Volume 5, Issue 3, March 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Data Mining in Smart Grids-A Review

Manju Khanna
Asst. Professor
Department of CSE
MVJ College of Engineering
Karnataka, India

Dr. N. K. Srinath
Professor
Department of CSE
RV College of Engineering
Karnataka, India

Dr. J. K. Mendiratta
Professor
Centre for Emerging Technologies
Jain University
Karnataka, India

Abstract: The Heterogeneous databases are widely used in various applications, including the smart grid at the generation end of the electrical energy.Since the data being generated by various renewable energy sources is based on time, time series data mining is used.Since the data being generated by the various resources is very vast, the analysis of it has to be done using specific techniques such as Discrete Fourier Transform, Wavelet transform etc.The present work reviews various techniques applied and compared with their limitations and advantages .Application of these methods can be used in hybrid smart grid for its optimal performance.

Keywords: Smart grid, Datamining, renewable, Heterogeneous, Distributed Data Mining, Discrete Fourier Transform

I. INTRODUCTION

With the digitization of processes,management,space, ocean exploration and other fields, data production is increasing in all these fields.In order to draw conclusions on such huge amount of data has become extremely difficult.So, Data mining [1] plays a very important role in decision making and analysis.

The Electric grid [2] has been a centralized system for electricity generation and distribution.Over a period of time with emergence of clean energy sources, such as solar, wind, biogas [3] for the generation of electricity operating in isolation or connected to the central grid power systems, it is transformed to a decentralized architecture .Further factor in decentralization of the grid has been total black outs in central grids, leading to loss of economy of the total connected systems.These source of energy can be used separately or all can be combined to form the Hybrid [3] renewable sources of energy.

Integrated use of the renewable[3] source of energy as well as the effective use of digital technologies led to the Smart Grid[2].With different uncontrolled sources of energy, compelled to monitor these energy sources and operate them with optimal performance. To achieve this objective, a new control strategy has evolved for operation of these sources of energy with different generation components changing with time. For the optimal operation, different types of data is required to be monitored and various sources of energy utilized to their utmost factors of optimization. The processing of data is involved around two aspects, one which involves the database management systems and other data mining [4].

Database Management system is required for data storage, transaction processing and report generation while data mining is required for analysis and future optimal performance.

II. PROBLEM DESCRIPTION AND PRIOR WORK

The major problem faced to monitor such a heterogeneous data involves selection of a common indexing technique, which may be applicable for such an Hybrid type of data.

Data Mining deals with analyzing the data [5] and this data if is distributed across various sites is named as Distributed Data Mining(DDM).Literature is flooded with various techniques evolved in data mining for Homogeneous data, but not much work has been carried out on heterogeneous data.In case of distributed grid systems with hybrid sources of energy, with each source monitored

Data Mining also plays a key role in Smart Grid [2] which gathers information both on the supplier side as well as the consumer side.

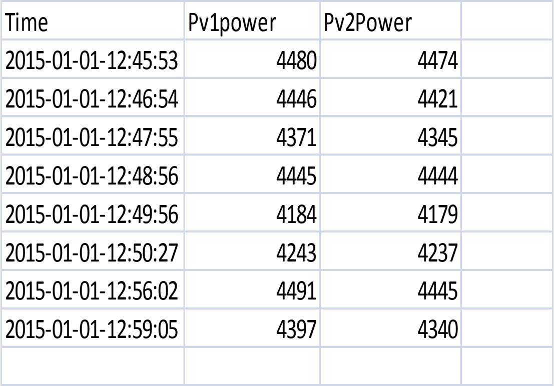
Since lot of data is involved at both ends decision making and analysis is required so as to improve the efficiency, reliability of electricity production and consumption.

Because the supply of energy is from various renewable resources which are distributed in nature we need to know how the datais distributed.

Since the data distributed is a structured database which has well defined columns, they can be horizontally or vertically portioned.

If they are homogenous database (horizontally partitioned) or heterogeneous database (vertically partitioned) then we are required to see how they need to be analyzed.In case of the heterogeneous database since the partitioning [5] is vertical each site of the energy resource will have different attributes, Since they have different attributes we need a unique identifier to facilitate matching.

