A Location-Based User Movement Prediction Approach For Geolife Project

[1,] Zhiyuan Chen, [2,] Bahran Sanjabi, [3,]Dino Isa

[1,2,3]University of Nottingham, Malaysia Campus

[1,2] School of Computer Science, [3]Department of Electrical and Electronic Engineering

Abstract

Recently obtaining knowledge from raw trajectory data has been an interest of many researches. Trajectory data set consists of thousands of records. To discover valuable knowledge from these records advanced data mining techniques must be applied. Models developed from these techniques will be useful for predication. In this paper data mining classification techniques are analyzed on trajectory dataset and Performance of these techniques is evaluated with recall, precision, kappa and accuracy.

Keywords- Classification Algorithms, WEKA, recall, precision, kappa and accuracy.

I. Introduction

We live in the era that smart phones and other GPS-enabled devices provide people the opportunity of capturing GPS trajectories everywhere at every time. Mining trajectory data leads in obtaining useful information e.g. prediction the user's behavior. Forecasting the behavior of users helps to have a better understanding of users' needs and it has many advantageous since it makes users' live simpler, more comfortable or even more secure. Considering situation that next preferred transportation mean of user is car then high quality location based services will provide to do list such as recommending car parks or gas stations on the way. Acquiring knowledge from raw trajectory data using data mining techniques has been an interest of many researches. (Zhou 2007) presented the approach that can detect important - frequent and important – non frequent locations. Using clustering algorithms and different classifiers he found the level of importance of locations in the trajectory.(Andrei Papliatseyeu, Oscar Mayora, 2008) used Naive Bayes, hidden Markov models and simple Neural Networks to analyze the performance of activity recognition from raw data collected by GPS, GSM and WIFI. The purpose of this project is applying data mining techniques (namely classification) on raw GPS records to predict the mode of transportation (such as taxi, bike, personal car and etc.) users choose once they arrive at a certain point. This paper concentrates on performance of classification algorithms. The classification algorithms considered here are Decision tree, Naïve Bayes classifier, Bayesian network, Neural Network algorithm and Support Vector Machines. These classifiers are compared based on statistical parameters such as Accuracy, Recall, Precision, Confusion matrix and Kappa. It will be shown that decision tree and Bayesian network are acceptable classifiers for classifying trajectory data set. Data mining software used in this project is WEKA (Waikato Environment for Knowledge Analysis) which is a collection of data mining algorithms. The structure of this paper is as follow: next section discusses about data mining technique. Section 3 explains the data set used and the process of preprocessing data .Section 4 is allocated to the result of experiment and section 5 describes a summary of future research.

II. Data Mining Technique

A. classification

Data mining is extracting valuable knowledge and useful pattern from raw data. One of the well-known data mining techniques is Classification which is a supervised learning algorithm. Data classification involves two phases; training phase where the classifier algorithm builds classifier with the training set of tuples and test phase where the model is tested on testing set of tuples

B. Different classifiers

Classifiers considered in this project are Decision tree, Naïve Bayes classifier, Bayesian network, Neural Network algorithm and Support Vector Machines.Decision tree: Decision tree is widely used in data mining project because it is easy to understand and gives a clear representation of how decisions are made. Decision tree consists of root node, in rnal node and leaf node. Internal nodes are between root node and leaf node. The condition is assessed at each

node if it has ositive result the data is sent to the leaf node otherwise it is sent to the non-leaf node and portioning process repeats until it reaches to leaf which assigns a class label to the data sample.Bayesian classifier: A Bayesian classifier is based on Bayes' theorem which states: P (Y|X) = ((P (X|Y) P(Y))/P(X). In order to determine a classification using Bayes theorem, P (Y|X) needs to be known for every possible value of X and Y. Two main type of Bayesian classifier are Naive Bayes Classifier and Bayesian Network. Naïve Bayes has a naïve assumption of independence between all attributes meaning that the presence or absence of one attribute has no impact on the next whereas the Bayesian Network allows conditional independence between attributes to be applied to only particular pairs of attributes.Artificial Neural Network: Artificial Neural Network is a system inspired by human neurology. The structure of ANN is like layer model. Each layer is made of numbers of interconnected nodes which connect to next layer via direct links with various weights. The first layer is called input layer that receives the input data and transmits it over next layers that are called hidden layers where the processing is applied. Hidden layers then shift the output to output layer. For a neural network to be useful, it must first be trained so that the weights of the links can be adjusted. Adjusting the weights of the links can be done in a couple of ways such as Back- propagation.

Support Vector Machine: It is based on the concept of decision planes where the training data is mapped in to a higher dimensional space and separated by a hyper plane to differentiate between two or more classes of data. The "support vectors" are those points in the input space which best define the boundary between the classes. The selected hyper plane for an SVM should be the one with the largest margin between the two classes because it creates clear boundary between them. (Bottou L., Chih-Jen Lin, n.d.)

