Big Data Driven Mobile Cellular Networks: Modelling, Experiments, and Applications

Shangjing Lin, Jianguo Yu and Ji Ma
Beijing Key Laboratory of Work Safety Intelligent Monitoring, Beijing University of Posts and Telecommunications, School of electronic engineering, room 421, 10 XiTuCheng Rd., Haidian District, Beijing, China. Email: linshangjing@bupt.edu.cn

Abstract. The proliferation and pervasive use of mobile devices results in the accumulation of massive amounts of wireless data. Mobile big data can be profitable only if suitable analytics and learning methods are utilized to extract meaningful knowledge and hidden patterns. In this article, we propose a novel mobile big data architecture consisting of five layers: the data storage layer, the data fusion layer, the data security layer, the data analysis layer and the data application layer. The functionality of each layer is presented. We consider one illustrative cases under this architecture, namely, user experience prediction by leveraging machine learning techniques. In practice, mobile big data analytics can be used for network planning and parameter dimensioning to facilitate network design, deployment and operation.

1. Introduction
The technological revolution has facilitated the proliferation and pervasive use of digital devices, such as smartphones, sensors, and the Internet of Things (IoT). Thereby, a massive amount of heterogeneous structured or unstructured data, called mobile big data (MBD), has been generated by those digital devices and carried by mobile cellular networks[1].

Historically, the value of such a great amount of MBD was underestimated until the introduction of big data analytics. Big data analytics can be used to extract meaningful knowledge and patterns from raw data by exploiting machine learning methods. The hidden knowledge and patterns revealed from mobile raw data can help to improve the performance mobile cellular networks and to maximize the revenue of operators. Compared with conventional big data problems, such as users profiling[2], sentiment analysis and opinion mining[3], etc., MBD analytics has the following distinctive features.

First, the volume of MBD is enormous. According to Cisco’s 2014 Visual Networking Index report, mobile data will exceed data from wired devices by 2018, constituting 61% of data traffic. By 2020, the amount of data traversing the Internet is expected to reach 1 billion gigabytes per month. Due to the contentious growth, the time duration for which collected data are processed for decision making can be relatively short. Therefore, MBD analytics should be rapidly executed to cope with the newly collected data samples. Second, MBD has temporal and spatial characteristics. The sources of MBD are mobile smartphones, sensors and IoT ends. Mobile devices and some types of nomadic IoT devices are free to move independently among many locations, which gives rise to the tempo-spatial features of MBD. Measurements on a CDMA2000 network revealed that wireless data traffic is bursty and exhibits strong diurnal patterns[4]. Third, in addition to the large data volume and the tempo-spatial characteristics of the data, MBD also differs from conventional big data problems because the data acquisition units of MBD are spread around complete logical network entities. The sources of traditional analytics, such as charging and billing systems and operation systems, are basically
centralized, whilst the sources of MBD are scattered across the infrastructure, such as cell sites, core network equipment, operation maintenance centres (OMC) of various vendors, and customer complaint departments. The higher dimensionality of the data implies better inference, which can provide enlightening insights. However, the high dimensionality also gives rise to the problem of data heterogeneity. The data are generated from different sources with disparate data formats, and the diversified granularity makes data fusion more challenging.

In this paper, we propose a novel mobile big data architecture consisting of five layers: the data storage layer, the data fusion layer, the data security layer, the data analysis layer and the data application layer. The functionality of each layer is presented. Under this architecture, we consider one illustrative cases, namely, user experience prediction, by leveraging machine learning techniques. In practice, the MBD analytics can be used for network planning and parameter dimensioning to facilitate cellular network design, deployment and operation.

2. Datasets Description
Our research in this work is based on real-world mobile datasets collected from Chongqing, one of the largest cities in China. This city has a population of approximately 3 million, and the operator has a user penetration of 2/3. This dataset is a record of 1.6 million anonymous mobile users' data traffic on a Saturday in 2014. Raw mobile data can generally be categorized into five types: application-related data, network-related data, link-related data, user-related data and operation-related data.

- Application data are the profiles related to user applications, such as application types (instant message, video, web services, etc.), the rate of flows, the volume of flows, and the number of TCP segment retransmissions associated with the flows. These data are usually collected by deep packet inspection executed at application servers.
- Network data contain information such as the coordinations of the base stations, the system bandwidth allocated, the key performance-related indicators (accessibility, retainability, quality, etc.) and various signalling interactions between users and networks. These data are usually collected from the base stations and core network equipment.
- Link data include channel quality information between the users and the base station, which is obtained by channel measurement performed by the users or base stations.
- User data include the behaviours and the preferences of users, e.g., locations, mobility, routines and experiences. Compared with the other three types of data, user data are not directly obtained from wireless networks but via data analytics based on the three previous types of data. For instance, the locations of users can be estimated via positioning algorithms based on the channel information contained in link data. Additionally, user locations might be filtered from the hypertext transfer protocol (HTTP) request records in application data when subscribers are using real location-based social apps or services (Yelp, Google maps, Foursquare, etc.).
- Operation-related data include customer care/ticket information, provisioning data, handset agent records, and turn-up/test records.

