International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Data Mining for Agriculture Electricity Consumption Forecasting for Rural Area of Gujarat

Rajnikant Pandya
Student-Research Scholar (Computer Science), Gujarat Vidyapith-Ahmedabad,Gujarat, India
Dr. Ajay Parikh
Professor & Head, Department of Computer Science, Gujarat Vidyapith-Ahmedabad, Gujarat, India

Abstract
In present days, the increasing of electricity consumption trend is very important in decision making for electricity production planning for government, particularly in Gujarat. For fulfilling the demand of electricity, company has to plan the production, and to know the demand in advance, accurate prediction of electricity requirement is needed. Accurate prediction is major task for future planning of any Company.In order to effecting forecasting , this paper present the design and development of model for agriculture electricity load forecasting. In this paper, we present the study of various data mining techniques. The steps required for this model include data collection, preprocessing, handling missing values etc. The final data after preprocessing is fed into Multiple linear regression(MLR), Support vector regression(SVR), KNN and Multilayer perceptron (ANN) to predict agriculture electric power requirement. At the end we compare and conclude the result of these techniques.

Keywords:- ANN, Data Mining , Electricity consumption forecasting, KNN, MLR, Prediction, SVR.

I. Introduction

India is an agriculture land country. In Most of cases, Indian rural areas are depends on agriculture. Development dependencies of rural areas are on agriculture. So to develop rural area, it is mandatory to develop agriculture. Development of agriculture is depends on seasonal crop. In agriculture, there is main three crop i.e. winter crop, summer crop and monsoon crop. If farmer yield crop of all seasons then it is good economical contribution in the development of rural area. Except monsoon crop, other two required water resources like well and bore well water. If there is sufficient underground water then there is a need of water pumping and this task required much more electric power supply.

If electricity supply is sufficient in agriculture then farmer can get good yield of various crop of every season and this will effect development of rural area.

In this paper, we forecast electric power requirement of agriculture in rural area. The main objective of this paper is, to develop one model which forecast electricity usage of agriculture. For prediction of electricity requirement, we use historical data of electricity consumption and weather data. We collect data for timeline Apr-2006 to Jan-2018.

Data mining is the process of analyzing data from different perspective and summarizing it into useful information[11]. Data mining can discover knowledge as well as some patterns and relationship among a variety of parameters. This is very supportive in decision making for company. Data mining has been effectively used in electricity consumption forecasting.

Received: 8 September Revised: 15 September Accepted: 20 September
Index in Cosmos
October 2018 Volume 8 Number 4
UGC Approved Journal

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

The main contribution of this paper is, using the different technique of data mining, we predict electricity demand and develop good fit model for prediction of agriculture electric power requirement for rural area.

A precise forecast becomes more constructive in developing power supply tactic, scheduling and management.

II. Background study and related work.

There are various studies done on the relationship among power consumption and prediction of power usage as well as forecasting the electricity price. There are many studies that carry the use of data mining techniques for electricity load forecasting.

Hossein Daneshi ,Mohammad Shahidehpour, Azim Lotfjou Choobbari et al [1] have taken the issue of Long-term load forecasting in electricity market they used fuzzy set to ANN and compare with traditional methods to enhanced forecasting result.
BahmanKermanshahi, HiroshiIwamiya et al [2] have taken the problem of pick electric load forecasting in Japan up to 2020. The data used for this forecasting includes parameter gross national product, gross domestic product, population, number of households, number of air-conditioners, amount of CO2 pollution, index of industrial production, oil price, energy consumption, and electricity price. The data used are: actual yearly, incremental growth rate from the previous year, and both together (actual and incremental growth rate from the previous year). They used two ANNs, a three-layered back-propagation and a recurrent neural network and achieved result for 2010 and 2020 are predicted to be 225.779 and 249.617 GW. With structure reform, the demands for 2010 and 2020 are predicted to be 219.259 and 244.508 GW.
H.Y.Yamina , S.M.Shahidehpourb , Z.Lic et al [3] have taken the issues of adaptive short – term electricity price forecasting. They include time factors, load factors, reserve factors, and historical price factor for solving problem. They used ANN and performance matrix are MAPE.
Jun Hua Zhao ; Zhao Yang Dong ; Zhao Xu ; Kit Po Wong et al[4] solve interval forecasting of the electricity price using support vector machine (SVM) and Maximum likelihood estimation (MLE) is used to estimate model parameters.
E. Gonzalez-Romera, M.A. Jaramillo-Moran,D. Carmona-Fernandez et al [7] have solve the issues of Monthly Electric Energy Demand Forecasting Based on Trend Extraction. They used two neural networks to forecast two new series trend and fluctuation separately and got better result than one neural network and ARIMA.
Ching-Lai Hor ; S.J. Watson ; S. Majithia et al [8] have taken the issues to Analyzing the impact of weather variables on monthly electricity demand in England and Wales. They used multiple regression model and could achieve MAPE 2.60% for 1989 to 1995 and 2.69% for 1996 to 2003.
H.S. Hippert ; C.E. Pedreira ; R.C. Souza et al [9] have taken the problem of short term load forecasting. They examine the various papers which use the ANN algorithms to solve this problem. They used collection of papers published between 1991 and 1999.

