Data Mining in IoT

Data analysis for a new paradigm on the Internet

Peter Wlodarczak
University of Southern Queensland
Faculty of Business, Education, Law and Arts (BELA)
Toowoomba, Qld
wlodarczak@gmail.com

Mustafa Ally
University of Southern Queensland
Faculty of Business, Education, Law and Arts (BELA)
Toowoomba, Qld
Mustafa.Ally@usq.edu.au

Jeffrey Soar
University of Southern Queensland
Faculty of Business, Education, Law and Arts (BELA)
Toowoomba, Qld
Jeffrey.Soar@usq.edu.au

ABSTRACT

This paper provides an overview on Data Mining (DM) technologies for the Internet of Things (IoT). IoT has become an active area of research, since IoT promises among other to improve quality of live and safety in Smart Cities, to make resource supply and waste management more efficient, and optimize traffic. DM is highly domain specific and depends on what is being mined for. For instance, if IoT is used to optimize traffic in a Smart City to reduce traffic jams and to find parking spaces quicker, different types of data needs to be collected and analysed from an eHealth solution, where IoT is used in a Smart Home to monitor the well being of patients or elderly people. IoT connects things that can collect numeric data from smart sensors, streaming data from cameras or route information on maps. Depending on the type of data, different techniques need to be adopted to analyse them. Also, many IoT applications analyse data from different devices and correlate them to make predictions about possible machine failures in production sites or looming emergency situations in Smart Buildings in a home security application. DM techniques need to handle the heterogeneity of IoT data, the large volumes of data and the speed at which they are produced. This paper explores the state of the art DM techniques for IoT.

KEYWORDS

Internet of Things, Data Mining, Machine Learning, Predictive analytics, Smart City

ACM Reference format:
Peter Wlodarczak, Mustafa Ally, and Jeffrey Soar. 2017. Data Mining in IoT. In *Proceedings of 2nd Int. Workshop on Knowledge Management of Web Social Media, Leipzig, Germany, August 2017 (KMWSM '17),* 4 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

The Internet of Things (IoT) and Cloud computing are core technologies that enable new paradigms such as Smart Cities or Smart Homes. The goal of IoT is to connect all objects on the basis of networked individuals to form a ubiquitous network, which is called the Internet of Tings [16]. IoT forms a Mobile Ad hoc Network (MANET) of connected things. Things in terms of IoT are everyday objects such as sensors, actuators, Internet enabled mobile devices, cars, household appliances or wearable devices. They provide ubiquitous sensing enabled by Wireless Sensor Network (WSN) technologies [6]. IoT cannot only connect physical devices but also services or other data sources to form the Internet of Everything (IoE).

Cloud computing provides shared computing resources and data on demand, typically in a distributed environment. Cloud computing relies on multiple data centers that span multiple domains and geographical areas [7]. Smart Cities and digital cities retrieve information in a collaborative environment and store it to the Internet cloud [11]. Smart Cities and Smart Homes promise to improve living conditions, safety, optimize resource supply and waste disposal, optimize traffic and economic, social and cultural development [6]. A Smart City contains a set of integrated services, a bunch of applications, and a large number of things [12].

The Internet of Things integrates multiple wired and wireless communication, control, and IT technologies, which connect various terminals or subsystems under a unified management platform that employs open and standardized data presentation technologies such as XML/Web Services/SOA [16]. The IoT building blocks can be represented in several layers to describe its functionality. The bottom layer is the sensing layer where the physical devices are located. They are connected directly to each other and the Internet to form the sensor network layer. Often there is an intermediate layer, sometimes called fog or edge computing layer. Edge computing is located at the edge of the sensor network and optimizes the cloud infrastructure by processing data near the source, i. e. close to the sensors. Data can be pre-processed and filtered in order to save bandwidth and processing power. Only relevant data is then transmitted to the cloud, where it is stored, analyzed and visualized. Figure 1 summarizes the IoT layers and its functionality.

