Classification of a trajectory with different unknown activities based on established rules from known tracks

Galindo, Diana
d_gali01@uni-muenster.de

Drate, Pamella
p_drat01@uni-muenster.de

August 17, 2012

Abstract

Background: The study of moving objects (animals, people, and objects among others) is increasingly being done by many organizations for various reasons ranging from service provision to security reasons. Since the trajectories, (moving objects) have properties of space and time; this information is of interest for scientific studies and research. This project aimed at exploring tools of classification from data mining using decision trees and segmentation to understand activities along different modes of transport. The data included tracks of separate activities to establish rules for a track with mixed activities.

Method: We established rules for the individual tracks; these rules were centered on the speed and the stops that were made in these activities. The RWeka package [Hornik et al., 2009] was used to classify the training data thus the individual tracks(separate activities) joined to form one track and a prediction was made on the single track with mixed activities. We used the unsupervised classification to have an overview of the data for supervised classification.

Results: The visualized results of the predicted track with multiple activities gave a clear impression of the activities since they were correctly classified. Approximately 75% of the classifications were correct as per the confusion matrix which was derived from the J48 decision tree method.

The results from the methods were very similar with only minor variations.

Conclusions: The activities were classified correctly basing on the results visualized. However we obtained some errors since the rules were not explicit and only speed as a variable was considered in the establishment of the rules. The accuracy levels may have been compromised by rules only based on outdoor activities and indoor periods were not considered for the track with multiple activities.

Keywords: Trajectory, classification, RWeka, GPS, R, decision tree, segmentation, GPS.

1 Introduction

Classification of trajectories is becoming increasingly popular around the world; many people track themselves using gps devices and mobiles phones and provide the data freely on internet for those interested in analyzing the movements. Accurate classification of the GPS tracks is still very demanding since the rules established vary according to the purposes for the classification and the conditions in which the data was collected.

The tracking processes are always influenced by many factors such as signal loss, faulty equipment, diary mistakes, among others. These highly impact on the classifications. In this study we considered these aspects to classify our own tracks based on set rules from known activities. The project was done using R software and supportive packages of *rgdal, sp, adehabitatLT, RWeka, rpart, reshape* and *lattice* were used.

2 Research question and hypothesis

Our interest in this study is to classify and asses the levels of accuracy in classification of the track with many unknown activities basing on known activity rules.

2.1 Research question/statement

To assess the classification of a single track based on rules obtained from multiple tracks and evaluate its accuracy from the diary of track information.

The research will compare 2 methods: The research will compare classification of regular and irregular tracks. We will consider unsupervised classification in R (using segmentation) to give the first impression of the results for analysis.

2.2 Hypothesis

1. Do rules from individual tracks correctly classify a track with multiple activities.
2. Does speed accurately classify movements in space and time for different activities.
3. There is a difference in the classification using regular or irregular tracks to establish rules and to use the decision tree.

3 Methodology

The processes followed in the execution of the project are explained by the figure 1 which shows the different steps undertaken.

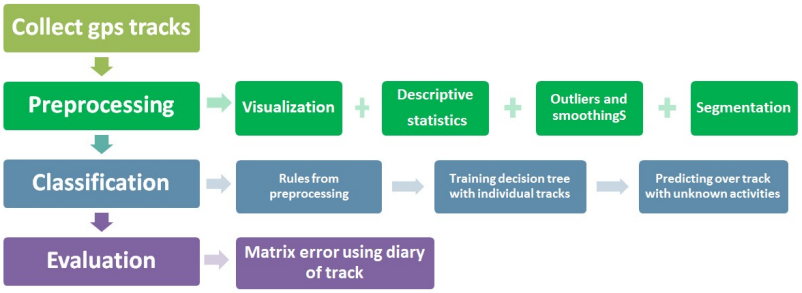


Figure 1: An illustration of the methodology frame work

The steps followed in the methodology figure above are further explained in detail below:

3.1 Data collection and the spatial temporal data

The data was obtained in the course of the semester using a hand held GPS, around the city of Muenster, Germany and Enschede, Netherlands. The data categories collected are in two forms: one single activity per track and mixed activities per track. The tracks with single activities were collected in two ways, regular and irregular time stamp. All the tracks have diary information describing the details and changes in activities, including those with mixed activities. The data is further categorized as:

3.1.1 Single activity tracks

1. **Walking:**In the project context, walking means, the path followed by a person including some short stops as traffic lights. This data set consists of one track, regular time stamp of 30 seconds with 38 points.
2. **Joging:** Means the activity of running at a certain speed interval. This dataset has a regular track of 72 points and time interval of 30 seconds.
3. **Bike:** The activity of riding a bike with different speeds, considering small stops for walkers or traffic lights. This dataset is in two forms, irregular and regular, with 150 and 30 points respectively. The time interval for the regular track is 30 seconds.

4. **Bus:** This activity information is based on the bus network of city of Münster, the stops were also considered. It comprises of two datasets, irregular track with 130 points and regular track with 30 points and 30 secs of time interval.

5. **Train:** Includes two tracks, the irregular track with 3411 points and regular with 716 points and 30 seconds time interval.These datasets were captured during the trips in the train and include the stops in the different stations.

3.1.2 Multiple activity track

The second data set consists of different activities all in one track: It is composed by two tracks, irregular with 833 points, has different activities as bus, walking and biking and a regular one with 5915 points which comprises of walking, jogging, bike, bus and train activities.

3.2 Exploratory analysis and data preprocessing

In this stage, explorative analyses were done including: visualization in *Google Earth*, descriptive statistics and segmentation for un supervised classification. The preprocessing was based on detection of outliers, smoothing using Moving Averages. The preprocessing gave a starting point in understanding the data trends and therefore provided a basis for the establishment of the rules that subsequently were used in the classification process.

3.3 Visualization

The first exploratory analysis of the data was visualization of the different tracks in Google earth in order to verify the spatial coherence. The figure 2 shows tracks of the individual activities: bus (red), train (yellow), biking (orange) and walking (green). The visualization provided us with an idea of the data display which was accurate in most communication paths with only a few deviations from the roads or the railways.

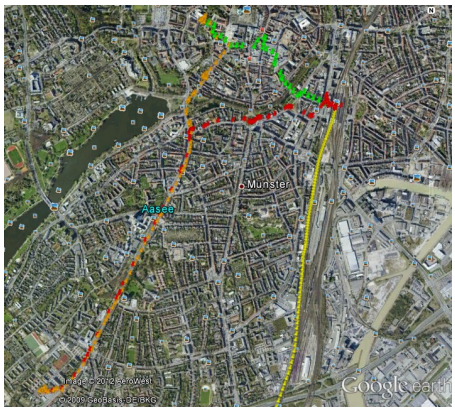


Figure 2: Collected tracks with different activities. Source: Google Earth and GPS tracks collected.

3.4 Descriptive statistics

The subsequent step was to obtain the summary of descriptive statistics that were used as a basis to establish the rules to classify the track with unknown activities, this step includes the calculation of speed for each track as one of the most important variables we considered for establishing the rules of behavior per activity.

The table 1 shows the summary statistics per activity. The descriptive statistics were taken as reference to establish the rules of the classification.

Type	Variable	Walking			Jogging			Bike			Bus			Train		
		1stQu	3rdQu	Median	1stQu	3rdQu	Median	1stQu	3rdQu	Median	1stQu	3rdQu	Median	1stQu	3rdQu	Median
Regular	speed	3,41	5,189	4,817	9,155	10,14	9,88	4,479	17,69	14,06	4,886	24,16	16,21	12,6	90,13	70,2
	distance	28,47	43,24	40,14	76,2	84,49	82,34	37,33	147,4	117,2	40,701	201,3	135,1	17,61	79,21	97,54
Irregular	speed	-	-	-	-	-	-	4,353	15,53	9,437	1,189	20,7	12,41	29,34	127,6	80,29
	distance	-	-	-	-	-	-	3,025	45,24	8,901	2,054	33,05	6,061	16,55	71,15	44,81

Table 1: statistics per activity and type of track

3.5 Outliers and smoothing

The velocity time series approach was used to identify outliers for the single tracks, the threshold varied according to the tracks. Moving window average method was used **in regular tracks** to derive threshold from standard deviation and to remove the outliers by way of smoothing. The window size varied according to the amount of noise from the tracks. **For regular tracks was used boxplot diagrams to overview presence of outliers**, only train track showed presence of extreme data. The figures 3 and 4 illustrate the result of the moving averages procedure for different activities in regular and irregular tracks .