One of the unique identifier is Timestamp, which can be just time, combined of data.In our case of renewable energy source at a specific time some power is generated whether it is solar, wind or any other source of energy.
The Figure 1 shows how data of a solar panel is heterogeneous and vertically partitioned based on Time.
The columns shown are Time, power generated by solar panels



Time	Pv1power	Pv2Power	
2015-01-01-12:45:53	4480	4474	
2015-01-01-12:46:54	4446	4421	
2015-01-01-12:47:55	4371	4345	
2015-01-01-12:48:56	4445	4444	
2015-01-01-12:49:56	4184	4179	
2015-01-01-12:50:27	4243	4237	
2015-01-01-12:56:02	4491	4445	
2015-01-01-12:59:05	4397	4340	

Figure 1 Heterogeneous solar panel data

Since the matching is done based on time stamp we need to use different paradigms of data mining other than the usual applied for homogeneous data .The other type of Data mining techniques which Karugupta[5]hasproposed is Collective Data Mining(CDM) for predictive data modeling .This technique is a framework for inducing patterns in the machine learning, statistics , distributeddatabases. Other extensions of CDM include distributed decision tree construction(Park,Ayyagari&Karugupta 2001)[5] and collective hierarchical clustering(Johnson and Karugupta,1999)[5].We use the time series data mining [4][6] which is of much relevance in the smart grids.
With the present data available, we can estimate its future behavior which helps in our decision making. With time stamping of data from various energy sources, optimization of the total power generation can be achieved, thereby helping in supplying power to different loads with their load forecasting.
The forecasting could be done for short term (as weeks or months) or for long term (as years).So we need to find a data mining technique which could enable us to achieve maximum utilization of energy.
Figure 2 shows Data mining technologies and their applications in power grids [6].

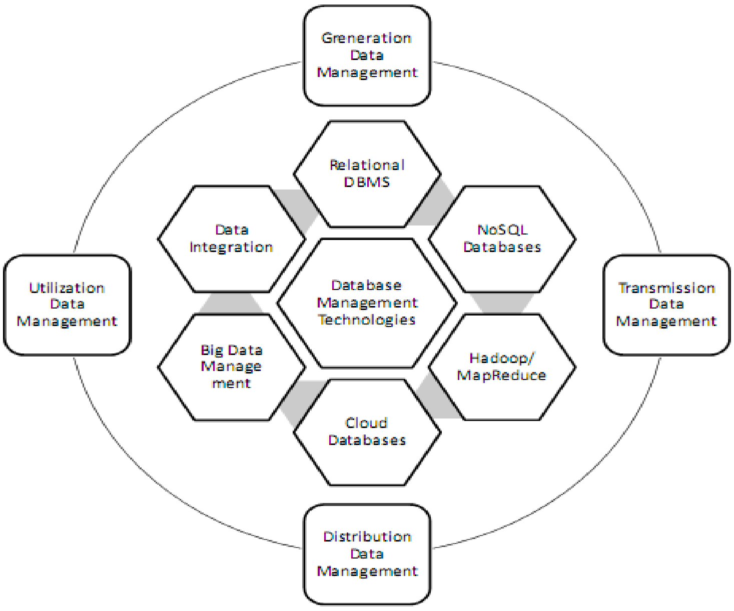

Greneration Data Management
Relational DBMS
Data Integration
NoSQL Databases
Utilization Data Management
Database Management Technologies
Transmission Data Management
Big Data Manage ment
Hadoop/ MapReduce
Cloud Databases
Distribution Data Management

Figure 2 Data mining technologies and their applications in power grids

III. BACKGROUND ON TIME SERIES DATA MINING

In many fields including the smart grid the measurements are carried out over a period of time. This results in data in time series [6] format. In the time series data mining the data is extracted from the shape of the data.
The Time series data [8] generally have two goals (1) modelling time series data (2) forecasting time series data. There are four components to characterize time series data. [8]

(1) Trend or long term movements – to determine a trend curve we use least squares and moving average methods.
(2) Cyclic movements which may be periodic or aperiodic.
(3) Seasonal movements that is movements which vary with season that for example in the solar panel, during the summer the direct radiation is more than the diffused radiation[11]
(4) Irregular or random movements.

The Figure 2 shows the time series data plotting of a solar panel where it is plotted against the power which is generated.