Received: 8 September Revised: 15 September Accepted: 20 September
Index in Cosmos
October 2018 Volume 8 Number 4
UGC Approved Journal

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

I. Moghram, S. Rahman et al [10] Analysis and evaluate five short-term load forecasting techniques. These are multiple linear regression, stochastic time series, general exponential smoothing, state space and Kalman filter and a knowledge-based approach. They compare all this techniques

III. Suggested model for electricity consumption forecasting

A. Model Design

The data used in the model was collected on monthly bases. The forecast model consists of following phases.

Phase 1: Data Collection
Phase 2: Preprocessing of Data
Phase 3: Handling Missing Value
Phase 4: Prediction

B Phase

Phase 1: Data Collection

In this section what we have done to collect proper data. Following things are done.
Data Collected from PGVCL circle office for the timeline of April 2006 to January 2018

1.1 Electricity Usage Data:

Actual document file that include details like power generation, power consumption and power loss are collected for amreli rural sub division. These data are collected feeders wise (total feeders are 38) . in these document, data has been collected with various parameters like feeder category, feeder name ,total sent, sum of HT sold , sum of LT sold, sum of AG ASSES , sum of AG RECORD, sum of TEMP units, and difference. Then these parameters are converted into three variable i.e. total power generation, power consumption and power loss. The measurement of these variables are in Mus(Million Unit). Finally these three variables include in final data set.

1.2 Weather Data

Another type of data has been collected was qualitative in this sense it was about the weather condition of the of various villages which actually gathered from various WEB resources, those data generally represent details about high temperature , low temperature , the environment over there etc.

Weather data Collected from online resources for two time line April 2006 to July 2014.and august 2014 to january 2018 [12][13]. Following are various weather attributes

-High temperature (Celsius)
-Low temperature (Celsius)

Received: 8 September Revised: 15 September Accepted: 20 September
Index in Cosmos
October 2018 Volume 8 Number 4
UGC Approved Journal

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

-Humidity (Fraction)
-Wind Speed (m/s)
-Precipitation (mm)

All the data values of above variables are monthly average. Finaly these are included in final data set.

Phase 2: Preprocessing of Data

In this particular section refers to the phase that comes after data collection phase. In this phase we have matched up to following things for processing the actual data. In the data collection section whatever data has been collected about Amreli region which was daily data rather we converted those data into monthly average.

The aforementioned data has been combined with the power consumption data of agriculture

The data of total power consumption, power generation, and power loss and their values are being replaced by particular month's average or say mean value. Based on above all this data processing we created .CSV (Comma Separated Values) file for final use.

Phase 3: Handling Missing Value

The data plays an important role in the process of data mining. The quality of the collected data can directly improve the efficiency of the subsequent mining process. However, data collected in the real-world tend to be incomplete due to some missing values.

Applying the relevant principles of missing value analysis, simple methods like Mean/Mode. Imputation can be easily used to handle missing values

As consider in our data preprocessing stage there were two missing records as of year September 2006 and December 2008 for electricity data usage. Values of these two records are replaced by mean value of particular month.

Phase 4: Prediction

This phase is an important phase to extract information that helps us for prediction. After preprocessing and handling missing values, the data become suitable for forecasting the target function. There are many data mining techniques used for forecasting. We use Support Vector Regression, Multiple Linear Regression, KNN and Multilayer Perceptron (ANN) methods. Here our Test mode is 10-fold cross-validation for all techniques.

