Human Behavior Challenge Winning Solution

Saket Kunwar
NARMA
Kathmandu, Nepal
saketkunwar2005@gmail.com

***Abstract*—This paper describes the winning method and results of the Human Behaviour Challenge 2018 [1]. GPS trajectories from smartphones and associated metadata of people involved in a game of finding hidden objects were collected at Nagoya Institute of Technology, Japan. The competition, hosted at codalab.org, challenged the participants to find a solution that could predict the future destination and past starting point from the given data. Knowing the future and past track data of users could enable the organizers to provide personalized and localized services. The source code [2] under open source license has been released at github.**

I. Introduction

The trajectory dataset, [3], is centered around Nagoya Institute of Technology, Aichi, Japan. There are 9 locations as shown in Figure 1, from which either a person starts a path or is a destination point of the path. Along with trajectory data set, the elapsed time from start to end, and metadata giving duration of the trip, age, and gender of the person was given. The goal was to predict from which location, out of the 9 possible locations the journey started and ended. Knowing the true start and end point of the trajectory, a high accuracy could be achieved by simply assigning start and end label based on the proximity of the goals to the start and end points, but the test trajectory is trimmed. Because of this addition of the trimming factor, We must consider the behavior of the person via the trajectory patterns to ascertain the starting and ending points to achieve high accuracy.

This is a multi-label and multi-class classification problem as start and end both are required to be predicted and there is 9 possible location. Instead, We can treat the combination of start and end label as a unique pair. This reduces the problem to a single label classification problem with 18 classes. At prediction time, We can convert this predicted class back to its constituent start and end location. This scheme matches with our motive of predicting human behavior from trajectories dataset. For example, people who start from the same location with similar trajectories could end up in the same destination. As shown by the distribution of start-end pair in Figure 2, a trajectory in some pairs are followed more often than others.

II. Method

The test dataset is trimmed randomly(5 to 60 seconds), but the training dataset is not. There are also only 263 data points in the training dataset. So to overcome the lack of data points

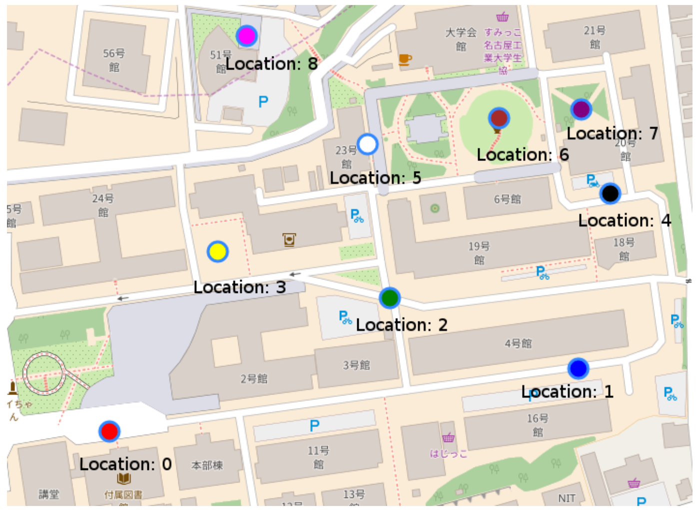

Fig. 1. Locations of start and end

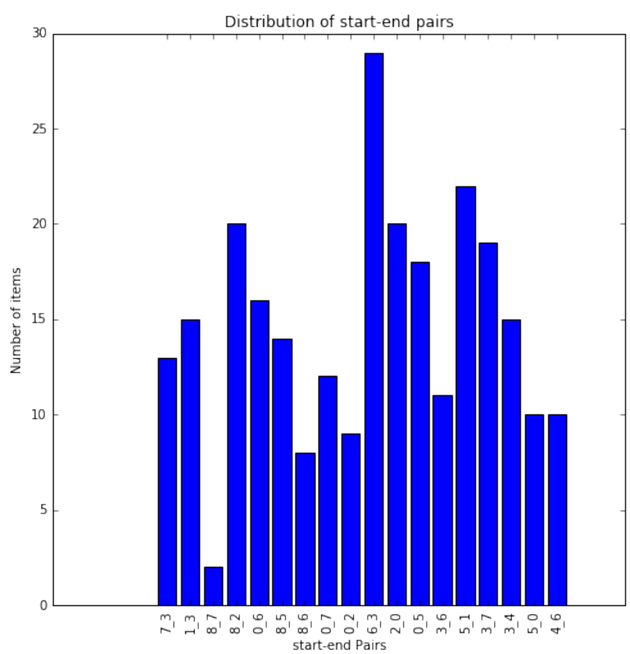

Fig. 2. The start-end pair distribution shows imbalanced class

We decided to aggressively augment the training dataset by randomly trimming it at the start and at the end. Twenty such samples were extracted from each trajectory making the new

[1]https://competitions.codalab.org/competitions/17401
[2]https://github.com/saketkunwar/hbc2018

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

TABLE I
DESCRIPTION OF EXTRACTED FEATURE VECTOR

Feature	Description
Duration	The elapsed time at the end of track
Total Distance Covered	The shortest distance from start to end
Velocity	Total Distance Covered / Duration
Age	Age of the person, NaN filled by mean
Gender	Gender of the person
Distance to goals from start	The shortest distance from the start to All the goals
Distance to goals from end	The shortest distance from the end to All the goals
Closest Distance	The shortest distance when a point on a track is closest to a goal for All the goals
Farthest Distance	The shortest distance when a point on a track is farthest from a goal for All the goals

augmented dataset have 5260 samples. The augmentation by trimming, as in the test set, results in a model that generalizes better to unseen samples.

