Data science in sensing machine generated data

Ana (Resulaj) Ktona
University of Tirana
ana.ktona@fshn.edu.al

Inva Bilo
University of Gjirokastra
ibilo@uogj.edu.al

Denada Xhaja
University of Tirana
denada.xhaja@fshn.edu.al

Xheni Melo
University of Tirana
xheni.melo@fshn.edu.al

Abstract

The increasing recent advances in hardware technology for mobile technology and sensor processing has resulted in greater availability of sensor generated data. For example, mobile devices contain many sensors such as GPS, accelerometers, gyroscope, magnetometer, thermometer, etc., which produce large volumes of data over time. This has lead to a need for principled methods for efficient sensor generated data processing. In this paper, we describe the application of data mining techniques in a case study of identifying patterns to m-health sensor generated data. These technique will be used to build a model for outlier analysis, pattern analysis, and prediction analysis.

Keywords: sensor, Internet of Things, big data, data processing.

1. Introduction

In recent years humans have much more interactions with things because modern devices contain more sensors than ever [ThKS14]. The addition of these sensors into everyday devices has becoming particularly apparent when reviewing the rate of global sales, because such devices are not constrained only to developed economies.

Sensors embed on devices we use help us monitoring almost every area of our lives through applications such as: healthcare, economy, telecommunication, etc. It is also important to note that the cost of sensors has been reduced considerably in recent years, which has made the process of collecting data easier. Nowadays people use these devices in their daily activities, even for most of them has become an inevitable routine. Since that there is an increasing awareness in physical and mental health, it has become much easier to monitor many health parameters through sensors embed on Smartphone-s or other related devices. It remains now that all this generated information, to be processed and to extract from them valuable information. In this paper, we describe the application of data mining techniques in a case study of identifying patterns to m-health sensor generated data. These technique will be used to build a model for outlier analysis, pattern analysis, and prediction analysis.

2. The Impact of the Internet of Things on Big Data

Big data existed before the Internet of Things and the Internet of Things is not the only source of big data. But, what is the impact of IoT on big data? This is seen first in the storage of the data. The Internet of Things and cloud storage make it easier to store the large amounts of data that flow into companies every day. IoT is also a source of data generation. The connected devices and sensors are responsible for collecting data, and that data joins other data to grow the amount of big data available to companies. Every day, sensors embedded into connected devices are gathering data and transmitting that data to central servers, which assist companies in making decisions. The Internet of Things (IoT) has been a major influence on the Big Data landscape. Now that millions of devices are connected and generate enormous volumes of data, should be considered the efficiency of data collection mechanism.

First, companies need to hire highly efficient data collection mechanisms. *Second*, companies are facing many security issues which are probably not

addressing with traditional ones. *Third*, not all data generated by these devices is useful. *Last*, IoT Big Data is changing our everyday lives at a fundamental level.

3. Sensor Data Mining and Processing

The enormous volume of data produced and transmitted from sensing devices is considered a big data challenge. Sensor generated data brings great challenges especially in the processing phase, because very often is needed real-time processing of a large volume of uncertain data. To deal with that, sensor data analytics is a growing field.

The large volumes of sensor data necessitate the design of efficient algorithms which require at most one scan of the data (known as data stream mining algorithms). A main characteristic of IoT data (sensor generated data) is the distributed storage, making thus data mining a challenge task. Quantity and quality of such data does not have the same rhythm; there is big quantity but low quality of data coming from heterogeneous sources. We have to deal with this variety and noise in data which makes it difficult to find and correct any errors. There is a need for modification of data mining algorithms to suit big data.

It is much easier to create than to analyze data. Data mining methods such as clustering, classification, frequent pattern mining, and outlier detection are often applied to sensor generated data in order to extract patterns from them [TLCY14]. This data usually needs to be filtered for more effective analysis. The challenge is that traditional mining algorithms are often not designed for real time processing methods. Therefore, new algorithms for sensor generated data processing need to perform the analytics in real time in order to make IoT more intelligent, thus providing smarter services.