IoT collects a lot of data from sensors, smart-phones, wearables or other Internet enabled devices and stores it in the cloud. To turn data into actionable knowledge, it has to be analyzed using suitable Data Mining techniques. For instance, sensor data from a Smart Home is used for security monitoring or for home automation for elderly or disabled people, or traffic data is analyzed to calculate an optimal route for an ambulance.

IoT data is typically heterogeneous, generated at high speed and in large volumes, and needs to be analyzed in real-time. Data can be for instance continuous numbers form a temperature sensor, streaming data from cameras or text. Traditional DM techniques

WI '17, August 23-26, 2017, Leipzig, Germany

ACM ISBN 978-1-4503-4951-2/17/08…$15.00
https://doi.org/10.1145/3106426.3115866

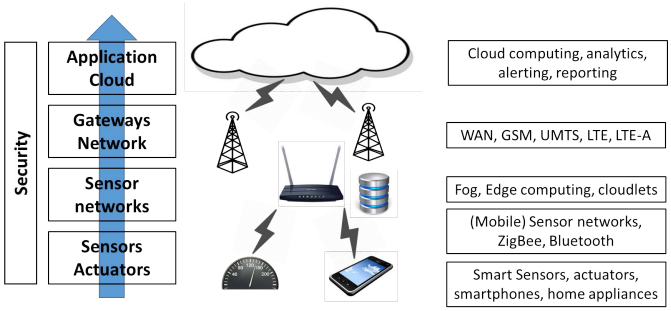

Figure 1: IoT layers

do not suffice to analyze the data generated by IoT. The data has to be pre-processed and correlated in a suitable way to get meaningful results. This paper explores the current state of the art DM techniques applicable for IoT.

2 DATA MINING TECHNOLOGIES FOR IOT

Extracting useful information from a complex sensing environment at different spatial and temporal resolutions is a challenging research problem in artificial intelligence [6]. To detect useful patterns in IoT data, the data needs to be analyzed with suitable DM techniques. DM is highly domain specific [14]. Whether the IoT platform has to predict optimal traffic routes or detect a machine that is about to fail and needs maintenance, different methods might apply. For instance, a predictive maintenance application that needs to detect machine failure before it occurs so it can be replaced before production is interrupted, collects and analyses sensor data from machines such as temperature, torsion or attrition. A home security application might use movement detectors and camera data to detect possible intruders. Both systems use different types of data and make predictions about different things. Also, IoT has some characteristics that influence the methods with which IoT data is analyzed:

- New devices can be added ad hoc to an IoT solution. This means that new data sources need to be analyzed, possibly in new data formats. For instance, an eHealth application can measure blood pressure and glucose levels. A patient using the solution has a new fitness tracker that can add fitness data to the eHealth solution.
- Devices might stop sending data. For instance, a car is driving into a tunnel and loses the GPS signal.
- A sensor might stop sending data at all because the battery is empty or the wireless communication is interrupted.
- A sensor or actuator can be part of several applications. For instance, a movement detector can be used to open an automatic door and to detect unauthorized intruders. If the detector fails, several applications are affected.
- An IoT application might have to show a different behavior in different situations. For instance, a home security system has to be able to differentiate between day and night, since during the day a greater number of human activities is detected, at night hardly any.

The ramifications are deviations in normal data traffic that might be interpreted as anomaly and the IoT solution might issue a false alert event. DM techniques for IoT need to be able to adapt to dynamic environments or changed data streams to avoid redesign of the DM rules each time a sensor is added or removed. Machine Learning (ML) techniques are well suited to handle the fuzziness in data streams and can adapt quickly when the environment changes. ML is a branch of Artificial Intelligence (AI) and aims to imitate human learning on computers without the need to be explicitly programmed.
ML techniques have several characteristics that make them favorable for IoT DM:

- ML techniques learn the DM rules from historic data so there, there is no need for a developer to program them manually
- ML methods can continue to learn new rules, for instance if a new smart device is added
- Many ML schemes calculate probabilities, which makes them robust against small changes in the data flow. For instance, when a device stops sending data and there are still others sending values, the probability changes only slightly and no false positive is issued.