The feature vector extracted from the trajectories and the metadata is shown in Table I. In particular, the features in the closest distance and farthest distance to the 9 goals can be considered to be an approximation to a trajectory. As the goals are fixed this feature becomes comparable. This feature proved discriminative and performed better than other means of describing trajectories such as clustering and zero-padding so that each track is of fixed length. Hausdorff-distance, to compare trajectory similarity, also proved not to have discriminative power for this dataset.

An ensemble of classifiers containing diverse models can generally yield a better score than a single classifier. For this dataset, an ensemble of 5 classifiers consisting of K-nearest Neighbors, Support Vector Classification, Random Forest Classifier, Extra Trees Classifier from Sklearn python package, [2], and Xgboost, [1], gave the best accuracy score. The hyper-parameters of the classifiers were tuned individually with class-weights set to balanced mode as our class labels are imbalanced. The prediction of all the classifier was averaged to obtain the final prediction.

III. RESULTS

On a 5-fold cross-validation evaluation, the best performing single model was Extra Trees Classifier with an accuracy score of 80 %, while that of Support Vector Classifier was only 70 %. The set of diverse ensemble achieved 79 %. On the test set the ensemble result was better than any single individual classifier, which is part of our final submission. The evaluation metric in the competition for the test set was weighted proportionally to the trimming while we during cross validation did not use this. The start and end confusion matrix from the cross-validation results using augmented dataset and ensemble is shown in Figure 3 and 4 respectively. We can see that location-6 and location-7 has higher miss-classfication due to them being relatively close to each other. The extracted features from augmented dataset combined with paired start-end class, and the use of ensemble gave us the best performance.

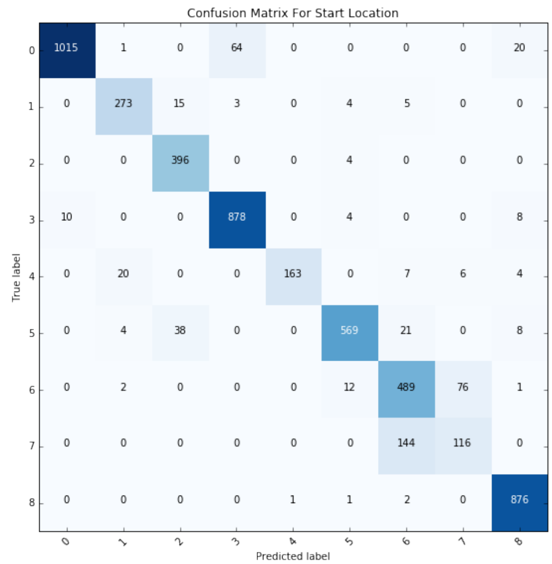

Fig. 3. Ensemble with augmented dataset cross-validation confusion matrix result for start

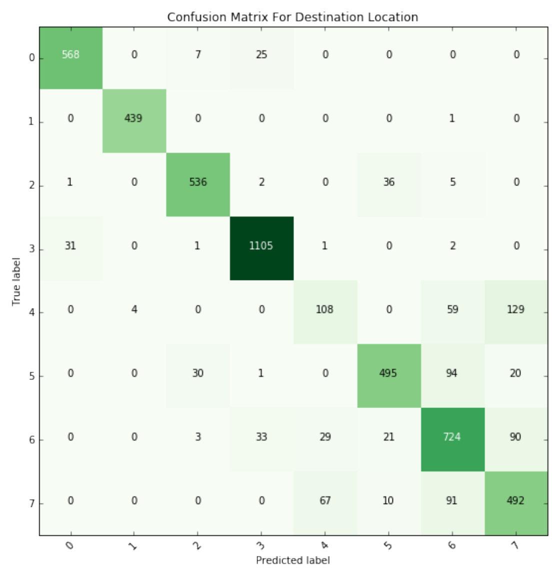

Fig. 4. Ensemble with augmented dataset cross-validation confusion matrix result for end

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] K. Nishi Y. Umezu M. Karasuyama T. Tamaki T. Maekawa T. Uchiya H. Matsuo S. Kajioka, T. Sakuma and I. Takeuchi. Comparative pattern mining of human trajectory data. *IPSJ SIG-UBI Technical Reports*, Feb 2018.