Received: 8 September Revised: 15 September Accepted: 20 September
Index in Cosmos
October 2018 Volume 8 Number 4
UGC Approved Journal

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

4.1 SMOreg - Support Vector Regression

The Support Vector Regression (SVR) uses the same principles as the Support Vector Machine (SVM) for classification. The idea of SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a nonlinear function.

One of the fundamental attributes of Support Vector Regression (SVR) is that as opposed to limiting the watched preparing mistake, SVR aim to limit the estimate error bound in order to accomplish execution.

A major benefit of using SVR is that it is a non parametric technique. Unlike other algorithms SVR does not depend on distributions of the underlying dependent and independent variables. Instead SVR technique depends on kernel functions.

SVM algorithms utilize different scientific functions that are characterized as the kernel. The capacity of bit is to accept information as information and change it into the required shape. Different SVM algorithms utilize different types of kernel functions. These functions can be different types. for example *linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.* The kernel functions return the inner product between two points in a suitable feature space.
Kernel functions have been introduced for sequence data, graphs, text, images, as well as vectors

A polynomial part, for instance, enables us to demonstrate highlight conjunctions up to the request of the polynomial. Circular assumption capacities permits to select circles (or hyper spheres) – in contrast with the linear piece, which enables just to choose lines (or hyper planes)

In machine learning, the polynomial kernel is a kernel function commonly utilized with support vector machines (SVMs) and other kernel models that represent the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized.
Polynomial Kernel (of degree d): k(x, z) = (x Tz) d or (1 + x Tz) d
SMOreg
weights (not support vectors):
+ 1.3133 * (normalized) Total_Power_Generation
- 0.3541 * (normalized) Power_Loss
+ 0.0008 * (normalized) High_Temperature(Celsius)
- 0.0002 * (normalized) Low_Temperature(Celsius)
- 0.0013 * (normalized) Wind_Speed(m/s)
- 0.0008 * (normalized) Humidity(Fraction)
+ 0.0015 * (normalized) Precipitation(mm)
+ 0.025

Received: 8 September Revised: 15 September Accepted: 20 September
Index in Cosmos
October 2018 Volume 8 Number 4
UGC Approved Journal

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Table 1 Coefficient and various errors of SVR

Correlation coefficient	0.9986
Mean absolute error	0.1468
Root mean squared error	0.7653
Relative absolute error	1.2334 %
Root relative squared error	5.3933 %

Above table show the correlation which is very near to 1. It show that all variable used in this model are much closed correlated. Other four error matrixes show the accuracy of prediction.

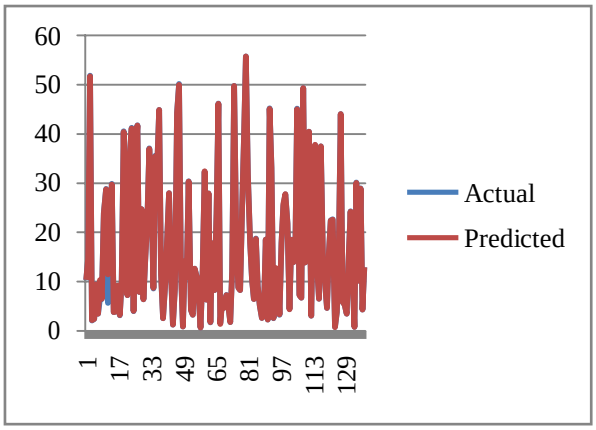

Figure 1. Comparison of actual and predicted agriculture power requirement using SVR

4.2 Multiple Linear Regression

Multiple linear regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or input variables).
The general form of multiple regression model shown as

$$Y_i = \beta\theta+\beta 1x1_i+ \beta 2x2_i+\beta 3x3_i+.....+\beta nxn_i+ \varepsilon_i$$

Where Yi is dependent variable, xi is the independent variables; βi is the regression coefficient of xi and εi is the random error.

Received: 8 September Revised: 15 September Accepted: 20 September
Index in Cosmos
October 2018 Volume 8 Number 4
UGC Approved Journal

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

To estimate the coefficients of the model, the predicted response is shown as

$$\hat{y}_i = b\theta + b1x1_i + b2x2_i + b3x3_i + \ldots\ldots + bnxn_i$$

The result of multiple regression models is shown as follows.