Data mining goes through several steps. They are divided into a data conditioning or data preprocessing phase and a predictive analysis phase [14]. In the data conditioning phase, data is collected and preprocessed. Not all data is useful for a specific DM task so selecting the observation points is an important preprocessing step. Other preprocessing steps include data transformation to have consistent data formats, data deduplication and outlier removal. In the predictive analysis phase, suitable DM methods have to be selected and trained. Depending on the problem, the data has to be correlated since data from only one data source might not be enough to make meaningful predictions. Often data is also visualized and reports are generated. DM is highly iterative and some steps might be run through many times. Figure 2 summarizes the DM steps:

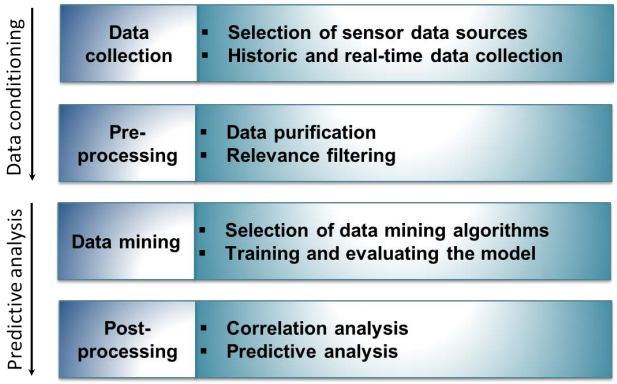

Figure 2: IoT Data Mining steps

Data collection

Data collection in IoT happens at the device or "thing" level. Smart sensors, smart phones or tablets measure different values from the environment and typically transmit them to an IoT cloud platform

for analysis and storage. Some devices produce potentially a lot of data and transmitting all data to the IoT cloud might not be practicable. To save bandwidth and computing power, often the data is stored and preprocessed in the fog by edge computers. Edge computing often also (pre-)analyses data and only some observation points are transmitted to the cloud.

Data pre-processing

Real life data is seldom in a format suitable for DM algorithms and is often poor in quality [13]. Data cleansing is thus a crucial step to get good results. Different sensors collect data in different formats. Data transformation is an important step to harmonize data. Relevance filtering is another important preprocessing step to get good performance of IoT applications. For instance, one application might only require the source and destination coordinates of an itinerary, for another application the whole route might be relevant. Data deduplication, outlier removal, entity resolution and feature selection are some important preprocessing steps. Feature selection will mean selecting the observation points that are used as input for the DM algorithms.

Data Mining

There are many DM techniques. Machine Learning (ML) techniques are adopted when the rules are getting too complex or if there are too many rules to be programmed by a developer. ML imitates human learning. Whereas humans learn from experience, ML algorithms learn the rules from historical data. ML techniques learn from past, historic data, to make predictions about future events. For instance, a predictive maintenance application uses historic sensor data that collects information about the state of a Smart Building to learn rules to predict if the air conditioning system or elevators are going to fail. ML is divided into supervised, semi-supervised and unsupervised learning. Supervised methods are used for classification and regression. They require labeled data for training. Typical supervised learners include Bayesian models, decision tree induction, Support Vector Machines (SVM) and Artificial Neural Networks (ANN). Unsupervised methods are adopted if no labeled data is present and semi-supervised methods are used when a small amount of labeled data and a large amount of unlabeled data is available. ML techniques are well documented in literature and are not explained in detail here.

To make predictions, for instance, about machine failure or a patient becoming symptomatic, the data has to be classified. The machine or patient starts to show deviant values or behavior. Ultimately we want to find a decision function f, that classifies a data set t as normal (N), or deviant (D). If we denote the whole data set by T, we search for a function:

$$f : T \rightarrow \{N, D\} \tag{1}$$

We use a set of randomly selected and pre-classified training observation points $\{(t_1, c_1), (t_2, c_2), ..., (t_n, c_n)\}$, where: $t_i \in T, c_i \in \{D, N\}$. Typically several learners are trained using this data set. The trained learners are then tested against new, unseen data and their performance is measured. During training, a loss function is minimized until it converges. Typical loss functions include the mean squared error or negative loss likelihood. Once the loss function converges, training is completed and the trained learners are evaluated. The best performing learner, i. e. the learner with the highest accuracy, is usually selected to make predictions on real life data.

