International Journal of Computer Engineering and Information Technology

VOL. 8, NO. 11, November 2016, 213–219
Available online at: www.ijceit.org
E-ISSN 2412-8856 (Online)

Application of Text Mining for Faster Weather Forecasting

Suvendra Kumar Jayasingh[1], Jibendu Kumar Mantri[2] and P Gahan[3]

[1] Biju Patnaik University of Technology

[2] North Orissa University

[3] Sambalpur University

[1]sjayasingh@gmail.com, [2]jkmantri@gmail.com, [3]pgahan7@gmail.com

ABSTRACT

Weather forecasting is a challenging problem in predicting the state of the climate for a future time at a given location. Weather is the state of atmosphere which is measured based on a scale of hot or cold, dry or wet, storm or calm and cloudy or clear. Human kind has tried a lot since ancient times to anticipate the future climate. It is why climate change prediction has become a very important task to the scientists and researchers of many countries. The weather is a continuous, data-intensive, multi dimensional, dynamic and chaotic process [6] and these characteristics made the weather forecasting a challenging job in the world. In order to make accurate prediction, many scientists have tried to forecast the meteorological behaviors. The objective of the research is to predict more accurately the meteorological characteristics. This article gives importance on using the fuzzy field and Natural Language Generation (NLG) that will make the weather prediction in a better and faster manner.

Keywords: *Weather Forecasting, J48 Decision Tree, Text Mining, Multiple Linear Regression (MLR), Root Mean Square Error (RMSE), Waikato Environment for Knowledge Analysis (WEKA), Natural Language Generation (NLG), Support Vector Machine(SVM).*

1. INTRODUCTION

Weather forecasting is a very important and useful area in the human day to day life. Forecasting weather for future is very necessary as human plans for the agriculture and many industries which are mostly dependent on weather conditions. For defence, shipping, aero navigations and mountaining purpose, we need to predict the weather conditions. Also to get prepared for forthcoming disasters and natural calamities, the abrupt change in climate condition needs to be forecasted.

The weather forecasting is done using the data collected from satellites. In this paper, the weather parameters like Average Temperature, Average Dew, Average Humidity, Average Air Pressure and Average wind speed of Cuttack city in Odisha state of India was taken to forecast the rain fall status of that place. We have drawn the decision tree based upon the J48 algorithm by using the numeric data for the weather parameters collected. Then the numeric data was converted to linguistic of textual form. Again the J48 Decision tree was drawn. A comparison is made based on the Root Mean Squared Error (RMSE).

This paper uses the Waikato Environment for Knowledge Analysis (WEKA) to design the decision tree for numeric and textual data after which the comparison concludes that the text mining for decision tree model design gives a faster and better way of predicting the weather. The Support Vector Machine was developed by taking the weather data of 365 days for training the model. Then the model was tested by a set of test data.

2. J48 DECISION TREE

Weka contains a data mining tool named J48 decision tree which is the implementation of C4.5 algorithm coded in Java. Ross Quinlan developed a decision tree by using C4.5 algorithm. This C4.5 is the extended version of ID3 algorithm. The decision tree generated by C4.5 is used for classification. It decides how the parameters behave to predict the nature of one dependent variable. In other words, the target variables are predicted by the algorithms designed by decision tree.

3. TEXT MINING

Text mining refers to text data mining which derives useful information from a database comprising of text data. The numeric data analysis which is generally used in data mining gives us useful information in data mining. But our proposed model is designed to use the database in textual

form so that in data mining the text data will be used for prediction of useful information by drawing the decision tree. After minute comparison between the data mining based on numeric data and text data, it is found that the decision tree built by using textual data gives quicker prediction than the decision tree built based upon the numeric data.

4. SUPPORT VECTOR MACHINE

Support Vector Machine is basically built upon the statistical learning theory. It is used to map the target data X into some feature space F with high dimension with a non linear mapping function to construct the non linear hyper plane [9]. In this article, the Support Vector Machine is uses the weather parameters like temperature, dew, humidity, air pressure and wind speed to predict the weather conditions like rain, no rain, Fog, Thunder Storm etc.