Linear Regression Model

Power_Consumption = -1.6027+ 1.0186 * Total_Power_Generation + -1.0272 * Power_Loss + 0.0579 * Low_Temperature(Celsius)

Table 2 Coefficient and various errors of MLR

Correlation coefficient	0.9985
Mean absolute error	0.2622
Root mean squared error	0.7657
Relative absolute error	2.2026 %
Root relative squared error	5.3963 %

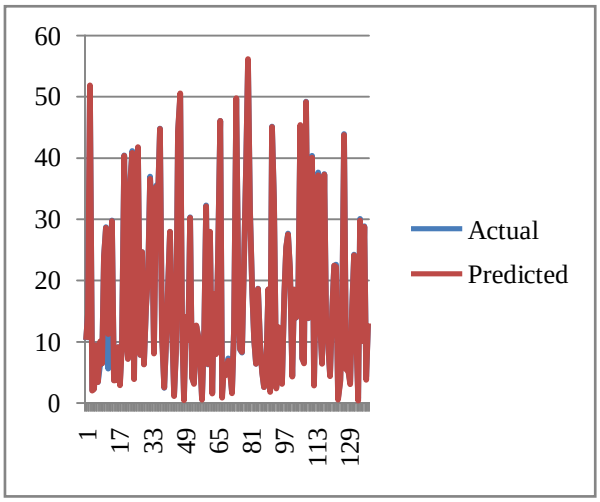

Figure 2. Comparison of actual and predicted agriculture power requirement using MLR

Received: 8 September Revised: 15 September Accepted: 20 September
Index in Cosmos
October 2018 Volume 8 Number 4
UGC Approved Journal

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

4.3 K Nearest Neighbors

K nearest neighbors is one of the non-parametric techniques used for classification and regression it implies a straightforward calculation that stores every accessible case and assume the numerical target in view of a comparability measure using distance functions. A basic implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors.

The implementation of kNN is purely depending on how you choose k factor for the prediction. If k is chosen to be too small it is sensitive to noise points if k is chosen larger works well, but too large may include majority points from other classes.

For the prediction, kNN used various distance function based on type of training sets. for the continuous data sets mostly used Euclidean distance function to find the distance between two neighbors other popular distance function for continuous data sets Manhattan Distance and Minkowski Distance and for the categorical data sets it used hamming distance.

Table 3 Coefficient and various errors of KNN

Correlation coefficient	0.9356
Mean absolute error	3.9298
Root mean squared error	5.0883
Relative absolute error	33.0084 %
Root relative squared error	35.8581 %

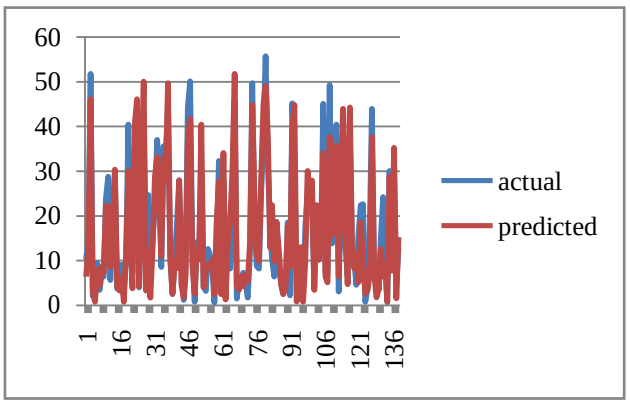

Figure 3. Comparison of actual and predicted agriculture power requirement using KNN

Received: 8 September Revised: 15 September Accepted: 20 September
Index in Cosmos
October 2018 Volume 8 Number 4
UGC Approved Journal

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

4.4 Multilayer Perceptron (Artificial Neural Network)

Tthis algorithm basically takes reference from previous instance and its inputs and readout the result, based on the result it calculates and find out proper conclusion. This algorithms concept arrived from the working of brain neurons that are connected by a network called Neural Network.
In this particular situation we are going to forecast or measure future energy usage based on past usage patterns.

Table 4 Coefficient and various errors of ANN

Correlation coefficient	0.9945
Mean absolute error	1.0606
Root mean squared error	1.6155
Relative absolute error	8.9086 %
Root relative squared error	11.3845 %

Above table show the correlation and other four error matrixes to show the accuracy of prediction.

