Detection of Missing Values from Big Data of Self Adaptive Energy Systems

MVD tool detect missing values in timeseries energy data

Muhammad Nabeel
Computer Science Department, SST
University of Management and Technology (UMT)
Lahore, Pakistan
muhammad.nabeel@umt.edu.pk

Malik Tahir Hassan
Computer Science Department, SST
University of Management and Technology (UMT)
Lahore, Pakistan
tahir.hassan@umt.edu.pk

***Abstract*— Information technology is playing an important role in the development of self adaptive energy systems by providing reliable, healthy and secure services to the smart grid. Automated metering is deployed which captures energy consumptions in detail. Meters collect the data at different rates mostly vary from few milliseconds to seconds. Data which are coming from smart meters is high magnitude of time series data. Time series is a series of data which is measured over time containing timestamp in it. Data is integral part of modern grid and quality of data should be ensured. Energy Systems have different important requirements like reliability, flexibility, load balancing, efficiency, sustainability etc. Goal of missing value detection (MVD) tool is to find out missing values in energy data to improve data quality of smart grid. Quality of data is important to have accurate analysis, prediction and self-healing of network.**

Keywords—SCADA; Big Data; Missing Values, Energy Management System

I. INTRODUCTION

Information technology is playing an important role in modern grid by providing robust, secure and efficient energy management. Automated metering infrastructure which is deployed at homes is providing near real-time energy consumption to the service provider. Different applications have been developed like load monitoring, energy forecasting, non-intrusive load monitoring, fault detection and demand side management, etc. All these application can provide accurate results only when we have accurate data with no anomalies. It is important to find and fill missing values to run accurate grid applications. Collection of data from meters and transmission of data to the service provider is critical to make the grid reliable. Between transmission and collection some data can be lost due to various reasons such that faculty equipment, lost records, human mistake etc. To ensure quality of data it is important to find missing values from raw data, estimate the missing values and fill the missing values with suitable values.

Smart grid contains large amount of data. Data mining is the process of extracting useful from large datasets. Data mining enables smart users to explore the data from different dimensions, summarize and categorize the data. Data mining helps in explore large volumes of unstructured data to identify relationships and regularities which lead to better understanding of the data.

One of important step, of data mining is data preprocessing. There are different data preprocessing stages like data selection, attribute selection, data cleaning, data integration, summarization and transformation to constructs final dataset from a data set. This includes an important step of identifies and predicts missing values. Once data transform into final dataset then identification of missing values becomes complex. Data cleaning is involved in various fields like weather, stock market, scientific data analysis etc. This involves missing value identification, prediction and noise removal.

Temporal data mining is extraction of knowledge from database with respect to time. Ideally we are assuming that data which is transmitted and collected are clean with no missing value and outliers, but in reality this does not happen. It is important to check the quality of data at initial stage. Common problem for time series data is frequently missing values in the data.

In Self managing systems values are coming in successive order thus making it in time series data which is collected at regular intervals over a period of time. Time series is a sequence of numerical data values in successive order. Common notation for defining time series is.

$$Y = \{Y_t : t \in T\}$$

Y is the reading record at particular time t and T is the index set.

A. Data Management in Energy Management System

Smart grids are very large interconnected information systems where computing agents, utility providers and consumers deliver and use various services. We present see in data flow section regarding which services need which type of time series data. Range of incoming stream of data vary from

