João C. Ferreira, Vítor Monteiro, José A. Afonso, João L. Afonso

"Methodology for Knowledge Extraction from Mobility Big Data"

Advances in Intelligent and Soft Computing, 1st ed., Sigeru Omatu, Ali Selamat, Grzegorz Bocewicz, Pawel Sitek, Izabela Nielsen, Julián A. García-García, Javier Bajo, Ed. AISC Springer Verlag, 2016, Part I, pp.97-105.

http://link.springer.com/chapter/10.1007%2F978-3-319-40162-1_11

ISBN: 978-3-319-40161-4 (Print) 978-3-319-40162-1 (Online)

ISSN: 2194-5357

DOI: 10.1007/978-3-319-40162-1_11

João C. Ferreira, Vítor Monteiro, José A. Afonso, João L. Afonso. "***Methodology for Knowledge Extraction from Mobility Big Data***", in Advances in Intelligent and Soft Computing, 1st ed., Sigeru Omatu, Ali Selamat, Grzegorz Bocewicz, Pawel Sitek, Izabela Nielsen, Julián A. García-García, Javier Bajo, Ed. AISC Springer Verlag, 2016 Part I, pp.97-105. ISSN: 2194-5357 DOI: 10.1007/978-3-319-40162-1_11 ISBN: 978-3-319-40161-4 (Print) 978-3-319-40162-1 (Online)

Methodology for Knowledge Extraction from Mobility Big Data

João C. Ferreira[1,3], Vítor Monteiro[1], José A. Afonso[2], João L. Afonso[1]

[1]Centro ALGORITMI, University of Minho, 4800-058, Guimarães, Portugal,
[2]CMEMS-UMinho, Guimarães, Portugal, and [3]ADEETC at ISEL, Lisbon, Portugal

Abstract. The spread of mobile devices with several sensors, together with mobile communication, provides huge volumes of real-time data (big data) about users' mobility habits, which should be correctly analysed to extract useful knowledge. In our research we explore a data mining approach based on a Naïve Bayes (NB) classifier applied to different sources of big data. To achieve this goal, we propose a methodology based on four processes that collects data and merges different data sources into pre-defined data classes. We can apply this methodology to different big data sources and extract a diversity of knowledge that can be applied to the development of dedicated applications and decision processes in the area of intelligent transportation systems, such as route advice, CO_2 emissions reduction through fuel savings, and provision of smart advice for public transportation usage.

Keywords: Big data, data mining, Naïve Bayes, mobile device, sensor information.

1 Introduction

Knowledge discovery (KD) from the big data available in databases has been explored with success in several areas, and usually is better known through the more popular term "data mining" [1]. Data mining (DM) allows the search for knowledge in large volumes of data [2] and it has been applied to different areas to discover hidden profitable patterns in databases (business opportunities). It is also applied in major research areas, as the example of application in bioinformatics, in genomics and proteomics identification [3].

Some works on this novel research area include traffic analysis and prediction [4], road accidents analysis [4], driver behaviour prediction [6], and range prediction for electric vehicles [7]. This paper is oriented to establish a working methodology that can reuse processes and algorithms in many cases of KD from the big data available from tracking users' mobility activity. This goal results from our experience in several mobility projects, where the DM processes can be reused, avoiding the costs associated to the development of solutions to new problems from scratch. The intention is to create a tool tailored to the increasing data available in the mobility area, which can be used with little effort in different application cases.

This information can be easily collected from mobile devices or based on commercial products. Our work on mobile devices uses GPS and accelerometer sensor data to passively track the users' mobility activity, as described in Section 2. This is a personalized data tracking process from where we can extract useful knowledge related with user mobility habits, like the transportation mode (bus, train, car, bike, walking or other) and carrier number associated, identify main mobility patterns, and generate a mobility invoice. Passive tracking of user activity using mobile devices [8] has been used in a diversity of studies applied to activity recognition [9] and transportation mode detection [10], among others [11][12]. Through the proposed approach, our goal is to establish a common process to handle the collected data into a diversity of KD, in order to facilitate the development of dedicated applications for intelligent transportation systems (ITS).

