Predictive Model Based Electrical Consumption

Diana Paul[1], Dhivya V[2], R.Kanagaraj[3]
[1, 2, 3] Department of Computer Science and Engineering
[1, 2, 3] Sri. Ramakrishna Engineering CollegeCoimbatore

Abstract- *The data mining is the task of analyzing of large quantities of data to derive previously unknown, interesting patterns such as groups of data records, unusual records , and dependencies. It has the ability to turn raw data into useful information. One of the important invention of mankind is electricity. It is considered as a blessing. The availability of this power is very much needed for development and economical stability of the nations. The electricity board in various states and countries perform tasks such as generation, transportation and distribution of electricity to its customers effectively. In this study the various data mining techniques are applied to derive information from the electricity consumption databases.*

Keywords- Data mining , Electricity consumption ,k-means, time series.

I. INTRODUCTION

Every human in this earth is dependent on electricity. The importance of electricity can be understood during the few minutes of power outages we encounter. The life of people almost stops when power outages. In this modern world one cannot imagine a life without electricity. The electricity is the used in many areas such as Communication, Entertainment, Work, Transportation, Food and Home. Electricity is a constantly developing technology, there would have been less technology in hospitals if electricity was not present, thus the human life is indirectly dependant on electricity. At home, electricity acts as the primary source of energy for water heating or heating, cooking, air conditioning, lighting and various other purposes. The electricity is considered the lifeline of the world due to the reasons mentioned above. The study about the electricity consumption can be done by using techniques and methods of data mining. Data mining is referred as Knowledge Discovery in Data.

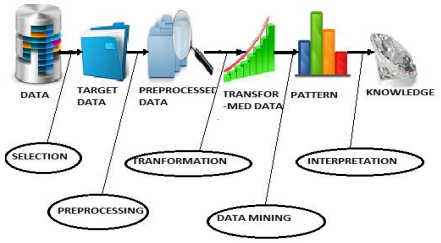

Figure 1. Steps in Knowledge Discovery of Data

Data can be mined from flat files, spreadsheets, database tables, or some other storage format. The most important criteria for the data mining is not the storage format, but its applicability to the problem to be solved. Data mining would not be an easy task, as it has complex algorithms and data will not always available at a single location. It has to be integrated from various heterogeneous data sources. A few widely used techniques include Characterization, Association and Correlation Analysis, Classification, Prediction, Cluster analysis, Outlier analysis, and Evolution analysis.

The data mining techniques are used to forecast the consumption of electricity. Forecasting is an art and science. According to Neils Bohr "Prediction is a very difficult art, especially when it involves the future". In the present world, an underestimate of electrical consumption may lead to localized brownouts or even blackouts. In case of overestimate could lead to authorization of plants which may not be needed for years. Short term forecasting helps in regulation. The forecasting of data various from simple extrapolation method to time series techniques. The forecasting of electricity using the data mining techniques is the objective of the proposed system.

II. LITERARY SURVEY

Ratore et al, proposes that by using data mining techniques the discovery of the electrical consumption pattern at regional level of the city can be determined[1]. The relationship between the electrical consumption and the atmospheric temperature along with geographical patterns such as river, farm, ground and highway are to be mined. The mining techniques included in the system are clustering and association rule mining techniques. The validation of the proposed system are done for a Sangli city. The clustering is the technique which is applied to form groups or clusters of data representing a common property. The association rule mining is used to establish relationship between the different entities.

The clustering is done for the electrical consumption data and atmospheric temperature. The k-means algorithm is being used for the clustering. One of the conclusion obtained from clustering is that electrical consumption increases as with the increase of temperature. Thus electrical consumption is

directly proportional to atmospheric temperature. The "Apriori" algorithm is used to derive the association rules by using the support and confidence. A set of five association rules have been derived. The data mining model depicted is generalized and can be derived for various other geographical region by analyzing the spatial data of the region and the electrical consumption data.