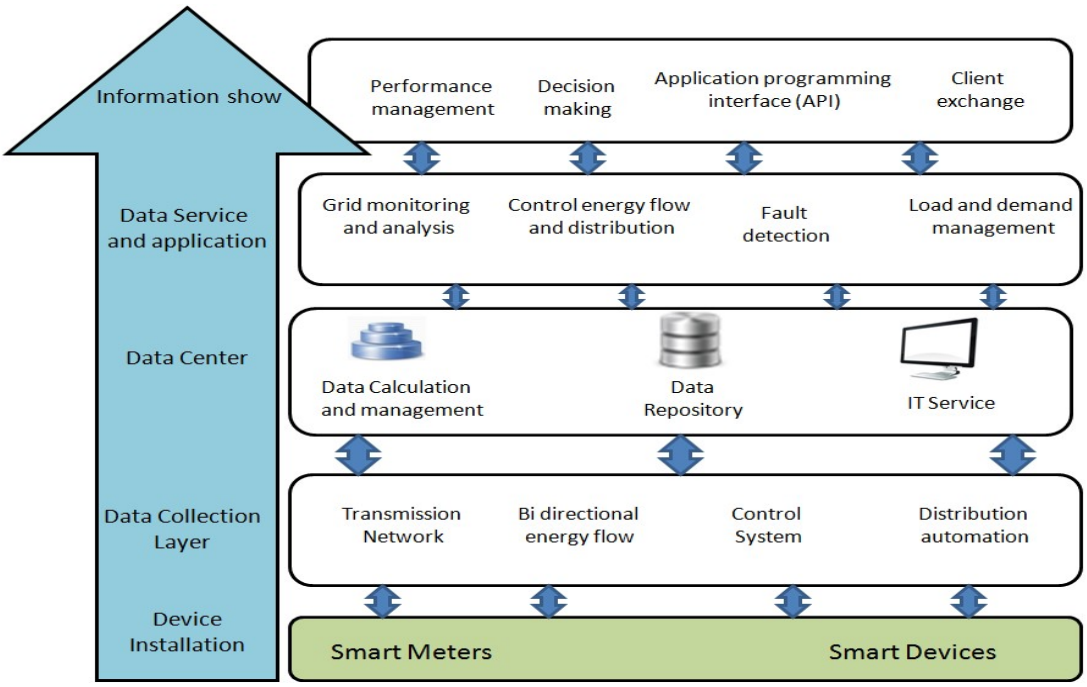

Figure 1 Data Flow in Self Managing Energy System

the fundamental property of data that time series constant. Time range of data varies from 16 KHz for non-intrusive, 1 KHz for visualization and few seconds to minutes for customer level aggregate report. Managing this magnitude of data is a challenging problem. First challenge is to store high speed of smart grid time series data, so that maximum data can be saved for long period of time with minimum storage space which has been discussed in [4]. Second challenge is to ensure data quality for accuracy of result which we discuss in next sections. Our first goal regarding data quality is to find missing values of data, so that we can predict the missing values and suggest the missing values for accuracy improvement. Smart Grid features of accuracy heavily rely on data accuracy. In this paper, we aim at solving the missing values problem in smart grid data.

There are two important aspect regarding data, first is storing in hardware. Store data in such a way such that data processing becomes fast. Second important aspect is to ensure quality of data. First step for quality of data is to ensure how many missing values are present in data. In Further section will present the importance of finding missing data and experimental results.

B. Data Flow in Energy Management System

Smart grid monitors real time transmission, distribution and power generation. Figure 1 shows the overview of data flow in smart grid. Sensors sending the data to smart grid with milliseconds to second delay which is time series in nature. Accuracy of data is crucial because energy management system control the grid, solves problems of detection and predictions and have a capability of fault detection and self healing of the system. Integrated transmits data and other information between the data collection layer and the application layer through enterprise service bus [4].

Smart grid applications mainly are deployed at the application layer. Some of the smart grid application are fault tolerance, load disaggregation [5], forecasting [6] and advanced visualization [7].

Need for time series data for each application vary greatly. For example load disaggregation algorithm requires 16 KHz data. Visualization application time series data is sufficient to visualize on 1Hz. Rate of propagation of data may vary. Data flow through different channels is shown in figure 1. On each layer semantics of time series data change according to the need but the important task is to make sure, that when semantics change the system does not lose information.

II. MOTIVATION AND PROBLEM STATEMENT

Time series is a collection of readings or values recorded sequentially through time over regular interval. Time series occur in variety of fields, ranging from engineering to economics and methods of analyzing time series constitute an important area of statistics [9]. A time series in which observations are made continuously through time is said to be continuous time series [10].

Smart grid data depends upon time series quality. Data from smart grid comes from different meters and devices and granularity of data varies from few milliseconds to minutes. Variation in time property is because of different time granularity of data needed by different applications of EMS.

As discussed above two major issues occur when such high velocity and low granularity data is coming. First issue is of storage, as sometimes volume of data is large that it easily overflows from the storage capacity. Second issue is of accuracy of data which is coming from different meters and smart devices.

Accurate results in smart grid rely on accurate data. To keep accuracy of data, data pre-processing step is required which is shown in figure 2. Data pre-processing have different stages. Step no 2 in data processing is the important step which clean the data and identifies missing values. First step towards accuracy of data is the data cleaning and in data cleaning missing value identification is the major step towards accuracy of data. Data size of EMS is large and one cannot easily find missing values in the data, so we need to develop effective software which can easily detect and suggest the missing values. Goal of this paper is to identify a missing value, so we develop MVD tool which intelligently identifies missing value in EMS data.

