ISSN: 2321-8169
1750 – 1753

Short Term Load Calculating for Smart Meter Data using Weighted KNN for Classification & Prediction

Ajinkya M. Chavan
Student, dept. Computer Engineering
JSPM's ICOER
Pune, India
Ajinkya.chavan@gmail.com

Prof. R.N.Phursule
Dept. of Computer Engineering
JSPMS's ICOER
Pune, India
rnphursule@gmail.com

Abstract— Load forecasting has always played important role in determining the operational efficiency of electricity utility companies. An unsatisfied customer and regulator due to insufficiently planned load forecast leads to direct hit on the bottom-line of these companies. Sensing this opportunity, researches have applied different data mining and forecasting techniques. Data explosion due to rollout of smart meters have further provided additional edge required to these techniques. A simple and efficient solution, however, always remains a challenge. In this paper, we have discussed an approach that takes into consideration weighted KNN using Euclidian and dice coefficient based classification and forecasting for predicting the power demand.

Keywords-Data *Mining, Forecasting, Smart meters, Weighted KNN, Dice, Euclidian*

*****

I. Introduction

Since the deregulation in major developed markets, electricity utility companies have always thrived to improve demand forecast accuracy. An inadequate power purchase would force these companies to buy the power at higher rates. Additionally, it would invite for penalties from regulators. Power demand is complex function and is dependent on various variables like region type (residential/commercial), day type (weekend/weekday/long weekend), weather (temperature, humidity, wind chill, visibility etc), time of the day (morning, afternoon, evening). A different combination of these variables results in different output. For example a power demand on a Monday morning in a residential region will be different that on a weekend morning. Similarly a power demand on a hot summer afternoon will be different than that on an average temperature day. Thus these variables show a close association between them.

Smart meter rollout has gained a considerable momentum in the developed countries since past few years. These meters generate reads at an interval of every 15 minutes and communicate the same to the company for billing and other purposes. As a result of these, now the companies have near real time information about the load cycle and consumption patterns of the customers. The real BIG data generated by these meters have significant information contents which can be revealed using focused data mining approaches. Asset analytics, theft analytics, behavior analytics are some these areas which have gained focus. However what interests these companies more is the demand analytics to ensure an interrupted flow of supply to the customers.

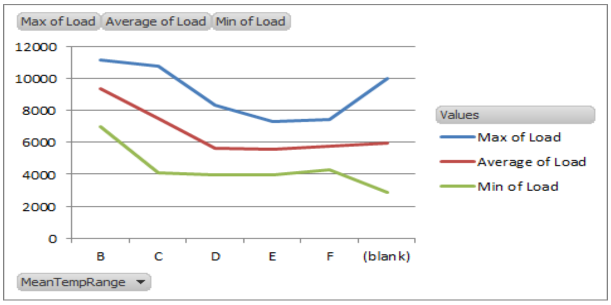
II. LITERATURE SURVEY

Load Demand prediction, as such, was there since earlier times however it is hardly surprising that there is considerable work happening in the areas of data mining & forecasting for electric demand prediction since smart meters have been put to use. Various approaches have been suggested for the same using range of data mining techniques like SVM, KNN, ANN, Wavelet transform and Curve Fitting methods.

Ranganathan & Nygard suggested an approach using M5 decision tree classifiers to predict the demand. However it didn't take into consideration other factors like weather, nature of the day. Wen-Chen and Yi-Ping Chen applied heat index (temperature & humidity) along with ANN to predict the demand but there was no mention of the smart meter data in the approach. Hourly load forecasting using ANN used only average daily loads and thus limiting accuracy. Chakhchoukh & Panciatici considered stochastic characteristics of load and proposed use of RME (ratio of median based estimator) as against traditional double exponential smoothing but it didn't considered any data mining techniques to its advantage. Apparently the RME approach worked well for normal days. Xiaoxia Zheng implemented a modelling approach based on least squares support vector machine (LS SVM) within the Bayesian evidence framework for short-term load forecasting. Under the evidence framework, the regularization and kernel parameters can be adjusted automatically, which can achieve a fine tradeoff between the minimum error and model's complexities.Yan Cao, Zhong Jun Zhang & Chi Zhou proposed SVM based model that takes weather factods into considertion to improve the accuracy. Koo & Kim used smart meter data along with KNN & forecasting models to further improve the accuracy but didn't consider the weather factors into consideration. Besides the KNN was on stationary pool of data and didn't consider any continuous flow