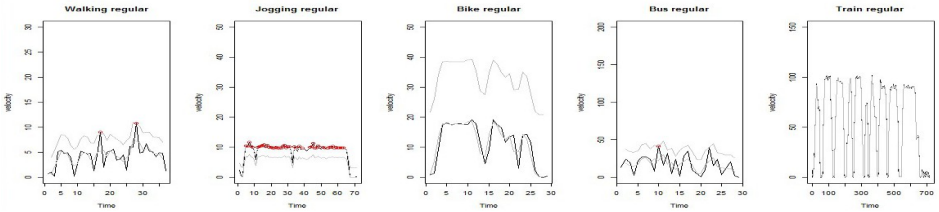


Figure 3: Activities in regular tracks

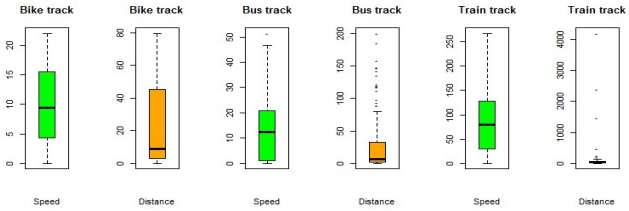


Figure 4: Activities in irregular tracks

Figure 5: Points highlighted in red show the outliers which caused noise in the track

3.6 Unsupervised classification

We used segmentation to perform unsupervised classification; this was done to give a general impression of the track with mixed activities and later on used for understanding the supervised classification. This step involved the use of the gueguen method in R software to segment the trajectory into segments characterized by a homogenous behavior. We could establish similar patterns in the trajectory which meant similar activity but could not say which activity it was.

This method only gave us an impression of the trajectory with multiple activities for guidance in the supervised classification process.

4 Supervised classification

In the supervised classification procedure, the decision trees were used to illustrate the rules that were administered. Decision trees are hierarchical models for classes according to established rules or conditions of the data. The training dataset consisted of the single data set per activity joined to form the training data set in order to obtain the model to predict the track with different activities.

The procedure was developed in R statistical software using RWeka package to implement the decision tree.

4.1 Variables used

We considered speed as the major variable in the classification; it's from this that the rules were established. Speed being a result of distance and time, therefore the time and distance were also considered in the classification process.

4.2 Rules set

The rules established as mentioned earlier were mainly derived from speed for each trajectory. We calculated the speed for each individual track, looked at the maximum and minimum to set the rules for that trajectory. A set of rules were establish for each type of tracks: regular and irregular.

We established logical rules despite the results we obtained from the tracks, especially for the activities we considered could have higher speed than the others. For example if an activity like jogging had a lower minimum than walking; we had to use logic to make the minimum for jogging higher.

The rules were only done for outdoor activities for the single tracks we therefore did not consider any rules for the indoor activities.

The single tracks were then joined together as a preparation for the classification. The first set of rules is summarized in the table 2.

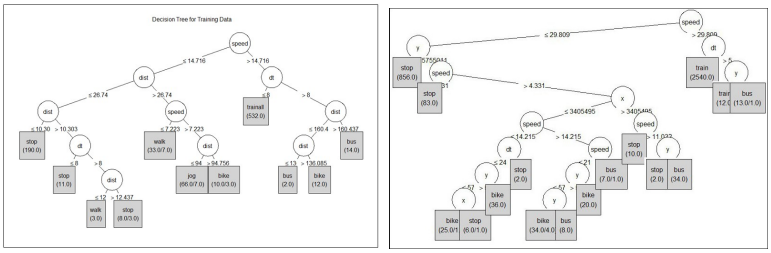
Rule of speed(s)		
Activity	**Regular**	**Irregular**
Walking	1< s <= 11	-
Jogging	2< s <= 12	-
Biking	4< s <= 20	4< s <= 22
Bus	10< s <= 42	12< s <= 50
Train	15< s <= 103	30< s <= 270

Table 2: Rules per activity

The established conditions / rules were added to the trajectories as categorical variables, for those points in which the condition was satisfied, the variable took the name of the activity, while in the opposite case the data points were then labeled as stop.