4. Pattern Analysis

This paper focuses on the data generated by sensors in mobile-Health field. Such sensors devices are related to a server and are the source for big data. Afterwards, there is a need for extracting patterns from these data through different mining techniques [BS15, MPPFM10, DR13, RKDT14, BYX10,MS10]. There exist many international companies that have developed applications for tracking physical activity throughout the world, among which are: Google Fitbit, Apple HealthKit, Samsung SAMI, etc. The difficulty is on availability of such data, since most of them are confidential information and therefore cannot be publicly available. In this paper we have analyzed two realistic datasets: **(1)** Heterogeneity Dataset for Human Activity Recognition [SBBPKDSJ15]; and **(2)** PAMAP2 dataset (Physical Activity Monitoring) [SS12, RS12], both of them are publicly available for the research community. The first dataset is a dataset devised to benchmark human activity recognition algorithms (classification, automatic data segmentation, feature extraction, etc) containing heterogeneous sensors; while PAMAP2 dataset provides a good basis to develop and evaluate data processing and classification techniques for physical activity monitoring. Data mining algorithms applied on the two dataset are: J48 (C4.5) and Naive Bayes. Data set is divided into two parts: training set and testing set. By the application of these data mining algorithms is seen how these sensor generated data are classified, and their generated errors respectively.

4.1. Heterogeneity Dataset for Human Activity Recognition

The dataset contains the readings of two motion sensors[1] commonly found in smart-phones, recorded while nine users executed activities scripted in no specific order carrying smart-watches and smart-phones. Activities performed by users are: biking, sitting, standing, walking, stair up and stair down. Dataset contains 10 attributes together with the activity performed by users (what we are going to predict). Attributes taken into account for analysis are arrival time, correlation time, axes X, axes Y, axes Z.

4.2. PAMAP2 dataset

The PAMAP dataset contains data from 24 activities[2] and 9 subjects, wearing three IMUs (inertial measurement units) and a HR-monitor. The dataset contain 54 attributes and the one we are going to predict is what activity users do based on other parameters. These activities are: lying, sitting, standing, walking, running, cycling, Nordic walking, watching TV, computer work, car driving, ascending stairs, descending stairs, vacuum cleaning, ironing, folding laundry, house cleaning, playing soccer, rope jumping.

[1] Sensors used to gather activity data are gyroscope and accelerometer.

[2] everyday household and sport activities.

4.3. Data Mining Process

The Data Mining process consisted of the following steps:

1-*Detection of adequate data*: This involved the finding of the sensor generated data that may be relevant to this study for the data mining process.
2-*Data Pre-Processing*: The collected data was transformed in order to be processed by the data mining algorithms: some unnecessary columns and rows were removed from the data set according to data mining best practices. This resulted in a data set of 938086 rows with 6 columns of Heterogeneity Dataset for Human Activity Recognition; and 249957 rows with 53 columns of PAMAP2 Dataset.
3-*Definition of Training Set*: Classifiers were independently trained for two datasets.
4-*Algorithms Selection*: The selected algorithms are J48 (C4.5) and Naive Bayes.
5-*Training*: Classifiers were produced by training the J48 and Naive Bayes data mining algorithms on the historical sensor data.
6-*Evaluation:* Classifiers were evaluated using training set and percentage split train/test set. The performance metrics produced for each classifier include Correctly Classified Instances (CCI) and Root mean squared error (RMSE).

5. Results and discussions

Classifiers were independently trained using training data and percentage split train/test set for two datasets. Below are the tables with regarding performance metrics (CCI - Correctly Classified Instances, RMSE - Root Mean Squared Error) of applied algorithms for each of datasets.