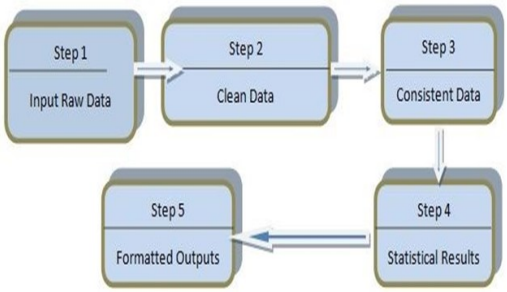

Figure 2: Data processing stages

We have picked different smart grid datasets from different sources. Table 1 shows the record in these datasets. We remove the actual name of data and call the dataset as dataset 1 and dataset 2 to protect the privacy of the data.

Table 1: Size of Different Data Sets

Data Set	Records
Data Set 1	17 Thousand Approximately
Data Set 2	0.2 Million Approximately

All above datasets are power datasets having different observations and being heterogeneous in nature. Both datasets have different ways of storing the power data according to their needs. We have developed MVD tool which intelligently finds missing values from any power datasets.

III. Experimental Results

Task of data preprocessing is to improve data quality. Data inconsistencies in data gathering stage can be due to manual entry, no uniform standard for contents and formats, parallel data entry, measurement errors, etc. Inconsistent data leads to inconsistent results. More accurate data means more accurate analysis and predictions.

Figure 3: GUI of Missing Value Detection Tool

We have developed smart missing values detection tool, which intelligently detects missing values from power data. GUI of tool is presented in figure 3. Most of power systems have their own structures of storing the data. We have developed a tool which identifies missing values from any power datasets. Our system will intelligently detect the time interval, identify missing values. Figure 4 shows sample of datasets 1 and dataset 2. We identify missing values from these datasets by using our MVD tool. These datasets are real time datasets taken from real systems and to protect the privacy of data we blur some of attributes and some other attributes are not shown in the figure.

Grid Station	Feeder Name	Date	Time	kWh
		25-Aug-15	0:00:00	16403928
		25-Aug-15	0:15:00	16404192
		25-Aug-15	0:30:00	16404432
		25-Aug-15	0:45:00	16404696
		25-Aug-15	1:00:00	16404936
		25-Aug-15	1:15:00	16405200
		25-Aug-15	1:30:00	16405440
		25-Aug-15	1:45:00	16405680
		25-Aug-15	2:00:00	16405944
		25-Aug-15	2:15:00	16406184
		25-Aug-15	2:30:00	16406424
		25-Aug-15	2:45:00	16406688
		25-Aug-15	3:00:00	16406928
		25-Aug-15	3:15:00	16407192
		25-Aug-15	3:30:00	16407432
		25-Aug-15	3:45:00	16407720
		25-Aug-15	4:00:00	16407984
		25-Aug-15	4:15:00	16408224
		25-Aug-15	4:30:00	16408272
		25-Aug-15	4:45:00	16408272
		25-Aug-15	5:00:00	16408296
		25-Aug-15	5:15:00	16408320

Figure 4 (a) Sample of Dataset 1

Grid Station	Feeder Name	Feeder Code	Feeder Category	Date	Time	kWh			
		115	Exempted Feeders	24-Aug-15	14:30:00	2718808		2.38	56024896
		115	Exempted Feeders	24-Aug-15	14:45:00	2718912	416	2.53	56024296
		115	Exempted Feeders	24-Aug-15	15:00:00	2719024	112		
		115	Exempted Feeders	24-Aug-15	15:15:00	2719120	96		
		115	Exempted Feeders	24-Aug-15	15:30:00	2719224	104		
		115	Exempted Feeders	24-Aug-15	15:45:00	2719320	96		
		115	Exempted Feeders	24-Aug-15	16:00:00	2719424	104		
		115	Exempted Feeders	24-Aug-15	16:15:00	2719536	112		
		115	Exempted Feeders	24-Aug-15	16:30:00	2719640	104		
		115	Exempted Feeders	24-Aug-15	16:45:00	2719752	112		
		115	Exempted Feeders	24-Aug-15	17:00:00	2719856	104		
		115	Exempted Feeders	24-Aug-15	17:15:00	2719960	104		
		115	Exempted Feeders	24-Aug-15	17:30:00	2720064	104		
		115	Exempted Feeders	24-Aug-15	17:45:00	2720176	112		

Figure 4 (b) Sample of Dataset 2

We have provided the dataset 1 and dataset 2 to our tool. When we give dataset 1 to our system it provides us results as in figure 5.