4.3 Decision Tree

The dataset with the joint individual activity track was used to train a decision tree. We chosen the default options of the J48 algorithm, this means, the process determines in which section to prune the tree and assign the points to classes according to the information provided by the training data. The resulting structure of the tree is shown in the figure 6:



(a) regular (b) irregular

Figure 6: An illustration of the decision trees from the J48 formula

The tree shows the distribution of the activities basing on speed: distance and time were used in the model since they are a function of speed. The activities with more speed intervals over distance and time are train and bus while those with less speed interval are walking and the stops.

The prediction formula was later used, which involves the trained data set used for the decision tree and the track with multiple activities which was the target. The prediction was done basing on the output from the training data.

As a result, we obtained a new variable with the category of the corresponding activity in the track with multiple activities (3 and 4).

id	x	y	date	dist	dt	speed	cat
1	3364930	5788617	2012-07-14 09:08:24	0.358885	5	0.258397	?
2	3364930	5788617	2012-07-14 09:08:29	0.405418	5	0.291901	?
3	3364929	5788617	2012-07-14 09:08:34	1.132649	5	0.815507	?
4	3364929	5788618	2012-07-14 09:08:39	0.917207	5	0.660389	?
5	3364928	5788617	2012-07-14 09:08:44	2.731477	5	1.966664	?
6	3364928	5788620	2012-07-14 09:08:49	0.843423	5	0.607265	?

Table 3: shows the initial values of the track with multiple activities with cat(category left blank for prediction)

id	x	y	date	dist	dt	speed	Prediction
1	3364930	5788617	2012-07-14 09:08:24	0.358885	5	0.258397	stop
2	3364930	5788617	2012-07-14 09:08:29	0.405418	5	0.291901	stop
3	3364929	5788617	2012-07-14 09:08:34	1.132649	5	0.815507	stop
4	3364929	5788618	2012-07-14 09:08:39	0.917207	5	0.660389	stop
5	3364928	5788617	2012-07-14 09:08:44	2.731477	5	1.966664	stop
6	3364928	5788620	2012-07-14 09:08:49	0.843423	5	0.607265	stop

Table 4: Showing the prediction values from the trained data set.

Considering that the evaluation of the method will be performed through the information of the diary, there was no validation dataset for the model of the tree.

5 Model evaluation

5.1 Evaluation

To measure the accuracy of the model, an error assessment was done using cross validation with the classified track per activity and the diaries of the two tracks.

5.2 Diaries influence in the procedure

The diary was considered as the ground truth and it was followed as the accurate guide. The rules and the classification algorithms were mainly based on the diaries especially in areas were the tracks went off the roads on main transport channel we considered the diary to have the actual result. We compared the classified track with multiple activities, with the unsupervised classification, to establish some similarities with guidance from the diaries.

According [Wu et al., 2011], "There is no gold-standard" against which to compare the classification of GPS-derived time activity data". The information of the diary also was supplied as a categorical variable in order to compare real and predicted activities, howewer, the the diary constraint its usability for error estimation of the model.

5.3 Error matrix

The statistics from the confusion matrix was considered in the process of assessing the error. The analysis focused on the classes that were wrongly classified and the classification ratio obtained to ascertain the levels of precision for the model.

Below is an illustration of the confusion matrix obtained after the prediction:

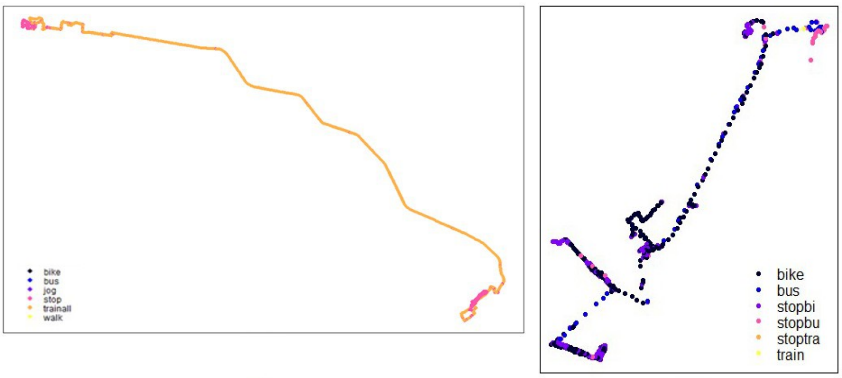
From the table 5 we were have train and bus activities that didn't have wrong classifications among them, we think this was mainly due to these activities having higher speed intervals than the others.