Table 1 : Evaluation of Heterogeneity Dataset for Human Activity Recognition

	J48		Naive Bayes	
	CCI (%)	**RMSE**	**CCI (%)**	**RMSE**
Using training data	*84,22*	*0,1905*	*66,1*	*0,2689*
Using percentage split (% of training set)	77.15 (70%)	0.276	66.30 (70%)	0.2701

Table 2 : Evaluation of PAMAP2 Dataset

	J48		Naive Bayes	
	CCI (%)	**RMSE**	**CCI (%)**	**RMSE**
Using training data	*99.99*	*0.0033*	*96.84*	*0.0483*
Using percentage split (% of training set)	*99.93 (70%)*	0.0079	*96.88 (66%)*	*0.0479*

Experiments indicated that J48 (C4.5) algorithm produces best results comparing to Naive Bayes algorithm for two datasets and sensor generated data in PAMAP2 Dataset serves better for physical activity recognition techniques.

6. Conclusions

Nowadays people use sensor devices in their daily activities and the enormous volume of data produced and transmitted from these devices is considered a big data challenge. The implementation of Data Mining Techniques and Internet of Things in healthcare will provide good health conditions to people without the necessary presence of doctors and will influence motivation for change in physical activity behavior.

In this paper, we presented two case studies on applying data mining algorithms on historical sensor data (realistic data) in order to evaluate data processing and classification techniques for human physical activity monitoring.

References

[ThKS14] B.Thirunavukarasu, T.Kalaikumaran, S.Karthik. Integration of Data Mining and Internet of Things – Improved Athlete Performance And Health Care System. *International Journal of Technical Research and Applications* e-ISSN: 2320-8163, Special Issue 11, 28-31, Nov-Dec 2014.

[TLCY14] Ch.W.Tsai, Ch.F.Lai, M.Ch.Chiang, L.T. Yang. Data Mining for Internet of Things: A Survey. *IEEE Communications Surveys & Tutorials*, Vol. 16, No. 1, First Quarter 2014.

[BS15] Sh.Bhatia, S.Patel. Analysis on different Data mining Techniques and algorithms used in IOT. *Int. Journal of Engineering Research and Applications*, ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 1), 82-85, November 2015.

[MPPFM10] Alexandra Moraru, Marko Pesko, Maria Porcius, Carolina Fortuna, Dunja Mladenic. Using Machine Learning on Sensor Data. *Journal of Computing and Information Technology - CIT* 341–347, 18, 2010, 4.

[DR13] Ch.Dule, K.M.Rajasekharaiah. Page Sensor Data Mining Model and System Design: A Review. *International Refereed Journal of Engineering and Science (IRJES)* ISSN (Online) 2319-183X, (Print) 2319-1821 Volume 2, Issue 6,), 16-22, June 2013.

[RKDT14] A.Rook, A.Knauss, D.Damian, A.Thomo. A Case Study of Applying Data Mining to Sensor Data for Contextual Requirements Analysis. 978-1-4799-6355-3/14, *IEEE AIRE* 2014, Karlskrona, Sweden, 2014.

[BYX10] Sh.Bin, L.Yuan, W.Xiaoyi. Research on Data Mining Models for the Internet of Things. *IEEE* 978-1-4244-5555-3/10.

[MS10] A.Mannini, A.M.Sabatini. Machine Learning Methods for Classifying Human Physical Activity from On-Body Accelerometers. *Sensors* 2010, 10, 1154-1175; doi:10.3390/s100201154.

[SBBPKDSJ15] A.Stisen, H. Blunck, S.Bhattacharya, Th.S. Prentow, M.B.Kjærgaard, A.Dey, T.Sonne, M.M.Jensen . Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition. *In Proc. 13th ACM Conference on Embedded Networked Sensor Systems* (SenSys 2015), Seoul, Korea, 2015.

[SS12] R. Stricker, D. Stricker. Introducing a New Benchmarked Dataset for Activity Monitoring. *The 16th IEEE International Symposium on Wearable Computers (ISWC)*, 2012.

{RS12] A. Reiss, D. Stricker. Creating and Benchmarking a New Dataset for Physical Activity Monitoring. *The 5th Workshop on Affect and Behaviour Related Assistance (ABRA)*, 2012.