(a) IF (River AND Farm, Near) THEN Electricity consumption (Medium OR Low)
(b) IF (River OR Farm, Near) THEN Electricity consumption (Medium OR Low)
(c) IF (Ground OR Highway, Near) THEN Electricity consumption (High OR Medium)
(d) IF ((River OR Farm, Near) AND (Highway, Near) THEN Electricity consumption (Medium)
(e) IF (Ground AND Highway, Near) THEN Electricity consumption
(High)

Figure 2. Association rule set derived for Sangli city[1]

Fu et al, provides the glossary for the researches on how the current time series data mining deployed and a guideline for the research direction for future investigation[3]. The time series is the collection of data made chronically. The characteristics include the large data, high dimensionality and update regularly. Thus the data mining on time series data need be summarized into a series of steps to derive knowledge from it. The steps include Time series representation and indexing, Similarity measure, Segmentation, Visualization, Mining in time series. The time series representation and indexing includes the reduction of dimension by one of the techniques namely sampling,piecewise aggregate approximation(PAA),adaptive piecewise constant approximation (APCA),linear interpolation, perceptually important points (PIP). The representation of the time series can be from numerical to symbolic form. The efficiency of the indexing in depending on the precision of approximation in reduced dimension space.

The similarity measure is fundamental step for time series analysis and data mining. The similarity measure can be done by whole sequence matching and subsequence matching. The segmentation can be considered as preprocessing step in time series analysis. Visualization is the mechanism of presenting the proposed time series for the future analysis. A few tools used for visualization are cluster and calendar-based visualization tool, Time Searcher tool, spiral visualization tool and VizTree. The mining in time series is done as Pattern discovery and clustering, Classification, Rule discovery and Summarization[3].

Qian et al, proposes a novel data-representation scheme and a framework for clustering algorithms [4]. Clustering is an important tool in data mining. It is used to discover the grouping structure which is present in a set of objects or data sets. It has applications in area, granular computing, text mining, information retrieval, bioinformatics, customer analysis, Web data mining, and scientific data exploration. The objective of clustering is to group a set of objects into meaningful clusters. The clusters have objects which are similar to each other and other clusters are very different. Various clustering algorithm are to deployed for this purpose.

For the numeric data, the k-means-type algorithms are very representative, that effectively and efficiently organize the objects into several clusters. In k-means algorithms, all objects are depicted by the features with numeric domains. Therefore, the objects described can be considered in a Euclidean space, and the similarity or the dissimilarity between two objects can be measured by a Euclidean distance, a cosine distance, and so on.

Pseudocode for k-means algorithm
Make initial guesses for the means m1, m2, ..., mk
Until there are no changes in any mean
Use the estimated means to classify the samples into clusters
For i from 1 to k
Replace mi with the mean of all of the samples for cluster i
end_for
end_until

Figure 3. Pseudocode for k-means algorithm.

Devi et al, proposes a survey on forecasting electrical consumption in Tamil Nadu [2]. In this paper the various data mining techniques are discussed. The importance of the forecast of the electricity in Tamil Nadu are being stated. The prediction techniques such as Fuzzy Logic, Particular Neutral network and Expert Systems are highly efficient. The survey details the load prediction in Tamil Nadu. The procedure of data mining is also explained in this paper[2].

III. PROJECT DESCRIPTION

The proposed system uses individual household electricity consumption data to gain knowledge about usage of electricity. The drawback of the previous established system is inability to predict the electricity consumption. The proposed system predicts the electricity consumption. The clusters are formed in the household electricity data and using the clusters the comparison with input data is performed and current

consumption levels are predicted. The cluster comparison with other clustering methodologies is also been established. The electricity consumption datasets are used to forecast the electrical consumption in the project. The prediction of the electrical consumption is done along with the prediction of usage levels. The clustering methods such as k-means and hierarchical clustering has been deployed. The project implementation is been done in three modules. The three modules include

- Data collection and preprocessing.
- Clustering module
- Prediction module

The detailed description of the various modules are been given

- **Data Collection and Preprocessing**

The data needed for the processing and prediction of electricity has to be collected from a database which has details about the voltage and intensity of the consumption. The data are collected from one such data base and are been processed to be used.The dataset are been collected for an individual house. The preprocessing of the data collected has led to be used in the algorithms such as k-means to produce the required output.

Data preparation has been done by Handling messy, inconsistent and unstandardized data.
Combining data from multiple sources.
Reporting data that has been manually entered.
Dealing with data from unstructured source.

Global_active_power	Global_reactive_power	Voltage	Global_intensity
2082	189	992	53
2654	198	871	81
2661	229	837	81
2668	231	882	81
1807	244	1076	40
1734	241	1010	36
1825	240	1017	40
1824	240	1030	40
1808	235	907	40
1805	235	894	40
2198	229	794	59
2680	215	786	82
2586	219	807	78
2608	179	799	79
2001	191	1032	49
1666	121	1222	32
1609	56	1181	30
1689	58	1214	33
1607	2	1221	30
1838	2	1092	43
2921	2	777	93

Figure 4. Few records of the dataset for prediction of electrical consumption.