Confusion Matrix						
a	**b**	**c**	**d**	**e**	**f**	**<-classified as**
19	0	2	0	0	1	**a = bike**
2	16	1	0	0	0	**b = bus**
1	0	59	1	0	3	**c = jog**
0	0	2	206	0	3	**d = stop**
0	0	0	0	532	0	**e = train**
0	0	2	2	0	29	**f = walk**

Table 5: Showing the confusion matrix of the trained data

5.4 Results

The prediction results obtained were visualized as a trajectory, where different colors were assigned to the categories predicted (Figure 7). The trajectory had clear descriptions for the different trends that resulted from the rules.



(a) regular (b) irregular

Figure 7: The trajectory from the prediction showing the different activities.

It can be observed the categories by trajectory show consistent results of activities, in detail, the classification losses accuracy, this could be due additional elements that are affecting the trajectories as traffic which are out of scope of regular diaries used.

The attached script shows that if the trajectory is segmented then we can visualize the different activities in detail.

6 Discussion

The discussion covers the research question and hypothesis we had in mind in the development of the project.

Our outstanding interest was to develop rules from single activity tracks and use these rules to predict multiple track activities. The prediction gave us understandable results; approximately 75% of the multiple track activities were correctly classified.

We had a hypothesis of considering whether speed is a good measure for establishing rules, with the results we obtained, speed which is a factor of distance and time: gave good predictions for the activities.

During the development of the project it could be concluded there is no significant difference establishing rules for classification according to irregular or regular trajectories, also, that the parameter dt feeds the algorithm to define the classification, property that could be lost in case of regularization of the tracks.

The wrong classifications, in our opinion were mainly due to rules which we based only on one variable thus speeding. We think if we had considered more variables then the accuracy levels would have been higher.

We did not consider periods indoor and outdoor in the establishment of the rules for the single activity tracks since we were aware that these tracks were collected during out-door periods.

However the multiple track activity had several periods in door which we did not consider in the rules establishment and this may be a reason for the compromise on the accuracy of the results.

We also experienced some challenges in the diary recordings, some recordings seemed to have been done wrongly or some changes in activities were not recorded or forgotten. This compromised the decisions to consider.

We had a limitation of no validation data set but based the validations on the diary, however we think this could also account for the low levels in accuracy which could have been higher with validation data and adjustments for the training data.

The challenges and limitations above, we suggest that should be considered in the future work development.

7 Conclusion

The study was success as shown by the results obtained: which were satisfactory giving real impression of the trajectories in space and time. For example the train mode of transport(for regular) and bike mode (for irregular) were correctly classified along the railways, unlike if we had train mode of transport on the road upon visualization of the prediction in Google earth.

Future work should focus on detailed studies on time-activity patterns and speed considering closely the periods in door and out door. We suggest that future work should be comprehensively administered to establish rules for many variables other than speed only to obtain more precise results and trends.

References

[Gerharz et al., 2009] Gerharz, L., Krüger, A., and Klemm, O. (2009). Applying indoor and outdoor modeling techniques to estimate individual exposure to pm2. 5 from personal gps profiles and diaries: a pilot study. *Science of the Total Environment*, 407(18):5184–5193.

[Hornik et al., 2009] Hornik, K., Buchta, C., and Zeileis, A. (2009). Open-source machine learning: R meets weka. *Computational Statistics*, 24(2):225–232.

[Wu et al., 2011] Wu, J., Jiang, C., Houston, D., Baker, D., and Delfino, R. (2011). Automated time activity classification based on global positioning system (gps) tracking data. *Environmental Health*, 10(1):101.