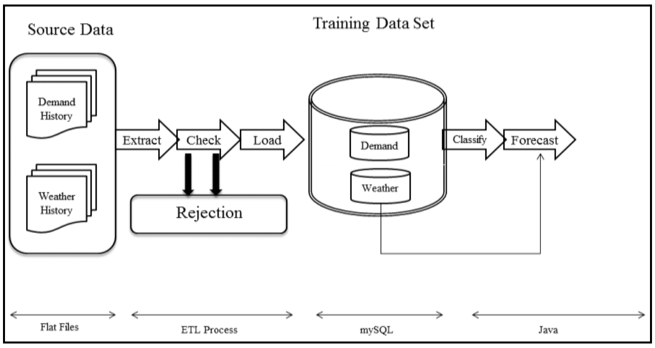
III. DATA ANALYSIS

In order to propose a robust system, we have done an in-depth analysis existing load patterns obtained from data publicly made available by NYISO. An analysis of 2+ years of consumption data of 6 regions (approx. 1M+ records) indicated that the load cycle varies more with the weather factors than the day type. Region type has its own complexity associated. Below picture represents the variation of consumption against temperature grouped in different categories.



IV. APPROACH

In line with the above discussion, we have proposed a simple yet powerful approach as depicted in the picture below.



The key components of this approach as discussed in the below section

A. Data Preparation

The primary source data for our approach is consumption data from NYISO website which is freely available. The dimensional data for the demand analytics is weather data which is available on the weather websites and local holiday data which again available on internet.

a) Consumption Data: A sample of consumption data is as shown in table below

TABLE I

Time Stamp	Time Zone	PTID	Load
1/1/2011 0:00	EST	N.Y.C.	61761
1/1/2011 0:05	EST	N.Y.C.	61761
1/1/2011 0:10	EST	N.Y.C.	61761
1/1/2011 0:13	EST	N.Y.C.	61761
1/1/2011 0:14	EST	N.Y.C.	61761
1/1/2011 0:20	EST	N.Y.C.	61761
1/1/2011 0:25	EST	N.Y.C.	61761
1/1/2011 0:30	EST	N.Y.C.	61761
1/1/2011 0:35	EST	N.Y.C.	61761
1/1/2011 0:40	EST	N.Y.C.	61761
1/1/2011 0:45	EST	N.Y.C.	61761
1/1/2011 0:50	EST	N.Y.C.	61761
1/1/2011 0:55	EST	N.Y.C.	61761
1/1/2011 1:00	EST	N.Y.C.	61761

b) Weather Data: For the purpose of this study we have considered only temerature and humidity as the prime factors from weather. The source weather data was rather in a crude format and was further transformed to obtain a format as shown in table below

TABLE I I

YYYY	MM	DD	HH	MI	Temp.	Humidity
2012	01	01	00	51	7.8	73
2012	01	01	01	51	7.2	76
2012	01	01	02	51	7.2	76
2012	01	01	03	51	7.2	74
2012	01	01	04	51	5.6	82
2012	01	01	05	51	6.7	76
2012	01	01	06	51	6.7	76
2012	01	01	07	51	6.7	76
2012	01	01	08	51	7.2	74
2012	01	01	09	51	8	71
2012	01	01	10	51	10	63
2012	01	01	11	51	10.6	59
2012	01	01	12	51	10.6	59
2012	01	01	13	51	10	66
2012	01	01	14	51	10	66
2012	01	01	15	51	10	68
2012	01	01	16	51	9.4	71
2012	01	01	17	51	9.4	77
2012	01	01	18	51	9.4	83
2012	01	01	19	51	9.4	83
2012	01	01	20	51	9.4	90
2012	01	01	21	51	10	86

B. Data Transformation

Both the data sets were aligned together so as to match the granularity and a single merged data file was prepared. Further, using standard ETL transformations, non-numeric attributes were converted into corresponding numeric representation.

C. KNN Classification & Forecasting

A nearest neighbor classifier is a technique for classifying elements based on classification of elements in the training set that are most similar to the test example. With K nearest neighbor technique, this is done by evaluating the K number of nearest neighbors. In pseudo-code, KNN can be expressed as:

```
For each object X in the set
    -   Calculate distance D(A,B) between A and every other
        object B
    -  Neighbors = the K neighbors in the training set closest to test instance A
    -  Get the majority vote from neighbors
End For
```

Above algorithm can be further modified and used for prediction (extrapolation) of quantitative data (e.g. time series). In classification, the dependent variable Y is categorical data. In this section, the dependent variable has quantitative values.