This approach also allows the joining and aligning of the work of field experts on generating the mobility big data with mining experts, in order to better extract knowledge from the collected data. This can be achieved by data discretization in predefined classes, performed by these field experts. This approach can also be applied for non-professional cases, such as personal data tracking, allowing its extension to the mobility habits of millions of users just by carrying cell phones in their pockets. User data privacy from this tracking activity is an important concern that increases in the context of big data. Thus, all personalized information concerning user mobility is stored in a central database with security parameters (login, encryption). Although we have individual data regarding users' movements, the only personalized information stored is the user email to exchange information.

The main contribution this paper is a common approach of Naïve Bayes (NB) application for KD, based on the normalization of different collected mobility data into predefined classes and sub-classes, which can be used for a diversity of mobility applications developed in a mashup approach. This approach makes possible an increase in the number of available, dedicated or personalized ITS applications, because the development is facilitated by the use of common parts and processes.

The next section presents the proposed work methodology. Section 3 presents examples of application development based on KD from mobile device data. Finally, in Section 4 we highlight the main conclusions.

2 Work Methodology

To extract knowledge we need a proper methodology to work with the big data generated by different sources. Due to analytical needs and the huge amount of data generated, this is a field in which it is possible to apply KD [13] based on a DM approach. The work methodology proposed in this paper involves four main processes:

1 – Data collection process associated with data cleaning (outliers identification and removal), using a common approach based on a dedicated data transformation;
2 – Data discretization process into predefined classes with the goal of data uniformization among different data collection conditions;
3 – Data discretization into sub-classes;
4 – KD based on a DM process.

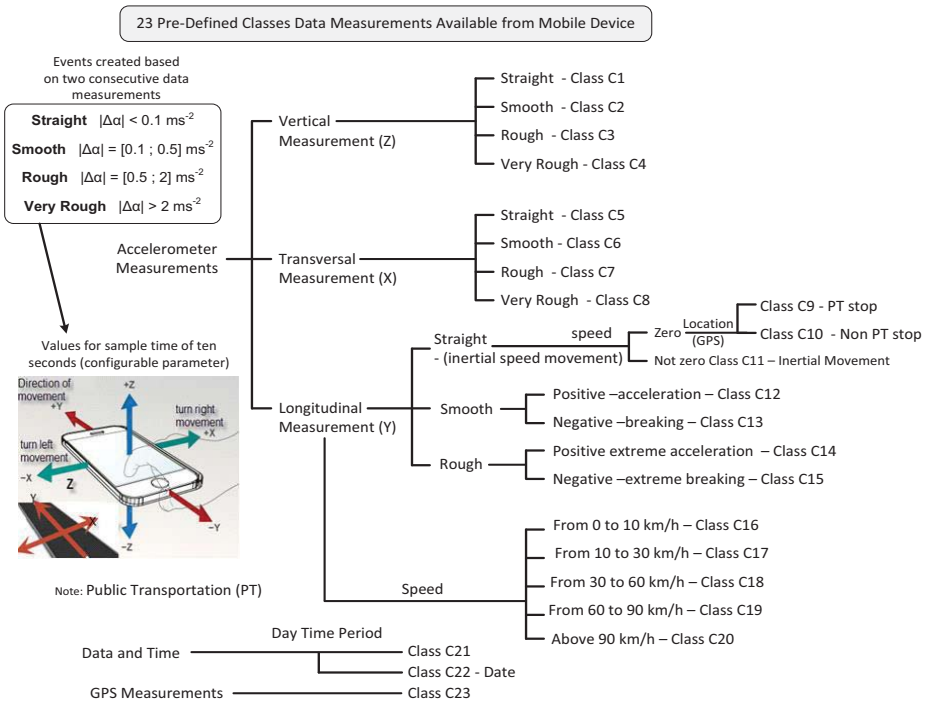

Fig. 1. Information about the creation of predefined classes and sub-classes based on a discretization process, with details for mobile device sensor data events creation based on 23 predefined data classes.