5. RELATED WORK

Hayati Mohsen and Mohebi Zahra have designed a model for predicting weather parameters before one day [1]. Zan Thet Chaw and Naing Thin Thu of University of Computer Studies have worked on Hidden Markov Models(HMMs) approach for rainfall forecasting out of the time series data during Myanmar Rainy Season [2]. Chattopadhyay Suarjit, Department of Mathematics, Techno Model School has predicted average summer – monsoon rainfall by proposing an artificial neural network model based weather prediction model [3]. Brian A. Smith, Ronald W.MClendo and Gerrit Hoogenboom have used Artificial Neural Network (ANN) for predicting air temperature by increasing the number of observations [4]. Oyediran O. F. and Adeyemo A. B. have made a comparison between Adaptive Neuro Fuzzy Inference System(ANFIS) and Multi Layered Perceptron(MLP) Artificial Nueral Network Models for analysing meterological data. The performance of the two models were evaluated and found that the ANFIS performed better than MLP ANN with lower error of prediction.

6. PROPOSED MODEL

We propose a model for prediction of rainfall at particular place i.e. Cuttack city of Odisha in India by taking the time series data of one year. The data of one year was used to train the model and then the data of one month, say January, was used to test the model. It is how the prediction of weather could be done. The data was collected from the website https://www.wunderground.com. The five weather parameters used for our model are Temperature, Dew, Humidity, Air Pressure and Wind Speed. The data present in the web site are numeric values. The J48 algorithm was used to draw the decision tree for predicting the rain fall. Then the data of the weather parameters are converted from numeric to linguistic form i.e. into text format [7]. Then the J48 algorithm was used to draw the decision tree for predicting the rain fall. The different statistical parameters were compared after taking the Decision tree output from the numeric data and textual data. It was found that the later was quicker and better in predicting the rain fall.

7. DATA PRE-PROCESSING

The data for the different weather parameters like temperature, dew, humidity, air pressure and wind speed collected are numeric in nature. With the context of our proposed model, the numeric data for the weather parameters are described in more natural and flexible summary. The numeric data is translated into flexible linguistic [8] data which is in the form of Natural Language. The entire range of present temperature is divided into five parts, namely, 'Very Low', 'Low', 'Medium', 'High' and 'Very High'.
The detail description of the different weather parameters for conversion from numeric to text are mentioned below.

Table 1: The data summarization during conversion of numeric data too text data

Temperature(°C)		Dew(°C)	
Numeric Range	Text Value	Numeric Range	Text Value
<=20	Very Low	<=15	Very Low
21-25	Low	16-20	Low
26-30	Medium	21-25	Medium
31- 35	High	25-30	High
>35	Very High	>30	Very High

Humidity(%)		Sea Level Pressure(hPa)	
Numeric Range	Text Value	Numeric Range	Text Value
<=60	Very Low	<=1000	Very Low
61-70	Low	1001-1005	Low
71-80	Medium	1006-1010	Medium
81-90	High	1011-1015	High
>90	Very High	>1015	Very High

Wind Flow(KM/H)	
Numeric Range	Text Value

<=5	Very Low
6-10	Low
11-15	Medium
16-20	High
>20	Very High

Table 2: Sample text data of 15 days after conversion from numeric data to text data

Date	temp_avg	dew_avg	humidity_avg	press_avg	wind_avg	Events
01-01-2015	Very Low	Low	Very High	High	High	Rain
02-01-2015	Low	Medium	High	High	Medium	Rain
03-01-2015	Low	Medium	High	High	Very Low	No Rain

8. ALGORITHM

The algorithm used to translate the numeric data to text equivalent to enable us for text mining of the weather forecasting is as follows:

if (temp_numeric<=20) set temp_text=very low
else if (temp_numeric>20 and temp_numeric<=25) set temp_text=low
else if (temp_numeric>25 and temp_numeric<=30) set temp_text=medium
else if (temp_numeric>30 and temp_numeric<=35) set temp_text=high
else set temp_text=very high
if (dew_numeric<=15) set dew_text=very low
else if (dew_numeric >15 and dew_numeric <=20) set dew_text =low
else if (dew_numeric >20 and dew_numeric <=25) set dew_text =medium
else if (dew_numeric >25 and dew_numeric <=30) set dew_text =high
else set dew_text =very high
if (humidity_numeric<=60) set humidity _text=very low
else if (humidity_numeric>60 and humidity_numeric<=70) set humidity_text=low
else if (humidity_numeric>70 and humidity_numeric<=80) set humidity_text=medium
else if (humidity_numeric>80 and humidity_numeric<=90) set humidity_text=high
else set humidity_text=very high
if (pressure_numeric<=1000) set pressure_text=very low
else if (pressure_numeric>1000 and pressure_numeric<=1005) set pressure_text=low
else if (pressure_numeric>1005 and pressure_numeric<=1010) set pressure_text=medium
else if (pressure_numeric>1010 and pressure_numeric<=1015) set pressure_text=high
else set pressure_text=very high
if (windflow_numeric<=5) set windflow_text=very low
else if (windflow_numeric >5 and windflow_numeric <=10) set windflow_text =low
else if (windflow_numeric >10 and windflow_numeric <=15) set windflow_text =medium
else if (windflow_numeric >15 and windflow_numeric <=20) set windflow_text =high
else set windflow_text =very high

The application of above mentioned algorithm converts the numeric data into text equivalent.

9. EXPERIMENTAL RESULTS

The J48 algorithms was used to draw the decision tree for predicting the rain fall based upon the weather parameters present in numeric form.
The decision tree generated by WEKA using J48 classifier for training set is shown below.

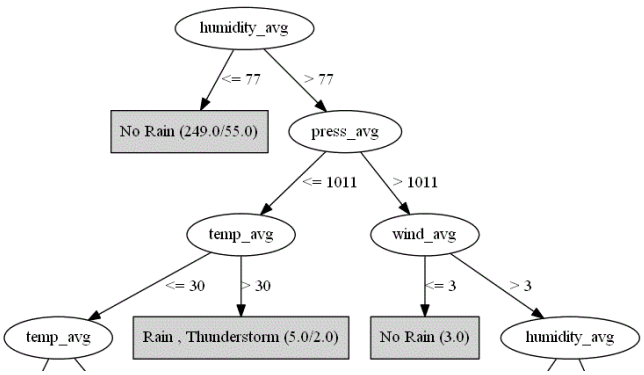

Fig. 1. J48 Decision Tree Generated With The Numeric Training Data Of 365 Days

The statistical parameters in the decision tree generated from numeric data for training is as follows.

Table 3: Statistical Parameters Of Numeric Data Used For Training The Model

Sl. No.	Parameters	Value
1	Number of leaves	21
2	Size of Tree	41
3	Time taken to build the model	0.08 sec
4	Correctly classified Instances	275(75.3425%)
5	Incorrectly Classified Instances	90 (24.6575 %)
6	Kappa statistic	0.503
7	Mean absolute error	0.1123
8	Root mean squared error	0.2369
9	Relative absolute error	67.5193%
10	Root relative squared error	82.5383%
11	Total Number of Instances	365

After the model was trained with the data of one year (365 days), the test data of one month (31 days) was fed to the model. After testing, the following parameters were found.

Table 4: Statistical Parameters Of Numeric Data Used For Testing The Model

Sl. No.	Parameters	Value
1	Number of leaves	21
2	Size of Tree	41
3	Time taken to build the model	0.05 sec
4	Correctly classified Instances	17(54.8387%)
5	Incorrectly Classified Instances	14 (45.1613 %)
6	Kappa statistic	0.503
7	Mean absolute error	0.1563
8	Root mean squared error	0.3192
9	Relative absolute error	85.4655%
10	Root relative squared error	101.4223%
11	Total Number of Instances	31

The decision tree generated by WEKA using J48 classifier for training set of text data is shown below.

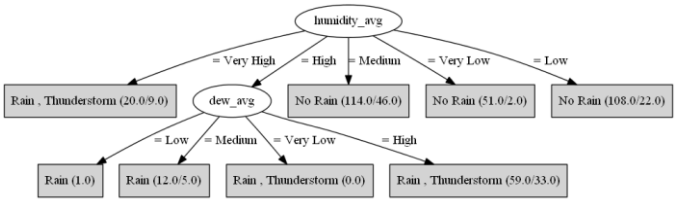

Fig. 2. J48 Decision Tree generated with the text training data of 365 days

The statistical parameters in the decision tree generated from text data for training is as follows.