Here is step by step on how to compute K-nearest neighbors KNN algorithm for quantitative data:

- Determine number of nearest neighbours to be used
- Calculate the distance between the test sample and all the training samples
- Determine nearest neighbours based on minimum distance
- Gather the values of the nearest neighbours
- Use average of nearest neighbours as the prediction value of the query instance

The accuracy of KNN also depends on relevance of the neighbors. An incorrect choice of the dimensions can skew the entire process and provide inaccurate results. Hence it is more important to provide the weights to these dimensions i.e. a dimensionally weighted KNN approach is required.

Additionally no. of neighbors to be used & similarity measure to be used also impact the final accuracy. While, no. of neighbors to be used is largely dependent on experimental results, the similarity can be measured using distance based approach or coefficient based approach. In a distance based approach, different approaches like Euclidean distance, Manhattan distance etc can be used. In coefficient based approach, Simple Matching Coefficient (SMC), Jaccard Coefficient, Rao's Coefficient, Dice coefficient etc. can be used. In this approach, we have used Euclidian Distance & Dice Coefficient

a) Euclidian Distance: In Euclidian Distance, distance of the test record is calculated using following formula

$$Ed = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Here X & Y are the two sets to be compared. In other words, this formula measures the distance between the two records based on distance between various attributes. The smaller the value of Ed, the better is match

b) Dice Coefficient: In Dice Coefficient method, distance of the test record is calculated using following formula

$$S(A,B) = \frac{2|x \cap y|}{|x| + |y|}$$

Here X & Y are the two sets to be compared. In other words, this formula measures the distance between the two records based commonality or the intersection points. The value lies between 0 and 1. Larger the value, better is the match

D. Accuracy Measurement:

Forecast accuracy can be measured using MSE, RMSE & MAPE. MSE is mean squared error, RMSE is root means squared error and MAPE is mean absolute percentage error. To compare the forecasted results with actual data and thus to compare both the models, MAPE will be used which is calculated using below formula

$$MAPE\ (\%) = \frac{1}{N}\left[\frac{|Z_t - X_t|}{X_t}\right] \times 100\ \%$$

Where Zt is Forecasted load, Xt is Actual load & N is Forecasting Number

V. CASE STUDY & RESULTS

The forecasting approach discussed above is used to predict the Jan 2013 load demand. Training data considered is two 2012 consumption data & weather data. Tests were conducted for different values of K. Below table shows the comparison of MAPE value for both Euclidian & Dice coefficient method based forecasting. After applying the weightage factor, the results are improved.

DAY	EUCLIDIAN	DICE
SUNDAY	9.19	4.41
MONDAY	3.30	5.63
TUESDAY	6.52	3.98
WEDNESDAY	3.86	5.71
THURSDAY	4.19	3.12
FIRDAY	3.94	3.31
SATURDAY	7.62	4.70

As shown in above results, the Dice coefficient based weighted coefficient is working better in majority of the cases as compared to the Euclidian based approach

REFERENCES

[1] JBon-Gil Koo, Min-Seok Kim, Kyu-Han Kim, Hee-Tae Lee and June-Ho Park, "Short Term electric load forecasting using data mining techniques" in *Proc. of 7th ISCO2013, 2012 IEEE*

[2] Prakash Ranganathan, Kendall Nygard, "Smart Grid Data analytics for Smart Meters" in *Proc. of 2011 IEEE Electrical Power and Energy Conference*

[3] Wen-Chen Chu, "Multiregion short term load forecasting in consideration of HI and load/weather diversity" in *Proc. of IEEE transactions on industry applications*

[4] Hongfei Li, "Usage analysis for smart meter management" in *Proc of 2011 IEEE Conference*

[5] Daswin De Silva, Xinghuo Yu,Damminda Alahakoon, and Grahame Holmes, “A Data Mining Framework for Electricity Consumption Analysis From Meter Data” *IEEE Trans.on Ind. Informatics*, vol. 7, no. 3

[6] Yang Wang,, Qing Xia, Chongqing Kang, “Secondary Forecasting Based on Deviation Analysis for Short-Term Load Forecasting” *IEEE Trans..on Power Systems*, vol. 26, no.2.

[7] XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda , Geoffrey J.McLachlan, Angus Ng, Bing Liu, Philip S. Yu, ZhiHua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, "Top 10 algorithms in data mining", Knowl Inf Syst, 2008 14, pp. 1-37

[8] Jiawei Han and Micheline Kamber, “Classification and Prediction”in *Data Mining: Concepts and Techniques 2nd ed., San Francisco, CA* The Morgan Kaufmann, 2006

[9] http://www.nyiso.com/public/markets_operations/market_data/load_data/index.jsp

[10] Report from Pike research, http://www.pikeresearch.com/research/smartgrid-data-analytics

[11] National Climate Data Center [Online]. Available: http://www.ncdc.noaa.gov/oa/ncdc.html