Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Forecasting Electricity Demand for Smart Meter Data Using KNN Based Classification and Weather Factors

Ajinkya M. Chavan* ICOER, Pune University
Prof. R.N.Phursule ICOER, Pune University
Prof. S.A.Chavan PVPIT, Pune University

Abstract— Accurate demand prediction for electricity is a very sensitive area for every electricity retailor. A gap in the same invites penalty in terms of last minute high price purchases and also leads to customer dissatisfaction. Recent implementation of Smart meters has resulted in an exponential rate of growth in the consumption data volumes. Instead of one meter read per billing cycle, now a meter read is generated every 15 minutes. This data explosion exhibits all sorts of trends and provides a good opportunity for these companies to improve the accuracy. The power demand is generally driven by weather factors like temperature, humidity, wind-chill. Also nature of the day, region, customer type etc. adds to the complexity. This paper proposes an approach to prepare the raw data and improve the forecast accuracy using KNN based classification. It compares the results using both Euclidian distance & Dice coefficient based similarity measures

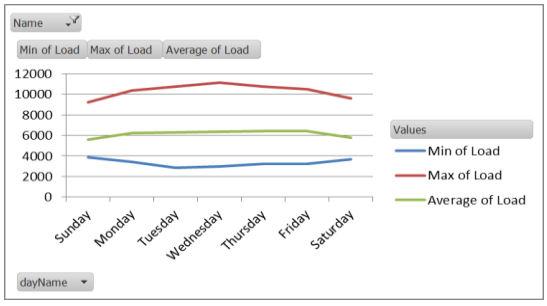
Keywords— Data Mining, Forecasting, Smart Meters, KNN

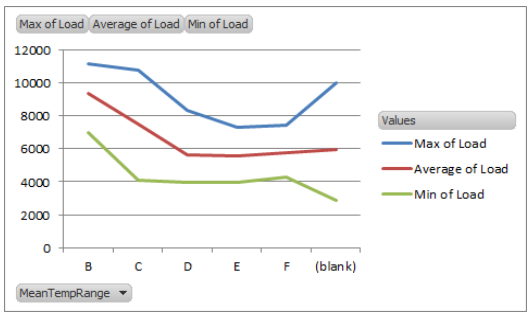
I. INTRODUCTION

One of the critical areas for utilities companies across the worlds is demand prediction. Future power demand, if not predicted correctly, will result in high cost power purchases at the last moment creating financial burden to the company. As a result, demand prediction has traditionally got more focus.

Weather conditions and Day type play important role in determining the power demand. For example a humid or sunny weather will demand more utilization of air coolers on the contrary a freezing weather will need more room heaters. For the Day type, power demand on a weekday will be different from a weekend. Also power demand on a holiday on a weekday will follow a different pattern. Power demand in residential areas will be different from industrial areas. Thus all these factors play key role in determining a correct predictions or in other words, there exists an association between these.

Since inception, smart meters have generated considerable interest to the utilities. These meters, unlike their traditional counterparts, can generate the meter reads at an interval as short as 15 minutes and have the capability to communicate this read back to the company for billing and maintenance purpose. This has resulted in a massive increase in the data volumes generated at the company, resulting in a data explosion. Graph 1 is consumption data as plotted against for individual day-type for each of the month for city of New York. Grpah2 is consumption data plotted against temperature





Both the above graphs clearly indicates more reliance on the weather factor as compared to the day-type while calculating the power consumption

Data mining is that branch of computer science which deals with analyzing this vast amount of data to arrive at meaningful patterns. Analyst across the globe, have already embarked on a journey to utilize smart meter data to derive these values. Though customer analytics, asset analytics, event analytics are few of the immediate benefits, what has gathered more attention is to use this data for correctly predicting the demand or in another words demand analytics for utility industry. In this paper, we will provide a novel approach to ensure data classification for the demand prediction using historical consumption data from smart meters along with weather factors.