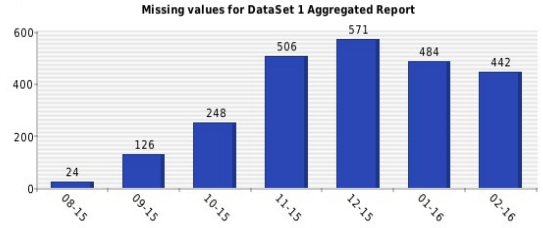

Figure 5 Result of dataset 1 month wise aggregated

Figure 5 represents the missing values in dataset 1 aggregated on monthly basis. X-axis shows month and year Y-axis shows missing values in each month. From the graph we can see that total missing value from month of August 2015 till February 2016 are 2401. System finds out that maximum missing values is: in month of December and minimum missing value is in month of the august. When we provide dataset 2 in our system it provide us output as shown in figure 5.

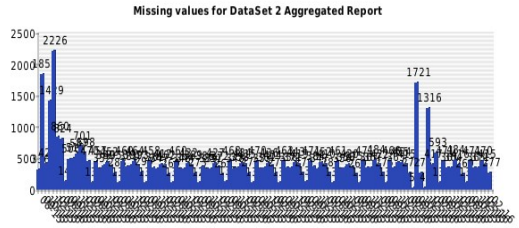

Figure 6 Result of dataset 2 month wise aggregated

We find very interesting fact when we provide dataset 2 that some data is of poor quality and even does not contain valid timestamps in it. Figure 6 shows result for dataset 2, X-axis represents the month and Y-axis represents the missing values in a graph. Graph dataset 2 shows repetition of time again and again. One of the major property of time series data is time always come in incremental manner mean once particular time and its reading recorded then that time cannot come again. Figure 6 shows dataset 2 is invalid data which did not fulfill time series property.

IV. Conculusion and Future Work

Information technology is playing its vital role in Smart Grid. To make sure the accuracy of different applications of Energy Management System quality of data is important. Many applications of power system depend on quality of data like load balancing, self healing, demand response, energy forecasting etc. To make sure all the applications of smart grids work, effectively we need high quality data. Data preprocessing is an important step to ensure data quality. We have presented in this paper a tool for the importance of finding missing values and developed a tool which finds missing values, aggregates the result on monthly basis. Our next goal is to publish our software so that power community can validate and find missing values from power datasets. Verify and validate our tool by applying more different power datasets.

References

[1] S. Sridevi, S. Rajaram, C. Parthiban, S. SibiArasan, and C. Swadhikar, “Imputation for the analysis of missing values and prediction of time series data,” in Recent Trends in Information Technology (ICRTIT), 2011 International Conference on. IEEE, 2011, pp. 1158–1163.

[2] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, “Missing value estimation methods for dna microarrays,” Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.

[3] Z. Zhang, W. Wu, and Y. Huang, “Mining dynamic interdimension association rules for local-scale weather prediction,” in Computer Software and Applications Conference, 2004. COMPSAC 2004. Proceedings of the 28th Annual International, vol. 2. IEEE, 2004, pp. 146–149.

[4] Nabeel, Muhammad, Fahad Javed, and Naveed Arshad. "Towards Smart Data Compression for Future Energy Management System." Fifth International Conference on Applied Energy, Pretoria, South Africa. 2013.

[5] Ali, U., Rana, Z. A., Javed, F., & Awais, M. M. EnerPlan: smart energy management planning for home users. In Neural Information Processing. Springer Berlin Heidelberg 2012; 543-550.

[6] J. Zico Kolter, Matthew J. Johnson. REDD: A Public Data Set for Energy Disaggregation Research. In proceedings of the SustKDD workshop on Data Mining Applications in Sustainability 2011; 1:6.

[7] Javed, F., Arshad, N., Wallin, F., Vassileva, I., & Dahlquist, E. Forecasting for demand response in smart grids: An analysis on use of anthropologic and structural data and short term multiple loads forecasting. Applied Energy 2012.

[8] Ali, U., Rana, Z. A., Javed, F., & Awais, M. M. EnerPlan: smart energy management planning for home users. In Neural Information Processing. Springer Berlin Heidelberg 2012; 543-550.

[9] Chatfield, Chris. The analysis of time series: an introduction. CRC press, 2016.

[10] Brillinger, David R. Time series: data analysis and theory. Vol. 36. Siam, 2001.

[11] Fox, Anthony J. "Outliers in time series." Journal of the Royal Statistical Society. Series B (Methodological) (1972): 350-363.