III. Description Of Data

The data set used in this project is a portion of GPS trajectory data set which was gathered for GEO life project. Recently number of researched have been done using this data set. For instance mining interesting location and travel sequence (Yu Zheng, 2009), finding similarity between users (Li, 2008) and learning automatically transportation mode (Zheng, 2008)

A. Raw data

A trajectory data set is a sequence of GPS records that are ordered by the timestamp of the records. This data set contains 17,621 trajectories that are gathered from more than 170 people and have a total distance of 1,251,654 kilometers and a total duration of 48,203 hours. Each trajectory folder is related to one particular user. These data are in PLT format and contain following fields:

Field 1: Latitude in decimal degrees.
Field 2: Longitude in decimal degrees.
Field 3: All set to 0 for this data set.
Field 4: Altitude in feet (-777 if not valid).
Field 5: Date - number of days (with fractional part) that have passed since 12/30/1899.
Field 6: Date as a string.
Field 7: Time as a string

B. Data Preprocessing

Data preparing is the vital step in data mining procedure. In this project the available data set was in PLT format. Since WEKA software accepts some distinct format the first step was converting data from PLT format to CSV format which is acceptable for WEKA. In second step impractical fields (fields 3, 4, 5) has been removed. Approximately 23% of users labeled their context by indicating the mean of transportation they used such as driving, taking a bus or taxi, using a subway, riding a bike, walking and in rare condition flying with airplane. In third step since files of trajectory and transportation label were stored separately from each other, long time has been spent to match the time and date of these files and create one complete data set. In step four, in Microsoft Excel environment the interval time user stayed in each pair of latitude and longitude point has been calculated using math function. Therefore the new data set contains latitude, longitude, date, time, transportation mode and duration. Step five was creating two tables from data set; stop table and move table. Stops can be assumed as important points of a trajectory if user stays more than a period of time. Using mathematical function in Microsoft excel the points that user stayed more than 10 minutes has been extracted and moved to stop tables. Stop tables can be used for creating location history and personal map of users. Other points have been moved to moving table. Classifying algorithms have been applied on move tables.

C. Weka

In this project WEKA "Waikato Environment for Knowledge Analysis" has been used. WEKA is an open Source Machine Learning Software that is written in Java and developed by the University of Waikato in New Zealand .It is a collection of machine learning algorithms and data preprocessing tools that helps researchers to mine different data sets. WEKA has four environments; simple CLI, explorer, experimenter and knowledge flow. In this project the explorer environment has been used. In WEKA, The results of classification is divided into several sub categories which is more human readable and easy for evaluating. First section shows the correctly and incorrectly classified instances in numeric and percentage value. Kappa statistic, mean absolute error and root mean squared error are presented also in this category.In second part parameters for measuring accuracy of each class is shown .These parameters are FP, TP, ROC area, F-measure, Recall and precision.The third section is confusion matrix which is one of best measurement for evaluating classifiers

IV. RESULT

In GEO life project, people who collected their GPS trajectories had a different period of collaboration. Some of them have a long collaboration and carried a GPS logger for several years while some others cooperated for just a few weeks hence the size of their trajectories were different. Small size with five class of transportation, middle size with 8 class of transportation and long size with 10 class of transportation were selected for applying classification algorithms. In this paper the result of classification on small data set will be shown. This data set has 44236 instances and 5 classes. Each classifier has been tried on two test options; 10 fold cross validation and percentage split 66%.

A. Result Of Classifiers

TABLE 1-PERCENTAGE SPLIT 66%

Classifier (%)	Correctly classified instance	Incorrectly classified instance	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Bayes NET	77.697	22.3023	66.79	10.96	24.51	41.5289	67.455
Naïve Bayes	58.1073	41.8927	39.14	0.19	33.45	71.9511	97.547
J48	87.296	12.7034	80.59	5.66	17.52	21.432	48.2095
Ann	66.401	33.5989	45.58	11.75	24.9	73.759	88.2569
SMO	51.491	48.5084	6.52	27.06	36.07	102.48	99.2469

TABLE 2-10 FOLD CROSS VALIDATION

Classifier (%)	Correctly classified instance	Incorrectly classified instance	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Bayes Net	77.7436	22.2564	66.83	10.77	24.36	40.798	67.042
Naïve Bayes	57.468	42.532	38.18	19.13	35.69	72.474	98.250
J48	87.0617	12.9383	80.26	5.87	17.71	22.218	48.7413
Artificial Neural Network	63.7762	36.2238	43	16.58	30.72	62.814	84.559
SMO	51.5474	48.4526	6.48	27.12	36.1	102.7	99.370

B. Comparing classifiers

The performance of classifier is evaluated by parameters like accuracy, precision, recall and kappa.