II. LITERATURE SURVEY

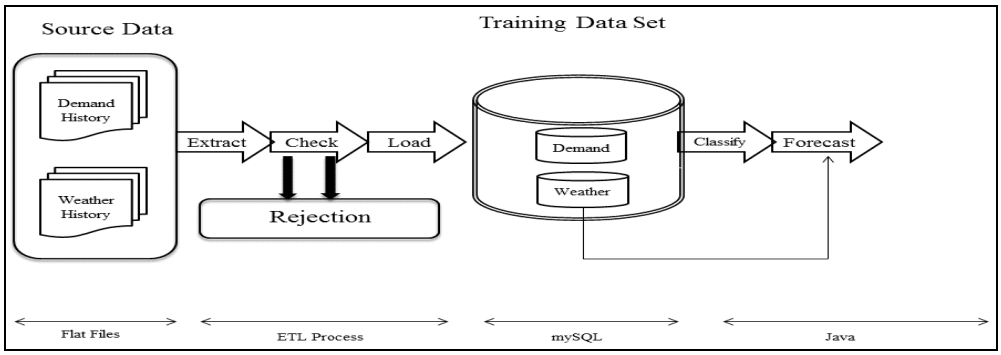
Load Demand prediction, as such, was there since earlier times however it is hardly surprising that there is considerable work happening in the areas of data mining & forecasting for electric demand

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. APPROACH DETAILS

In line with above discussion, picture below depicts our approach and various components in the same



A. 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(X,Y) between X and every other object Y
  Neighbors = the K neighbors in the training set closest to X
  Get the majority vote from neighbours
End For

We can extend the K-Nearest Neighbor (KNN) algorithm for smoothing (interpolation) and 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 neighbors to be used i.e. K
- Calculate the distance between the test sample and all the training samples
- Determine nearest neighbors based on minimum distance
- Gather the values of of the nearest neighbors
- Use average of nearest neighbors as the prediction value of the query instance

The accuracy of KNN depends on two important factors – no. of neighbors to be used & similarity measure to be used. 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

B. 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

C. 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

IV. CASE STUDY

A. Historical Consumption Data

Historical power demand at different hours of the day can be obtained using the smart metering data gathered at different hours of the day. This can be used as input data set to predict the demand on the system. For the purposed of this project, we have used 3 years historical consumption data (2011, 2012, 201) of New York City. This data is publicly available over the internet for analyst. A sample of data is as shown in the picture below

TABLE I: CONSUMPTION DATA

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

This consumption data was transformed to obtain required classification parameters such as Weekday, Month, Holiday factors.

B. Historical Weather Data

Historical weather data is downloaded from the weather website. For the purpose of this paper, we have considered historical values of hourly temperature & humidity as the driving factors. Further these weather factors were aligned to respective data items from consumption values

C. Data Transformation

Since KNN prediction needed numeric attributes for classification, all the non -numeric attributes were transformed into equivalent numeric values. Further, factors like year, wind direction etc. have no binding on the consumption and hence were ignored. A sample combined & transformed data file that served as input is as shown in the table below –

TABLE II TRANSFORMED SOURCE DATA

Month	Day	Hour	Minute	Week Day	Current Holiday	Prior Holiday	Temperature	Humidity	Load
12	7	12	35	5	0	0	4.4	93	6532.1
12	7	12	40	5	0	0	4.4	93	6519.8
12	7	12	45	5	0	0	4.4	93	6530.5
12	7	12	50	5	0	0	4.4	93	6520.3
12	7	12	55	5	0	0	4.4	93	6529.6

12	7	13	0	5	0	0	5.6	92	6537.4
12	7	13	5	5	0	0	5.6	92	6512.6
12	7	13	10	5	0	0	5.6	92	6527.6
12	7	13	15	5	0	0	5.6	92	6534
12	7	13	20	5	0	0	5.6	92	6550.2

D. Accuracy Measurements

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. RESULTS

The forecasting approach discussed above is used to predict the Jan 2013 load demand. Training data considered is 2012 consumption data & weather data. Tests were conducted for two different values of K (k=4 & k=6). Below table shows the comparison of MAPE value obtained using K = 4 & K = 6 for both Euclidian distance & Dice coefficient methods. After applying the weightage factor to distance, impact of value K is observed to be reduced

Month	K = 4	K = 6
Euclidian Distance	4.596%	3.572%
Dice Coefficient	4.208%	3.088%

VI. CONCLUSIONS & FUTURE SCOPE

Here, we have discussed an approach to improve the accuracy for predicting the power demand based on KNN method for smart meters data. The correctness of results largely depends on the accuracy of weather parameters of the test records. We have tested the KNN using both Euclidian distance & Dice coefficient similarity measures for different values of k. Here we have used a non-iterative method to predict hourly demand for next 24 hours i.e. demand prediction for a particular hour of day confines itself to the previous days consumptions and doesn't consider the calculated demand prediction of the prior hours of the same day

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