Predictive analysis

During the predictive analysis phase, data is often correlated. For instance, only blood levels might not be enough for an IoT-based eHealth solution to determine if a patient becomes symptomatic. It has to be correlated with movement data to make reliable predictions about a personâĂŹs health state.

There are many different techniques for correlation analysis. A common technique is time series analysis. Two time series x and y are correlated to determine if time series x has predictive information about time series y. The Granger causality test for two scalar-valued, stationary, and ergodic time series X_t and Y_t is defined as:

$$F(X_t|I_{t\text{-}1}) = F(X_t|I_{t\text{-}1} - Y_{t-Ly}^{Ly}), t = 1, 2... \tag{2}$$

Where $F(X_t|I_{t\text{-}1})$ is the conditional probability distribution of X_t given the bivariate set $I_{t\text{-}1}$ consisting of an L_x-length vector X_t and an L_y-length vector of Y_t.

It should be noted that ML techniques can also be used for correlation and predictive analysis [2]. Many ML schemes calculate probabilities, which makes them suitable for changing IoT environments. They output, for instance, the probability Pr, that a behavior x_i with corresponding label y_i belongs to class j:

$$Pr(x_i|y_i) = j \tag{3}$$

where class j can be *Normal* or *Deviant*.

3 CHALLENGES

One of the biggest challenges in IoT is the lack of standards which makes interoperability of different devices and connecting them to the Internet difficult. Standardization efforts and standard protocols such as Message Queue Telemetry Transport (MQTT) and Advanced Message Queuing Protocol (AMQP), which are lightweight message oriented middleware, have emerged. Nevertheless more data transformations are needed.

Many smart devices have limited resources, limited bandwidth and battery life. Also, mobile network coverage varies in different places which can be problematic especially for eHealth IoT applications. Security remains a major concern in IoT since many devices were not designed with security in mind and they remain vulnerable to cyber-attacks. Due to their limited nature, securing them with encryption and intrusion prevention mechanism remains challenging. Privacy is a major concern, expecially in places like Smart Homes where a great deal of personal data is collected. Privacy preserving DM techniques have been proposed [1], but due to the restrictions of the devices and all the other challenges in IoT DM they are often not applied. Privacy and AI are complex topics and more research in privacy engineering is highly desirable.

4 FUTURE DEVELOPMENT

Technology becomes an integral part of our environment and the volumes and heterogeneity of the data that need to be processed will ever increase. To be useful, IoT data often needs to be analyzed in real-time, e. g. to respond to traffic jams or to optimize energy consumption. The demand for more and faster resources will increase and more solutions will be operated in the cloud. Cloud solutions provide the scalability and high availability that many IoT applications require.

An emerging area in ML is Deep Learning [15]. Contrary to the learners described above, Deep Learners (DL) have multiple layers of abstraction to interpret data. As the signal processes through the layers, the data is represented at a higher abstraction level. This makes DL more robust if a data stream shows deviations from normal streams, which might be interpreted as an anomaly by traditional ML techniques, also called shallow learners. Also some DL can do feature extraction automatically, there is no need for feature engineering. This will lead to IoT solutions that can prioritize and thus learn what is relevant in data and what is not. This allows DL enabled IoT solutions to learn from new situations and react accordingly. For instance, in a traffic situation an IoT enabled car can decide if it should evade or break. DL have been particularly well performing for multimedia mining and natural language processing (NLP) and we will see more IoT solutions that can be operated using voice commands and object and face recognition capabilities.

We will see IoT expand into new areas. IoT is used to form Smart Homes and Smart Cities. Ultimately this will lead to the Smart Planet that will become smarter, more efficient, greener, more economical, and will self-regulate the energy consumption, waste disposal, improve security and overall quality of life.

REFERENCES

[1] Agrawal, R & Srikant, R 2000, 'Privacy-preserving data mining', ACM Sigmod Record, vol. 29, no. 2, pp. 439-50.