The first process, which varies from case to case, is responsible for the collection of huge amounts of data (big data). To show the application of the proposed methodology we use data from mobile device sensors. For testing purposes, we use data from 50 Lisbon area users, in a period corresponding to the first six months of 2013. The first phase involves the identification of outliers to reduce the number of records. Inconsistent data is also removed during this phase [14], where we remove distant data points based on statistical measurements. Process 2 consists on data transformation into predefined classes. This is a process that is specific to each case study. Taking into account the mobile device sensor data, Fig. 1 shows the 23 predefined classes (C1 to C23). These classes are created based on accelerometer measurements in the three orientation axis, as well as additional GPS data. The accelerometers' data is divided into three dimensions: Z (vertical) is the upright direction; Y (longitudinal) is the direction of movement and X (transversal) is the horizontal direction. The transformation process to merge original data into these predefined classes uses information of two consecutive accelerometer data measurements, where the data value difference is classified into four scales (as shown in the left box of Fig. 1). This number of scales (four) is a compromise between diversity and complexity, as selected based on an analysis of several tested

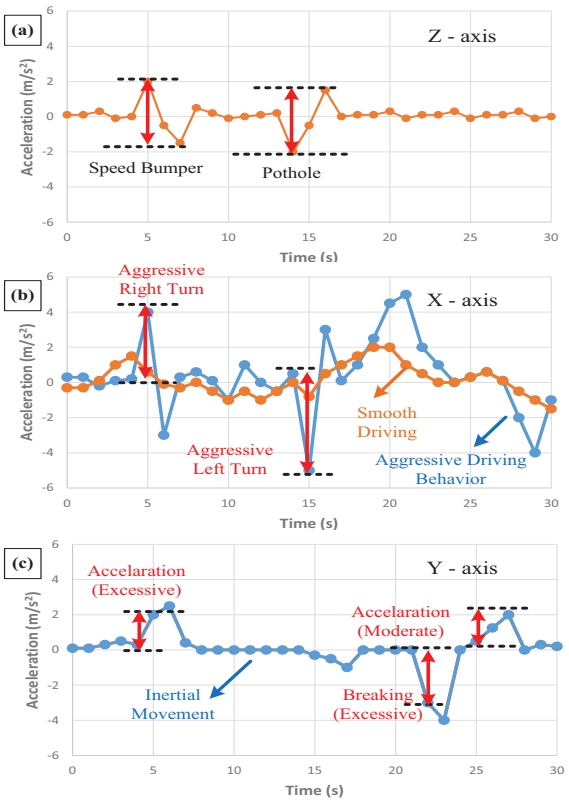

Fig. 2. Accelerometer data transformation process from collected data to predefined classes: (a) *C1* to *C4*, based on X-axis accelerometer measurements; (b) *C5* to *C8*, based on X-axis accelerometer measurements (data from an aggressive driver and a smooth driver); (c) *C9* to *C15*, based on Y-axis accelerometer measurements.

cases (ranging from two to nine scales). Fig. 2 shows the creation of data events (measurements performed by the tracking device or application in a sampling period). These data events can be allocated to a specific route or not. All of these data events are allocated into predefined classes. Classes *C1* to *C4* are based on accelerometer data on the Z-axis. This data can be used to identify, together with the GPS coordinates, the position of potholes and speed bumps. Classes *C5* to *C8* are based on accelerometer data on the X-axis. Fig. 2 (b) shows data from an aggressive driver, with aggressive left and right turns. For the Y direction (Fig. 2 (c)), we divide these data events into thirteen classes: *C9* when the speed is close to zero and the GPS coordinates match a public transportation stop; *C10* when the speed is close to zero and the GPS coordinates do not match a public transportation stop; *C11* when the speed is different from zero, which means inertial movement. The other events are used to identify moderate or aggressive acceleration and braking events. Speed values are taken from:

$$V_y[k+1] = V_y[k] + ta_y[k]\,, \tag{1}$$

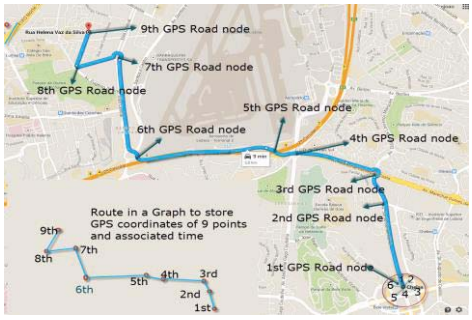

Fig. 3. Simplified process to generate route paths in geo-referenced graphs.

where, t is the elapsed time between samples, k is the sampling time and a_y is the measurement of acceleration in the Y-axis. The speed data is transformed into classes *C9*, *C10*, *C11* and *C16* to *C20*.

GPS data is also transformed in this process to route graph data, due to our interest in the route trajectories. This process eliminates the error associated to the GPS measurements and allows the saving of storage space. Each road is represented by its corners, and we match the GPS coordinates against the nearest road corner. As shown in Fig. 3, for the first GPS data collected we need to look for the nearest road corner located inside a circle centered on the current GPS position. We calculate the distances to these six corners (1 to 6) and choose the smallest one, which in this case is the corner number 1. This process is repeated several times, forming a route. After that, we use this route in a graph to perform the matching against other routes. The developed system adds more GPS nodes than those shown in Fig. 3, because each road corner generates a node, but for simplification purposes we show only nodes from the main roads. With this process, we can store route trajectories in a graph for future processing, where a circular distance function with a predefined radius is used from the current user position to route the graph trajectory.

Process 3 consists in the class discretization into sub-classes. This step can be applied to continuous or discrete data with the aim of dividing the collected data into sub-classes. Due to the diversity of application cases we present multiple options to the user so he may decide the best one to apply. This discretization process can be based on Heuristics, applied based on [15]: (1) Predefined criteria; (2) Equal area clustering method; (3) Data population division based on percentage.

Process 4 is oriented to KD based on the predefined sub-classes, to use a common approach developed as a tool for easy usage. Since the classes and related sub-classes are similar from case to case, the same approach is applied to different cases of knowledge extraction. We implement a NB algorithm that we use to extract different knowledge based on big data allocated in our predefined class. In both cases, the same NB approach is applied using the predefined class and sub-classes. This tool was developed on top of Microsoft SQL Server 2008R2 with SSIS (SQL Server Integration Services), SSAS (Microsoft SQL Server Analysis Services) and SEMMA (Sample, Explore, Modify, Model and Assess), but others platforms are possible, such as the following open source platforms: Rapid Miner [16], Weka [17] or R [18].

3 Application Based on KD from Mobile Device Data

Mobile devices are a rich source of mobility data because users carry them all the time. In this section we present the developed applications based on this data acquisition process and the proposed KD methodology. Technical development details are omitted because we are interested in the idea expressed by the proposed methodology, the usage of a common approach and the reuse of processes:

1) **User Mobility Patterns:** Other knowledge can be extracted from this mobile device sensor data, like the information of where persons spend their time at several distinct locations throughout the day: home, work, shopping centers, restaurants, etc. From the data that we have collected, we are particularly interested in identifying locations where people spend a great deal of time, and associate these locations with information about the environment obtained from geographic information system data sources. All GPS data is stored in a user mobility profile, in a cloud database, with the information about time and routes (XML graph with time and GPS coordinates). It is possible to present the route representation for that month with associated information of the transportation mode, the number of times the route was performed, and also the temporal periods. Thus, it is possible to represent the time that a user spent in a month in these locations. This information represents the user mobility activity captured by the mobile device sensors.
2) **Traffic Information - My Traffic Info App**: Based on the speed and location information (GPS data), it is possible to identify traffic situations as conditions that occur on road networks as the number of vehicles increase and are characterized by slower speeds and longer trip times. From the information of a current road (matching of position against available routes), it is possible to check traffic situations using the relation of current speed divided by the maximum road speed (we call this ratio *Vt*). The system tries to classify this traffic into four classes: (1) Green if the user's average *Vt* is above 0.25; (2) Yellow for average *Vt* between 0.1 and 0.25; (3) Red when the average *Vt* is below 0.1; (4) Black when average *Vt* is zero. We disregard $Vt = 0$ at public transportation stops if transportation mode is a public transportation, as well as at road intersections, because we assume (for simplification purposes) the existence of a traffic light on each road intersection. To avoid this we would need the GPS coordinates of the traffic lights, which are taken from road graph information. There may be times where the number of online users is small, which results in a reduction of information in the user database. Thus, it is necessary to use external information from traffic web services. The current average speed information of the users in a given route is used as input in the route advisor. Taking into account the traffic KD and since we have all users' positions in a graph, it is possible, based on a Dijkstra's algorithm [19], to propose alternative routes in a real-time approach, where we minimize the time between all possible graph nodes to reach the final destination. Since we stored past user routes based on the current user position, it is possible to generate personalized traffic alerts when traffic happened on users' past taken routes. We illustrate the app usage with a small example: The user A in the morning period always follows a certain route to go to work. As he starts the driving process, the system knows his position by GPS information,

and a traffic situation is detected three blocks away in his usual route. The app alerts the user on his mobile device, and the application shows an alternative route. We are aware of the danger of the use of a mobile device while driving, but this problem can be overcome with upcoming device interoperability standards that offer integration between a Smartphone and a car's infotainment system, such as MirrorLink [20]. The result is My Traffic Info, is an Android App that uses traffic KD and the information about a past route driven, in order to present traffic alerts to the driver and, based on a developed real-time route planner [19], propose an alternative route.

3) **Road Classification Application and Potholes Identification:** Accelerometer vertical measurement (Z-axis) can give out important information about road conditions. It is possible to identify potholes and, with GPS coordinates, mark them in a road map. This application ensures road surface quality assessment in a continuous monitoring process, and can be important information for road authorities and drivers. Road quality can be used as an input parameter of router planning. Basically, this detection process uses Z data from accelerometer (see Fig 2 (a)) and looks for an initial decrease of this Z-acceleration of more than 1.5 ms-2 (this value was selected based on the study of several cases) followed by an increase. It is possible to classify the severity of a pothole based on this Z-axis acceleration value and duration. When these events are detected, the GPS coordinates are stored for geographic representation. It is possible to identify the road side associated to the pothole based on the movement direction, and the GPS measurement error is reduced because we use data from multiple users. This application classifies the roads based on predefined criteria, like the number of potholes, and these are divided into three severity scales (small, medium and big) using the Z-axis accelerometer difference between two consecutive measurements.

4 Conclusions

Tracking user activity and associated mobility is available at low costs through the possibility of using mobile device sensor information. The data generated is huge and has great impact in the study of user mobility habits. In this research we show different applications of this big data tracking using a common mining approach for knowledge discovery (KD), which can be used for specific applications, based on predefined classes and sub-classes. The proposed work methodology provides a bridge between field experts in data collection and data mining (DM). This framework needs to be improved with more case experiences, but it is one of the first steps towards the establishing of a semi-automatic approach to KD in big data and a mashup approach for intelligent transportation systems web applications. Another important issue is the connection among these applications and others, in order to share knowledge and data. User tracking data is important for public transportation planning, and even for advertisement applications, because the information can be automatically personalized based on location and time. Other DM algorithm approaches can be applied, but the Naïve Bayes (NB) classifier has a simplified common approach application. The proposed methodology can be applied to different big data sources. Another important issue is the diversity of mobility

applications that can be easily developed using all this mobile device data. One example is a mobility invoice system, as a tool for citizens to be aware of their carbon footprint impacts, along with associated measures or suggestions to reduce this invoice. The passive tracking of data from citizens (with a considerable number of users) could generate useful data about mobility habits and be used to improve citizens' mobility.