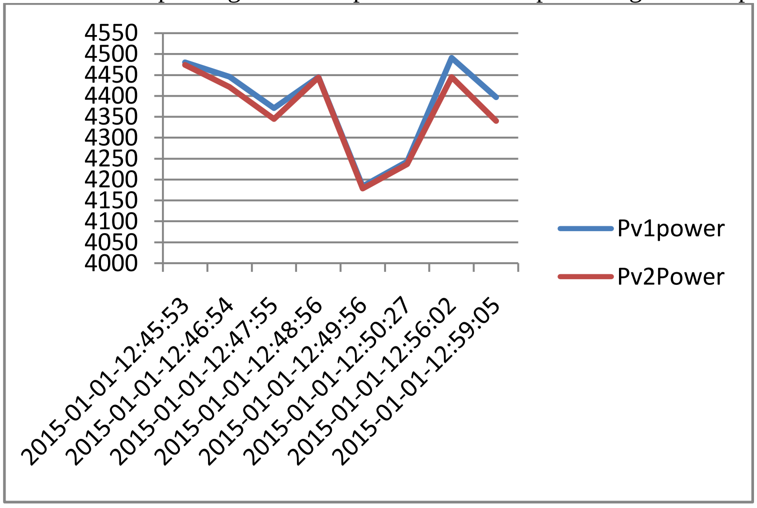


Figure 2 Time series data of a solar panel

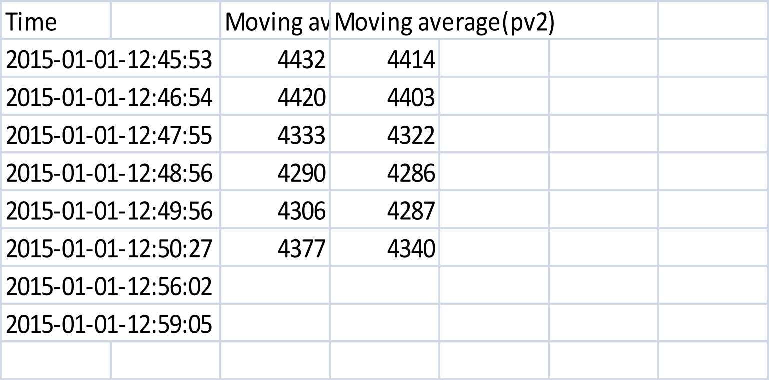
These movements are represented by variables (T,C,S,I) .They can be modelled as a product as (Y=T * C * S * I) or sum(Y=T+C+S+I) with notation details given at the end.

The trends can be determined by using the moving average method of order n as following sequence of means:

x1+x2+…..xn/n,x2+x3+…..xn/n this helps eliminates unwanted fluctuations.

For example let us see the data generated by the solar panel.

In the example we have shown how the plotcome because of the moving average for the 8 samples of data and a order of 3.



Time		Moving av	Moving average(pv2)			
2015-01-01-12:45:53		4432	4414			
2015-01-01-12:46:54		4420	4403			
2015-01-01-12:47:55		4333	4322			
2015-01-01-12:48:56		4290	4286			
2015-01-01-12:49:56		4306	4287			
2015-01-01-12:50:27		4377	4340			
2015-01-01-12:56:02						
2015-01-01-12:59:05						

Figure 3shows moving average of the sample data for solar panel

The moving average method smoothen the data, eliminates any cyclic, seasonal and irregular movements.

The seasonal variations are of much significance because in summer the direct radiation is more which will be due to the months of March, April.

Once the trend analysis has been done then we need to do the Data Reduction and Transformation.

For the time series forecasting the ARIMA (Auto –Regressive Integrated Moving Average) [8] model can be used. This model can be used to predict the future points in series.

IV. TIME SERIES DATA REDUCTION

Since the data we are analyzing is of tremendous size, data reduction and transformation needs to be done.

There are various techniques which are used in the data analysis, i.e. Discrete Fourier transform(DFT) and the Discrete Wavelet transform (DWT)[10]

Each time series is compressed with wavelet or Fourier decomposition.Instead of using only the first coefficient, a new method of choosing [10] the best coefficient are chosen from the time series data.

Discrete Fourier Transform: It is the projection of a signal from time domain to frequency domain, and is given by:

$$cf = \left(\frac{1}{\sqrt{n}}\right)\sum_{t=1}^{n} \mathrm{f(t)}\exp\left(\frac{-2\pi ift}{n}\right) \qquad (1)$$

Where f=1…n and i=$\sqrt{-1}$,

cf being complex number represent the amplitude and shift of a decomposition of signal into sinusoid functions.