[2] Arias, M, Arratia, A & Xuriguera, R 2014, 'Forecasting with twitter data', ACM Trans. Intell. Syst. Technol., vol. 5, no. 1, pp. 1-24.

[3] Cubo, J, Nieto, A & Pimente, E 2014, 'A Cloud-Based Internet of Things Platform for Ambient Assisted Living', Sensors (14248220), vol. 14, no. 8, pp. 14070-105.

[4] Dohr, A, Modre-Opsrian, R, Drobics, M, Hayn, D & Schreier, G 2010, 'The Internet of Things for Ambient Assisted Living', in Information Technology: New Generations (ITNG), 2010 Seventh International Conference on: proceedings of theInformation Technology: New Generations (ITNG), 2010 Seventh International Conference on pp. 804-9.

[5] Gachet, D, de Buenaga, M, Aparicio, F & PadrÃşn, V 2012, 'Integrating internet of things and cloud computing for health services provisioning: The virtual cloud carer project', in Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), pp. 918-21.

[6] Gubbi, J, Buyya, R, Marusic, S & Palaniswami, M 2013, 'Internet of Things (IoT): A vision, architectural elements, and future directions', Future Generation Computer Systems, vol. 29, no. 7, pp. 1645-60, <http://www.sciencedirect.com/science/article/pii/S0167739X13000241>.

[7] Guelzim, T & Obaidat, MS 2016, 'Chapter 12 - Cloud computing systems for smart cities and homes', in Smart Cities and Homes, Morgan Kaufmann, Boston, pp. 241-60.

[8] Islam, SR, Kwak, D, Kabir, MH, Hossain, M & Kwak, K-S 2015, 'The internet of things for health care: a comprehensive survey', IEEE Access, vol. 3, pp. 678-708.

[9] Konstantinidis, EI, Bamparopoulos, G, Billis, A & Bamidis, PD 2015, 'Internet of things for an age-friendly healthcare', in MIE: proceedings of theMIE pp. 587-91.

[10] Pang, Z, Zheng, L, Tian, J, Kao-Walter, S, Dubrova, E & Chen, Q 2015, 'Design of a terminal solution for integration of in-home health care devices and services towards the Internet-of-Things', Enterprise Information Systems, vol. 9, no. 1, pp. 86-116.

[11] Soyturk, M, Muhammad, KN, Avcil, MN, Kantarci, B & Matthews, J 2016, 'Chapter 8 - From vehicular networks to vehicular clouds in smart cities A2 - Obaidat, Mohammad S', in P Nicopolitidis (ed.), Smart Cities and Homes, Morgan Kaufmann, Boston, pp. 149-71.

[12] Tsai, CW, Lai, CF, Chiang, MC & Yang, LT 2014, 'Data Mining for Internet of Things: A Survey', IEEE Communications Surveys & Tutorials, vol. 16, no. 1, pp. 77-97.

[13] Witten, I.H., Frank, E., and Hall, M.A. 2011, âĂŸData MiningâĂŹ, Elsevier, 3 edn.

[14] Wlodarczak, P, Ally, M & Soar, J 2015, 'Data Process and Analysis Technologies of Big Data', in Networking for Big Data, Chapman and Hall/CRC, pp. 103-19.

[15] Wlodarczak, P, Soar, J & Ally, M 2015, 'Multimedia data mining using deep learning', in Digital Information Processing and Communications (ICDIPC), 2015 Fifth International Conference on: proceedings of theDigital Information Processing and Communications (ICDIPC), 2015 Fifth International Conference on IEEE Xplore, Sierre, pp. 190-6.

[16] Zhihao, X & Yongfeng, Z 2010, 'Internet of Things and its future', Huawei Technologies Communicate Beyond Technology, pp. 23-6.

[17] Zhong, N, Ma, J, Huang, R, Liu, J, Yao, Y, Zhang, Y & Chen, J 2016, 'Research challenges and perspectives on Wisdom Web of Things (W2T)', in Wisdom Web of Things, Springer, pp. 3-26.