5 References

1. G. Mariscal, Ó. Marbán, C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," The Knowledge Engineering Review, vol.25:2, n.25, June 2010.
2. Sholom M. Weiss, Nitin Indurkhya, "Predictive Data Mining: A Practical Guide (The Morgan Kaufmann Series in Data Management Systems)", MK Publ., San Francisco CA: Ed.1, Aug. 1997.
3. J. Li, L. Wong, Q. Yang "DM in Bioinformatics," IEEE Intelligent System, IEEE CS, 2005.
4. M. Baldi, E. Baralis, F. Risso, "Data mining techniques for effective and scalable traffic analysis," IEEE Int. Symposium on Integrated Network Management, pp.105-118, 2005.
5. Manuel Fogue, P. Garrido, Francisco J. Martinez, Juan-Carlos Cano, Carlos T. Calafate, P. Manzoni, "Using data mining and vehicular networks to estimate the severity of traffic accidents," In Manag.t Intelligent Systems, Springer Berlin Heidelberg, vol.171, pp.37-46, 2012.
6. Uwe Reiter, "Modeling the driving behaviour influenced by information technologies," Inter. Symposium on Highway Capacity. Highway capacity and level of service, pp.309-320, 1991.
7. João C. Ferreira, Vítor Monteiro, João L. Afonso, "Dynamic Range Prediction for an Electric Vehicle," EVS27 Int. Electric Vehicle Symposium & Exhibition, Barcelona Spain, 2013.
8. Wazir Zada Khan, Yang Xiang, Mohammed Y Aalsalem, Quratulain Arshad "Mobile phone sensing systems: A survey," Communications Surveys & Tutorials,v.15, pp.402-427, Feb13.
9. P. Turaga, R. C. Pavan, V. S. Subrahmanian, O. Udrea "Machine recognition of human activities: A survey," IEEE Trans. On Circuits and Systems for Video Technology, vol.18, no.11,Nov. 2008.
10. S.Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, M. Srivastava, "Using Mobile Phones to Determine Transportation Modes," ACM Transactions on Sensor Networks, vol.6, no.2, Feb. 2010.
11. Donald J. Patterson, Lin Liao, Dieter Fox, and Henry Kautz, "Inferring High-Level Behavior from Low-Level Sensors," UbiComp 2003: Ubiquitous Computing, v.2864, pp.73-89, 2003.
12. Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, Wei-Ying Ma, "Understanding mobility based on GPS data," International Conf. on Ubiquitous Computing pp.312-321, Sept. 2008.
13. Jamie MacLennan, ZhaoHui Tang, Bogdan Crivat, "Data mining with Microsoft SQL server 2008," Wiley Publishing, Inc., Indianapolis, Indiana, 2009.
14. Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy, "Advances in Knowledge Discovery and Data Mining," AA for Artificial Intelligence, USA, 1996.
15. José de Almeida, João C. Ferreira, "BUS Public Transportation System Fuel Efficiency Patterns," Int. Conf. on Machine Learning and Computer Science, Kuala Lumpur Malaysia, pp.4-8, 2013.
16. Markus Hofmann, R. Klinkenberg, "RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC DM & Knowledge Discovery Series)", CRC Press, 2013.
17. Waikato ML Group, "User Manual Weka: The Waikato Environment for Knowledge Analysis," Department of Computer Science, University of Waikato (New Zealand), June 1997.
18. Andy Bunn, Mikko Korpela, "R: A language and environment for statistical computing," R Foundation for Statistical Computing. Vienna Austria, [Online]. Available: http://www.R-project.org.
19. João C. Ferreira, "Green Route Planner,"Springer, Nonlinear Maps and their Applications," Selected Contributions from the NOMA 2011 Int. Workshop, vol.57, pp.59-87, 2013.
20. Fabian Huger, "User interface transfer for driver information systems: a survey and an improved approach," Int. Conf. on Automotive User Interfaces & Interactive Vehicular Applications, 2011.