9

IoT Analytics: From Data Collection to Deployment and Operationalization

John Soldatos and Ioannis T. Christou

Athens Information Technology, Greece

9.1 Operationalizing Data Analytics Using the VITAL Platform

The VITAL smart cities platform has been introduced in an earlier chapter (Chapter 4). It comprises a set of middleware libraries and accompanying tools, which facilitate the development, deployment and operation of smart cities applications, including IoT analytics applications. The platform supports functionalities across all the phases of the IoT analytics lifecycle, which have been presented in the introductory chapter. The rest of this chapter focuses on illustrating the practical implementation of the IoT analytics lifecycle as part of the VITAL internet-of-things (IoT) platform for Smart Cities, which has been already introduced in Chapter 4. Furthermore, it presents practical examples associated with the deployment and operationalization of advanced IoT analytics, over footfall datasets collected from a smart city. IoT Data Collection.

In terms of IoT data collection, VITAL enables the collection of data from heterogeneous IoT systems, notably systems that have been developed and deployed independently in the scope of a smart city. To this end VITAL defines the PPI (Platform Provider Interface) abstract interface, which enables the unification of data from diverse systems in terms of their format. In particular, VITAL enables the collection of IoT data from different systems and data sources as soon as the latter implement the PPI interface.

The VITAL platform provides also the means for managing, registering and de-registering IoT systems in its platform, based on PPIs. Furthermore, the

PPIs enable the data collection and retrieval based on a JSON-LD format, which facilitates the semantic unification of data streams from different systems and data sources. This boosts the application of IoT analytics over diverse IoT systems, through alleviating the semantic heterogeneity of the various streams. Hence, the VITAL platform addresses the variety of IoT data streams both in terms of their formats and in terms of their semantics.

9.1.1 IoT Data Analysis

VITAL supports the storage and processing of the semantically unified data within a datastore, thus facilitating IoT data analytics over data stemming from multiple data sources and systems. The VITAL datastore is supported by a NoSQL database. The VITAL platform offers a wide range of data processing functions over this datastore, including:

- Dynamic data discovery based on criteria such as sensor type and location.
- Filtering on specific data attributes and on the basis of appropriate thresholds for each attribute.
- Complex Event Processing towards producing events based on information contained in multiple IoT streams.

Moreover, VITAL supports the data analysis phase through its integration with libraries of the R project. The latter libraries enable the execution of machine learning schemes, such as regression, classification and clustering.

9.1.2 IoT Data Deployment and Reuse

The VITAL platform enables the deployment of data processing algorithms over semantically unified streams, which are stored in the DMS (Data Management Service) of the system. It also enables the management of registrations to the various IoT data sources (including IoT platforms and systems), which provide the data to the DMS. In this way, VITAL supports the deployment of IoT data and its integration within IoT analytics applications in-line with the third phase of the already presented IoT analytics lifecycle. The integration of IoT data within applications is supported in a way that enables the repurposing and reuse of IoT data across multiple applications. This is made possible on the basis of the semantic annotation of the IoT data streams according to the VITAL JSON-LD contexts.

9.2 Knowledge Extraction and IoT Analytics Operationalization

Based on the functionalities outlined above, VITAL can be used for knowledge extraction, as well as for the deployment and operationalization of IoT analytics. Prior to deploying an IoT analytics application, the discovery and testing of IoT data mining algorithms that are likely to extract the desired knowledge in a credible way is required. In this respect, IoT data mining (which is part of the second phase of the IoT Analytics lifecycle) is very similar to conventional data mining applications i.e. applications leveraging transactional data instead of IoT streams. Hence, mainstream models for data mining and analytics such as the Cross Industry Standard Process for Data Mining (CRISP-DM) Model for Knowledge Discovery [1] can be applied. CRISP-DM entails the following activities and phases:

- **Business Understanding**: This activity is the starting point of the process and refers to the need of understanding the business problem at hand. A sound understanding of the nature of the problem is a key prerequisite to identifying proper machine learning models.
- **Data Understanding**: This activity follows business understanding and aims at understanding the data. By inspecting and understanding the data experienced data scientists can gain valuable insights on the applicability of certain data mining schemes. Data understanding leads to identification of data patterns in the datasets, which can serve as basis for identifying candidate machine learning schemes.
- **Data Preparation**: This is tedious, yet indispensable task in the process, given that the collected datasets need to be transformed in a format appropriate for identifying appropriate data mining and machine learning models. In conventional data mining applications, the data preparation step involves multiple ETL (Extract Transform Load) processes. In the case of IoT analytics, data engineers will have to deal with a multitude of data sources and formats depending on IoT data streams involved. The data preparation process is in several cases tedious, as a result of the need to deal with heterogeneous data sources, formats and semantics. As already outlined, semantic interoperability solutions (such as VITAL) facilitate the data preparation process.
- **Modelling**: This activity leverages data sets collected from the IoT system in order to identify a proper machine learning scheme for the problem

at hand. This task is facilitate by data mining tools (such as RapidMiner[1] and Weka[2]), which can be used to produce a machine learning model (e.g., a classifier or an association) given a training dataset. The modelling phase interacts very closely with the data preparation phase in order to ensure that the available training datasets are appropriate for fitting the target/identified models.

- **Evaluation**: As part of this phase, the produced model is evaluated in terms of its efficiency as the latter is reflected in the speed of training, the speed of model execution, its noise tolerance, as well as its expressiveness and explanatory ability. The evaluation is based on metrics such as classification accuracy, errors in numeric prediction, lift and conviction measures and more. A validation datasets (which is different from the training dataset) is used in the scope of the evaluation process. In case of acceptable performance and accuracy, the data scientists and practitioners can move the model to deployment. However, in case of poor performance, the whole cycle (i.e. from business problem understanding to model evaluation) has to be repeated in order to identify a model that gives satisfactory results for the problem at hand.
- **Deployment**: Successful models (i.e. schemes providing acceptable performance for the business problem at hand) are deployed and operationalized. The VITAL platform and more specifically its development environment offers integration with the R project, as a means of easily programming and deploying IoT analytics schemes. Hence, identified data mining models and schemes (e.g., Bayesian classifiers, K-means clustering algorithms, logical regression schemes) can be flexibly programmed and integrated within an application workflow and accordingly deployed based on the VITAL middleware platform.

In following paragraphs we provide a concrete example of the knowledge extraction process based on IoT data streams.

9.3 A Practical Example based on Footfall Data

As a case-study of performing some useful analysis on sensor data using advanced data mining techniques, we analyze the Camden footfall dataset. The Camden dataset comprises a multi-dimensional time-series in a

[1] https://rapidminer.com/

[2] http://www.cs.waikato.ac.nz/ml/weka/

time-frame of 1 hour, for two months, of counts of people passing in front of 5 different cameras in Camden, London. The first two weeks of this data-set is depicted in Figure 9.1 where its periodic nature is revealed; in the figure, each time-series label InPlusOutX refers to the total number of people detected during the particular time-frame by camera X (in location X); the interested reader can find more information in http://www.spring-board.info/service/service-display/visitor-counting.

By visually inspecting the time-series it is clear that the 4-th location (corresponding to the camera no. 4) is usually the busiest. It is also clear that there are dependencies between all locations (that can be confirmed by computing the R values for any pair of time-series components.) Logistic regression (as implemented in the Weka suite of tools [2]) does not provide much more insight into the nature of the relations between the time-series components. In order to obtain some more insight into the relations between the given time-series components, we have performed Quantitative Association Rule Mining (QARM), introduced in [3], using QARMA, a highly parallel/distributed algorithm for mining all non-dominated "interesting" quantitative association rules in multi-dimensional dataset [4]. Quantitative association rules are association rules defined over quantitative attributes which they qualify over certain intervals. We define a rule to be "interesting" when its support and confidence exceed 8% and 85% correspondingly. The notion of non-dominance in quantitative association rules is formally defined in [4], but intuitively, a rule r dominates a second rule s if whenever s fires (i.e. all its antecedents are satisfied), r also fires, the consequent part of r covers the consequent of s, and r has equal or higher support and confidence than s.