The Fourier transform measures the global frequencies and signal assumed to be periodic in nature, thereby causing poor approximation at the border of the time series. Instead, a fast algorithm, fast Fourier Transform is utilized.

To eliminate above limitations, a Discrete Wavelet Transform (DWT) is much suited for the application. DWT measures frequency at different time resolutions and locations. The basis function used is:

$\Psi_{j^{\cdot}k}(t) = 2^{j/2}\Psi(2^{j}\,t\text{-}k)$ (2)

Where Ψis the mother wavelet function. Any square integrable real function f(t) can be represented in terms of this basis as:

$$f(t) = \sum_{j,k} c_{j,k}\Psi_{j,k}(t) \qquad (3)$$

and the $c_{jk} = <\Psi_{j,k}(t), f(t) >$ are the coefficients of DWT.

A simple and commonly wavelet transform used is Haar wavelet transform with the mother function:

$$\Psi_{Haar}(t) = \begin{cases} 1, if\ 0 \prec t \prec 0.5 \\ -1, if\ \ 0.5 \prec t \prec 1 \\ 0, otherwise \end{cases} \qquad (4)$$

Feature extraction of the heterogeneous data can be achieved using DFT and DWT by use of first k coefficients and discarding the rest. Such a rough scratch of time series data with first coefficients representing low frequencies of the signal. In DFT, each feature is influenced by all-time points, as the coefficients describe global frequencies.

Whereas, in DWT, coefficients are influenced by sub-series of different sizes, offering a multi-resolution decomposition. A better method can be by extracting the largest coefficient, along with its position for each time series, providing optimal energy per time.

V. CONCLUSION

The methods and procedures developed for homogeneous data fail to meet the requirements of heterogeneous data, which has led to new methods applied as seen above. With requirement of clean energy, stress is being laid by various governments through the world to harness the renewable energy sources. The biggest hurdle in utilization of these sources being uncontrollability of the various sources, which change with time, seasons and change in environment. For effective harnessing of these sources, a new approach is required to be applied for the purpose. Data mining of the data collected over the year is taken into consideration and a controller required to change its parameters according to the profile of these resources has to be implemented. This requires need of intelligence on the part of controller.

It is required to develop various controllers for different power sources, which should communicate with themselves, i.e. intelligent agents, which can adapt its parameters as required by various resources and feed the power to the grid for their optimal performance.

REFERENCES

[1] http://en.wikipedia.org/wiki/Data_mining

[2] http://en.wikipedia.org/wiki/Smart_grid

[3] Jose Maria Gonzalez de Durana , Oscar Barambones “Object Oriented simulation of Hybrid Renewable Energy Systems focused on Supervisor control” ,Emerging Technologies and Factory Automation ,2009.ETFA 2009.IEEE Conference

[4] Zeyar Aung,”Database systems for smart grid”,Springer chapter 7, 2013

[5] Byung-Hung Park and Hillol Kargupta, “Distributed Data Mining: Algorithms, Systems and Applications” ,Department of Computer Scienc and Electrical Engg. University of maryland,2003.

[6] http://web.engr.illinois.edu/~hanj/cs512/bk2chaps/chapter-8.pdf

[7] Fabian Morchen, ‘Time series feature extraction for data mining using DWT and DFT” Nov. 2003 , Data Bionics, Philipps. University Marburg, Germany

[8] Viejo et.a. “Simulation of energy system scenarios for regional planning decision-making using agent based modeling” , 11th International conference on Computers in Urbn Planning and Urban Management CUPUM, Hong Kong, 2009

[9] Enrique Kremers et.al “ A complex systems Modelling Approach for Decentralized Simulation of Electrical Microgrids” ,9th IASTED Europian Conference on Power and Energy Systems,EuroPES 2009, Spain 7-9, 2009

[10] Enrique Kremers, et.al. “Agent based Simulation of Wind farm Generation at Multiple Time Scales” Chapter 14 book by InchOpen

[11] M.K.Deshmukh,S.S.Deshmukh “Modelling of Hybrid Renewable energy systems”,Renewable and Sustainable energy reviews,Elsevier,2006