Correctly classified instance presents the percentage of instances which were classified correctly and this measure is often called accuracy. Precision is the fraction of instances which truly have class x among all those which were classified as class x. Recall is a fraction of instance which correctly classified as class x among all instances that belong to class x.

Kappa is a measure of agreement normalized for chance agreement. Kappa = P (A) – P (E) / 1 – P (E) Where P (A) is the percentage agreement between the classifier and ground truth and P (E) is the chance agreement. A value greater than 0 shows that classifier is doing better than chance.

TABLE 3- weighted average of recall, precision and accuracy

Classifier	Weighted average of Recall	Weighted average of Precision	Number of Correctly classified instance
Decision tree(j.48)	0.871	0.871	35953
Naïve Bayesian	0.633	0.575	23733
Bayes Net	0.796	0.777	32105
Artificial Neural Network	0.637	0.638	26337
Support Vector Machine	0.323	0.515	21287

It can clearly be seen that Decision tree (j.48) has the maximum accuracy and better recall and precision. On the other hand support vector machine is the least accurate classifier. High recall demonstrates that an algorithm correctly classified most of the instance of each class. High precision means that result of an algorithm is more correct than incorrect.

TABLE 3- COMPARING KAPPA

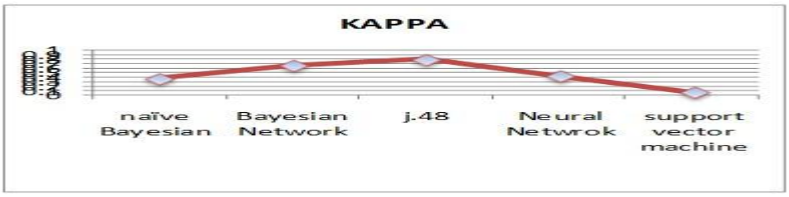

Figure 1- comparing Kappa

J.48 has the highest kappa. It means that the result is more close to truth than chance.Another important concept for evaluation the performance of classification is confusion matrix. A confusion matrix illustrates the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The confusion matrix is an array with n size where n is the number of classes.

TABLE 4- confusion matrix obtained using decision tree classifier

a	b	c	d	e	<-- classified as
5625	6	858	1383	22	a = bike
40	2813	204	43	14	b = subway
393	17	19724	143	31	c = bus
2031	12	70	7658	1	d = walk
19	8	39	9	133	e = car

It is obvious that majority of instances are classified correctly.

V. Conclusion

Data mining through different technique turn raw data in to meaningful information. In this research data mining methods have been used to mine trajectory dataset which were gathered from people who have collaborated with GEO-life project. The final goal of this research is prediction the mode of transportation users use based on geographic location they are. To achieve this goal considerable effort has been put to prepare the proper data set in preprocessing level. Three sample sizes of trajectories have been selected and each of them categorized in to stop and move data set. prediction the state of transportation is achieved by applying classification algorithms on move data set .Decision tree, naïve Bayesian, Bayesian Network, Support Vector Machine and Artificial Neural Network were used as classifiers and their efficiency were evaluated by precision, Recall, Accuracy and Kappa. Decision Tree achieved the highest score and Bayesian Network was in the second place. Support vector machine illustrated weak result and it might because of the structure of data set which has lots of classes and few attributes.

VI. Futur Research

As a future extension of this study we will create models for predicting the use of public transportation or personal one in dense and popular regions. Density based clustering will be applied on trajectory data set in order to find most dense region and then by using SQL commands in data base the probability of using public transportation (bus-subway-taxi) and personal transportation (car-bike) will be calculated. Then by applying classifiers namely decision tree and Bayesian network the model for predicting the use of public or personal transportation will be build.

VII. References

1) Andrei Papliatseyeu, Oscar Mayora, 2008. Inferring and predicting user activities with a location aware smartphone. Advances in Soft Computing , Volume 51, pp. 343-352.
2) Bottou L., Chih-Jen Lin, n.d. Support Vector Machine Solvers. s.l.:s.n.
3) Christine Körner,Jutta Schreinemacher, 2012. Analyzing Temporal Usage Patterns of Street Segments Based on GPS. s.l., s.n.
4) Quannan Li1,2, Yu Zheng2, Xing Xie2,, 2008. Mining User Similarity Based on Location History. s.l., 16th ACM SIGSPATIAL international conference on Advances in geographic information systems .
5) Yu Zheng, 2009. Mining Interesting Locations and Travel Sequences from GPS Trajectories, s.l.: s.n.
6) Zheng, Y., 2008. Learning Transportation Mode from Raw GPS Data for Geographic Applications on the Web. NewYork, 17th international conference on World Wide Web .
7) Zhou, C., 2007. Mining Personally Important Places from GPS Tracks. Istanbul, Data Engineering Workshop, 2007 IEEE 23rd International Conference.
8) Remco R. Bouckaert, E. F. H., 2012. WEKA Manual for version 3.6.7. s.l.:s.n.