HMM-based location prediction framework with location recognizer combining k-nearest neighbor and multiple decision trees. In Hybrid Artificial Intelligent Systems, pages 618–628. Springer, 2013.

Defu Lian, Xing Xie, Vincent W. Zheng, Nicholas Jing Yuan, Fuzheng Zhang, Enhong Chen, CEPR: A Collaborative Exploration and Periodically Returning Model for Location Prediction, ACM Transactions on Intelligent Systems and Technology, 2015.

Yingzi Wang, Nicholas Jing Yuan. Regularity and Conformity: Location Prediction Using Heterogeneous Mobility Data, 2015.

Yu Zheng, Lizhu Zhang, Zhengxin Ma, Xing Xie, and Wei-Ying Ma. 2011. Recommending friends and locations based on individual location history. ACM Trans. Web 5, 1, Article 5 (February 2011), 44 pages.