Running QARMA on the Camden dataset produces a total of more than 25.000 non-dominated rules, which entirely cover the dataset: every data point

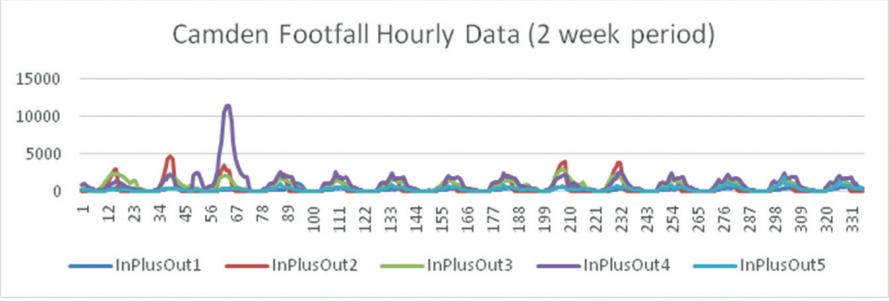

Figure 9.1 Camden footfall dataset.

in the dataset is covered by the application of at least one produced rule. The produced rules have a conviction [4] measure in the range [100.0%, 837.9%] which on average is 437.3%, while the lift of the rules [5] is in the range [1.0, 10.8], and averages at 2.87; both these measures indicate that the produced rules are far from being statistical flukes. Among the most interesting rules found, we list the following:

- Rule1: Time-of-day in [16:00, 19:00] $\longrightarrow$ InPlusOut1 $\geq$ 449 with support = 14.2%, confidence = 85.2%, conviction = 472.7%, lift = 2.82
- Rule2: InPlusOut3 $\geq$ 1874.0 $\wedge$ InPlusOut4 $\geq$ 423.0 $\wedge$ InPlusOut5 $\geq$ 262.0 $\longrightarrow$ InPlusOut2 $\geq$ 2051.0 with support = 8.6%, confidence = 85.1%, conviction = 579.9%, lift = 6.17
- Rule3: InPlusOut1 $\geq$ 265.0 $\wedge$ InPlusOut5 in [327.0, 1791.0] $\wedge$ Max{InPlusOut{1...5}} in [2309.0, 6554.0] $\longrightarrow$ InPlusOut2 $\geq$ 2334.0 with support = 8.9%, confidence = 85.06%, conviction = 603.7%, lift = 8.65
- Rule4: InPlusOut2 $\geq$ 2298.0 $\longrightarrow$ InPlusOut1 $\geq$ 365.0 with support = 8.811475409836065%, confidence = 85.4%, conviction = 415.85%, lift = 2.16

The first rule for example, shows that the camera in location 1 (InPlusOut1) whose average footfall is just under 336, in the 3-hour afternoon period between 16:00–19:00 increases above 449 regardless of the traffic in the other cameras or day of month.

The 2nd rule states that when locations 3, 4, and 5 are above certain footfall thresholds, then it is the second location that becomes the most crowded. This particular rule has among the highest conviction and lift rates, making it a very statistically significant and interesting rule.

The 3rd rule states that when footfall in location 1 is above a certain threshold, location 5 is within certain limits and the maximum of all locations is within certain limits as well, then location 2 exceeds its average value by more than 3 times (the average footfall measured by camera in location 2 is around 706).

Finally, the 4th rule provides an association between the footfall in location 2 and location 1, showing that when the footfall in location 2 is above a threshold that is (very significantly) above its average value, then the footfall in location 1 also increases above its average value. However, this last rule has a lift value of 2.16 and is thus not as strong as the previous three rules.

The produced quantitative rules fully describe the dataset, and show significant associations between the measured values of the time-series components; they also have the extra advantage of showing associations at "corner cases", that is they show what happens in one component when some other components significantly exceed their expected values, in fully quantifiable ways.

Acknowledgement

Part of this work has been carried out in the scope of the VITAL project (www.vital-iot.com), which is co-funded by the European Commission in the scope of the FP7 framework programme (contract No. 608662).

References

[1] Wirth, Rüdiger, and Jochen Hipp. *CRISP-DM: Towards a standard process model for data mining*. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. 2000.

[2] I. H. Witten, E. Frank, M. A. Hall, "Data Mining: Practical Machine Learning Tools & Techniques", Morgan Kaufmann, 3rd edition, Burlington, MA, 2011.

[3] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules", In: Piatetsky-Shapiro G., and Frawley, W. J. (eds), Knowledge Discovery in Databases, AAAI/MIT Press, Menlo Park, CA, 1991.

[4] I. T. Christou, E. Amolochitis, Z.-H. Tan, "QARMA: A Parallel Algorithm for Mining All Quantitative Association Rules and Some of its Applications", Under Review, 2016.

[5] M. Hahsle, B. Grün, K. Hornik: "arules – A computational environment for mining association rules and frequent item sets", Journal of Statistical Software 14(15):1–25, 2005.