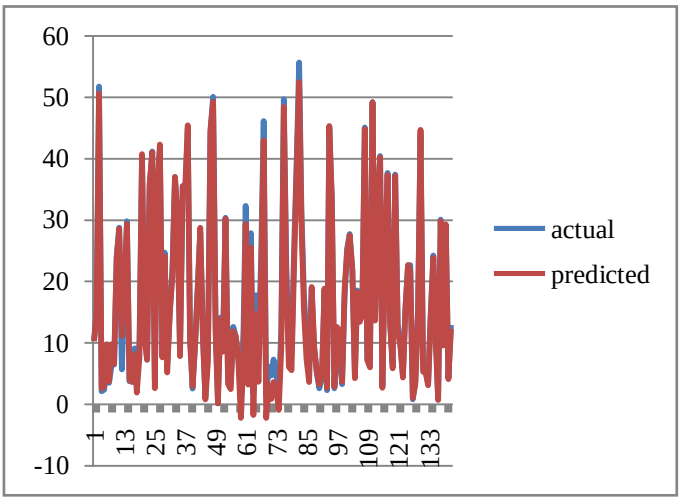

Figure 4. Comparison of actual and predicted agriculture power requirement using ANN

IV. Result and conclusion

There are no any exact rules to prove the best prediction model, but most appropriate one was selected by selecting the model with the lowest error. Here we used main four errors for selecting best fit model. All errors are negatively oriented scores, which mean lower values are better. The mean absolute error value is the average absolute error value. Root mean squared error is standard deviation of the prediction errors. Relative

Received: 8 September Revised: 15 September Accepted: 20 September
Index in Cosmos
October 2018 Volume 8 Number 4
UGC Approved Journal

absolute error is same like relative squared error. Root relative squared error is a square root of relative squared error. We compare the result of these different four matrixes to see the variation on result and the result shows that, SVR (Support Vector Regression) model is superior to other approaches.

References

1. Hossein Daneshi , Mohammad Shahidehpour , Azim Lotfjou Choobbari "Long-term load forecasting in electricity market" IEEE – 2008

2. BahmanKermanshahi , HiroshiIwamiya "Up to year 2020 load forecasting using neural nets" (International Journal of Electrical Power & Energy Systems) Vol 24 , Issue 9, November 2002, P.p 789-797

3. H.Y.Yamina , S.M.Shahidehpourb , Z.Lic "Adaptive short-term electricity price forecasting using artificial neural networks in the restructured power markets" IEEE-2004 Vol 26, Issue 8, October 2004, P.p. 571-581

4. Jun Hua Zhao ; Zhao Yang Dong ; Zhao Xu ; Kit Po Wong "A Statistical Approach for Interval Forecasting of the Electricity Price" IEEE – 2008 Vol: 23, Issue: 2, May 2008 P.p. 267 - 276

5. O. Landsiedel ; K. Wehrle ; S. Gotz "Accurate prediction of power consumption in sensor networks" IEEE – 2005

6. Tùng T. Kim ; H. Vincent Poor "Scheduling Power Consumption With Price Uncertainty" IEEE-2011, Vol:2, Issue: 3, Sept. 2011, P.p: 519 - 527

7. E. Gonzalez-Romera ; M.A. Jaramillo-Moran ; D. Carmona-Fernandez "Monthly Electric Energy Demand Forecasting Based on Trend Extraction" IEEE,Vol : 21, Issue: 4, Nov. 2006 ,P.p. 1946 - 1953

8. Ching-Lai Hor ; S.J. Watson ; S. Majithia "Analyzing the impact of weather variables on monthly electricity demand" IEEE 2005 Vol : 20, Issue: 4, Nov. 2005 P.p. 2078 - 2085

9. H.S. Hippert ; C.E. Pedreira ; R.C. Souza "Neural networks for short-term load forecasting: a review and evaluation" IEEE Vol : 16, Issue: 1, Feb 2001 P.p. 44 - 55

10. I. Moghram ; S. Rahman "Analysis and evaluation of five short-term load forecasting techniques" IEEE-1989 Vol : 4, Issue: 4, Nov. 1989,P.p. : 1484 - 1491

Received: 8 September Revised: 15 September Accepted: 20 September
Index in Cosmos
October 2018 Volume 8 Number 4
UGC Approved Journal