3. Big Data Architecture for Mobile Networks
To store, process and make full use of the MBD generated from real communication networks, we propose a big data architecture that consists of five layers: the data storage layer, the data fusion layer, the data security layer, the data analysis layer and the data application layer. Generally, the data storage and fusion layer is responsible for storing all types of data gathered from various core network nodes (i.e., base stations, serving gateways, mobility management entities). The data fusion layer filters, cleans, associates, and abstracts the types of data.

3.1. Data Storage Layer
The data storage layer handles a wide range of data types and sources, as described above. These mobile raw data have a size of approximately 6 TB and are stored in approximately $3 \times 10^6$ files for one-day logs. The value of the raw data is uncertain, and it is better to store them all without loss. Therefore, an infrastructure that can store the raw data at a low cost is desired. By adopting both
MongeDB and Hadoop distributed file system (HDFS), the database of our architecture can maintain low cost while achieving rapid response.

![Proposed big data architecture.](image)

3.2. Data Fusion Layer

The data fusion layer is designed to handle data extraction/transformation/loading (ETL), enrichment, pre-aggregation, and related functions. The layer is designed to support high availability and to robustly handle missing or corrupted data.

While operators have reams of data at their disposal, the data are usually trapped in disconnected tools spanning domains such as RAN performance, core signalling, application performance, and provisioning. Hence, when problems arise, engineers must “swivel chair” between systems to identify the root causes, which is a painstakingly slow and expensive process. The data fusion layer can overcome this problem by automating the process of fusing, analysing and extracting insights from data across these domains. Specifically, the data fusion layer maps enriched source data to a wireless data model spanning custom, network and reference information. This wireless data model consists of:

- Fact tables containing enriched time series records loaded by the ETL engine. These fact tables are optimized for specific types of records, such as 2/3/4G control plane or user plane;
- Dimensional tables covering invariant information, such as device types (make, model, other information), subscriber group affiliations, network element information (identifiers, geolocation information, topological relationships, vendor information, technical information, etc.), and user plane reference info (APN and URL groupings, etc.).

3.3. Data Analysis Layer

This layer ingests enriched data from the data security layer, applies a variety of analytical techniques, and exposes the results of these analytics to the data application layer for further data mining and knowledge discovery. The analytical techniques are based on three statistical analysis methodologies, which are described in the following.

3.3.1. Distribution-Based Analytics

The probability density function (PDF) is a statistical expression that defines the probability distribution for a random variable (R.V.). Another distribution-based analytic is the cumulative distribution function (CDF). These two divergence metrics can facilitate network anomaly detection by comparing the current probability distribution of a feature to a set of reference distributions that describe its "normal" behaviour.

For instance, we draw the distribution of traffic density generated at 6 AM and 12 AM, as shown in figure 1. From the figure, we know that the empirical data follow log-normal distribution, and at 6 AM, the distribution has an expectation of 3.23 and a standard deviation of 1.87, whereas these two parameters are 5.45 and 1.75 at 12 AM. These parameters can help us to detect network anomalies if
the real-time network performance in terms of traffic flow significantly deviates from the empirical value. Additionally, the divergence metrics can be applied to detect abrupt changes in the empirical PDFs of relevant features.

3.3.2. Entropy-Based Analytics
Entropy can summarize the changes in the distribution of a certain variable. Specifically, given an empirical distribution of a variable, its entropy represents a measure of the dispersion or concentration of the feature around a single value.

Consider another example: we discretize time into hourly segments and aggregate the traffic volume by hour. Then, the hourly traffic entropy of each base station is defined as $s = -\sum h(i) \times \log h(i)$, where $h$ is the traffic proportion of hour $i$. A lower value means that this cell has greater traffic concentration on a few hours, whereas a higher value means the cell traffic is mostly stable over time. We plot the density histogram in figure 2. The results follow a truncated Laplace distribution, which has a negative skew and more than 4 peaks.

3.4. Data Application Layer
The applications of big data analytics in mobile cellular networks can be divided into two categories: internal business supporting applications and external innovative business model developments.

The internal business supporting applications include the operational efficiency, subscribers’ experience enhancement, and tailored marketing. The external innovative business model developments cover location-based social applications to personalized recommendation system designs, from traffic dispersion to precision marketing.

Figure 2. Distribution of traffic density generated at 6 AM and 12 AM.

Figure 3. Distribution of hourly traffic entropy.

4. Application Case One: User Experience Prediction
In this section, we present user experience prediction as a case study of applying the proposed big data architecture in wireless network management. MBD can proactively identify customer experience issues tied to data and voice services by utilizing quality of experience (QoE) scores and linking to the underlying data model. Based on the QoE scores, operators can optimize network capacity planning, resource prioritization, self-care, device management, and other activities to minimize customer complaints and to predict them before they occur.