- **Clustering module**

The clustering module includes the implementation of k-means and hierarchical clustering to the obtain clusters which are used for the prediction. A comparison of the various clusters to obtain the higher accuracy is implemented. K-means is one of the simplest unsupervised learning algorithms that used to form clusters. The algorithm follows a simple and efficient way to classify a given data set through a certain number of clusters fixed a priori. The main objective is to define k centroids, one for each cluster. These centroids shoud be placed in a way so that different location causes different result.

Number of Clusters	Accuarcy
2	48.7 %
3	67.6 %
4	67.2 %

Figure 5. Accuracy Comparison of implementing k-mans

- **Prediction module**

The prediction module is the most important part of project. The prediction is done is in two ways. The prediction of usage level has been done and predict the electrical consumption using time series. The prediction of usage level such as low, high and medium.

The prediction of usage level has been updated to the load profile data. The update is done on the given data as well as prediction of usage level for an new data by comparison with the models formed by the clustering.

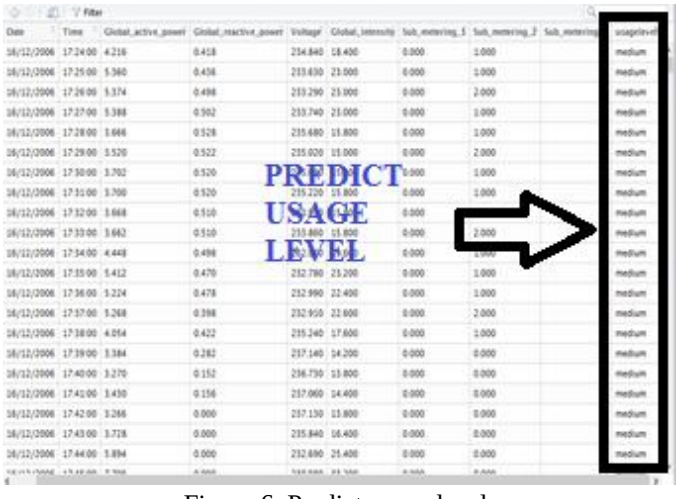

Figure 6. Predict usage level.

The prediction of next electricity consumption can be deployed using time series. Time series analysis comprises methods for analyzing time series data to extract meaningful statistics and other characteristics of data. Time series forecasting is the use of a model to predict future values based on previously observed values. The time series is used to predict the voltage and intensity of future electricity consumption.

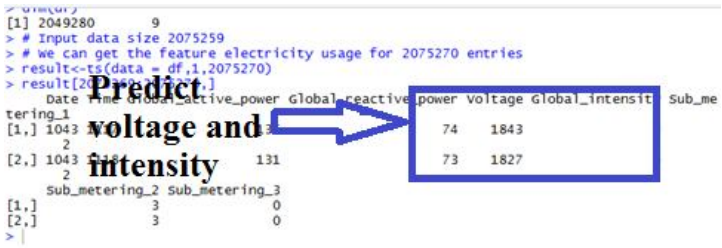

Figure 7. Predict electricity consumption.

V. CONCLUSION

The system has used individual household electricity consumption data to gain knowledge about usage of electricity. Another usage of the proposed system predicts the electricity consumption. The clusters are formed in the household electricity data and using the clusters the comparison with input data is performed and current consumption levels are predicted. The cluster comparison with other clustering methodologies is also been established. The comparison of the electrical consumption data to the clusters formed results in prediction of the usage levels.

REFERENCES

[1] Ravindra R. Rathod, Rahul Dev Garg," Regional electricity consumption analysis for consumers using data mining techniques and consumer meter reading data", Electrical Power and Energy Systems, Volume 78, June 2016, Pages 368–374.

[2] Dr.Mrs.M.Renuka Devi,R.Manonmani,"Electricity forecasting using data mining techniques in tamil nadu and other countries– a survey", International Journal of Emerging trends in Engineering and Development, Issue 2, Vol.6 ,September 2012.

[3] Tak-chung Fu, "A review on time series data mining",Engineering Applications of Artificial Intelligence, Volume 24, Issue 1, February 2011, Pages 164–181.

[4] Yuhua Qian, Bing Liu, "Space Structure and Clustering of Categorical Data" , IEEE transactions on neural networks and learning systems, vol. 27, no. 10, october 2016.

[5] Camilo Ernesto López Guarín, Elizabeth León Guzmán, and Fabio A. González," A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining",IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje, Vol. 10, No. 3, August