In QoE measurements, the parameters used to evaluate QoE (including the weights used in the KPI-KQI-QoE mapping process and the threshold used to distinguish unsatisfied and satisfied users) are mostly empirical values, which are subjective. Therefore, we resort to machine learning methods to overcome the subjective tendencies introduced by traditional QoE metrics. From the perspective of machine learning, the user experience prediction process is essentially a classification problem. That is,
we need to classify the mobile users in our dataset into two categories: satisfied users and dissatisfied users. This application process consists of three major steps.

4.1. Feature Abstraction
The first step in supervised learning is to determine the input feature representation of the learned function. A feature is a measurement attribute extracted from sensory data to capture the underlying phenomenon being observed and to enable more effective MBD analytics. The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector that contains a number of features that are descriptive of the object. The number of features should not be too large, due to the curse of dimensionality, but should contain sufficient information to accurately predict the output.

First, we introduce six mobile network performance features by artificial expertise, which are considered to potentially be related to user complaints, as listed in table 1. Feature 1 is mean opinion score (MOS). MOS is a popular indicator of perceived media quality. Feature 3 is the proportion of the number of disconnected calls to the total number of calls. Feature 4 is the proportion of the number of unexpected dropped calls to the total number of calls. Feature 5 is the number of calls originated by the end user, and feature 6 is the number of calls received by the end user.

| Complaint Features                  | $\eta^2$ value |
|-------------------------------------|----------------|
| Low MOS ratio                       | 0.0411         |
| User instability                    | 0.01407        |
| Disconnected calls                  | 0.00204        |
| Unexpected dropped calls            | 0.00151        |
| Originating call attempts           | 0.01793        |
| Receiving call attempts             | 0.0135         |

These abstracted features are verified by the correlation coefficient, which is an approach to measure the statistical relationship between two variables. We introduce the $\eta^2$ correlation coefficient to measure the overall relationship between a continuous dependent variable (DV) and a categorical independent variable (IV). $\eta^2$ ranges from -1 to 1, where 1 indicates the strongest possible agreement and 0 the strongest possible disagreement.

4.2. Data Labelling
The second step in supervised learning is to train the learning model with labelled data. Note that there is a difference between complaint users and unsatisfied users. Complaint users are group of extremely unsatisfied users. Therefore, the dataset we obtained is partially labelled, and we should label it manually. We adopt the cosine similarity method to fill the gap.

The similarity between users in the dataset and complaint users is calculated by the cosine similarity. The smaller the value of the cosine similarity is, the stronger the relationship between the unlabelled user and the complaint user.

4.3. Classifier Construction
In the final step, we feed labelled data to the selected machine learning algorithms. We partition the whole dataset into two parts: training dataset and testing dataset. The algorithm is trained on the training dataset and gives the desired output on the testing dataset. We test a variety of widely used classification models, including naive Bayesian (NB), Bayesian, support vector machine (SVM), and random forest (RF). As illustrated in table 2, SVM achieves the best performance.
Table 2. The performance of different classifiers

| Classifier       | Precision | Accuracy | Recall | Alarm |
|------------------|-----------|----------|--------|-------|
| Naive Bayesian   | 0.226     | 0.894    | 0.991  | 0.009 |
| Bayesian         | 0.0018    | 0.865    | 0.315  | 0.385 |
| Support Vector   | 0.976     | 0.929    | 0.996  | 0.004 |
| Random Forest    | 0.999     | 0.913    | 0.993  | 0.007 |

5. Conclusion
In this article, we propose a novel mobile big data architecture that consists of five layers. The data storage layer stores a large amount of data collected from different data sources. Then, the data fusion layer handles data ETL, and the data security layer guarantees data integrity, availability, and confidentiality. Consequently, the processed data are input into the data analysis layer, where a variety of analytical techniques are applied. Finally, the data application layer initiates machine learning methods to extract hidden knowledge and patterns. Under this architecture and with the leverage of machine learning techniques, mobile big data analytics can be used for network planning and parameter dimensioning to facilitate network design, deployment and operation.

6. Acknowledgments
This work was partly supported by National Natural Science Foundation of China under the Grant No.61701034 and No.61531007 and partly supported by China Postdoctoral Science Foundation under the Grant No 2017M620696.

7. References
[1] C. V. N. Index, “Global mobile data traffic forecast update, 2013-2018,” White Paper, February, 2014.
[2] S. N. Schiaffino and A. Amandi, “Intelligent user profiling,” Artificial Intelligence, pp. 193–216, 2009.
[3] B. Liu, “Sentiment analysis and opinion mining,” Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–10, 2012.
[4] C. Williamson, E. Halepovic, H. Sun, and Y. Wu, “Characterization of cdma2000 cellular data network traffic,” in Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on. IEEE, 2005, pp. Z000–719.