Table 5: Statistical parameters of text data used for training the model

Sl. No.	Parameters	Value
1	Number of leaves	8
2	Size of Tree	10
3	Time taken to build the model	0.08 sec
4	Correctly classified Instances	248(67.9452%)
5	Incorrectly Classified Instances	117 (32.0548 %)
6	Kappa statistic	0.3614
7	Mean absolute error	0.1326
8	Root mean squared error	0.2575
9	Relative absolute error	79.7444%
10	Root relative squared error	89.6998%
11	Total Number of Instances	365

After the model was trained with the text data of one year (365 days), the test data of one month (31 days) was fed to the model. After testing, the following parameters were found.

Table 6: Statistical Parameters of Text Data Used For Testing The Model

Sl. No.	Parameters	Value
1	Number of leaves	8
2	Size of Tree	10
3	Time taken to build the model	0.05 sec
4	Correctly classified Instances	17(54.8387%)
5	Incorrectly Classified Instances	14(45.1613 %)
6	Kappa statistic	0.503
7	Mean absolute error	0.1516
8	Root mean squared error	0.2999

9	Relative absolute error	82.8963%
10	Root relative squared error	95.2976%
11	Total Number of Instances	31

The numeric data collected with the weather parameters like temperature, dew, humidity, air pressure and wind speed are divided into 2 sets. One is training set another test set. The training set contains data of one year i.e. 365 days. The training set of data is fed to the model using J48 decision tree and Support Vector machine (SVM).
The statistical parameters in SVM generated from numeric data for training is as follows.

Table 7: Statistical Parameters Of Numeric Data Used For Training The SVM Model

Sl. No.	Parameters	Value
4	Correctly classified Instances	251(68.7671%)
5	Incorrectly Classified Instances	114 (31.2329 %)
6	Kappa statistic	0.3288
7	Mean absolute error	0.2119
8	Root mean squared error	0.3139
9	Relative absolute error	127.4627%
10	Root relative squared error	109.3591%
11	Total Number of Instances	365

After the SVM model was trained with the data of one year (365 days), the test data of one month (31 days) was fed to the model. After testing, the following parameters were found.

Table 8: Statistical parameters of numeric data used for testing the SVM model

Sl. No.	Parameters	Value
4	Correctly classified Instances	17(54.8387%)
5	Incorrectly Classified Instances	14 (45.1613 %)
6	Kappa statistic	0.3288
7	Mean absolute error	0.2115
8	Root mean squared error	0.3133
9	Relative absolute error	115.6636%
10	Root relative squared error	99.5624%
11	Total Number of Instances	31

10. ANALYSIS

After careful study of both the models, it was experimentally found that the decision tree generated by taking the text data is smaller in size than that generated using the numeric data. In case of text data, the size of the tree is 10 where as in case of numeric data, it is 41 for doing the same prediction of rain. Our ultimate objective is to predict the status of the rain such as 'Rain', 'No Rain', 'Fog', 'Thunderstorm', 'Thunderstorm, Tornado', 'Rain, Thunderstorm' or 'Fog, Thunderstorm'. Also, the number of leaves where the decision of the aforesaid events of rain is decided is 21 in case of numeric data whereas it is only 8 in case of text data. So, it is observed by taking many such sample data and test data that the decision obtained by taking the text data is faster and better than the numeric data. Table IX shows the parameter comparison between Predictions using Numeric and Text Data.

Table 9: Parameter comparison between Predictions using Numeric and text Data

Parameter	Prediction using Numeric Data	Prediction using Text Data
Number of leaves	21	8
Size of Tree	41	10
Time taken to build the model	0.08 sec	0.08 sec
Correctly classified Instances	275(75.3425%)	248(67.9452%)
Incorrectly Classified Instances	90 (24.6575 %)	117(32.0548 %)
Kappa statistic	0.503	0.3614
Mean absolute error	0.1123	0.1326
Root mean squared error	0.2369	0.2575
Relative absolute error	67.5193%	79.7444%
Root relative squared error	82.5383%	89.6998%
Total Number of Instances	365	365

If we compare between the parameters obtained by using J48 decision tree and SVM while testing, it gives the following analytical outcome.

Table 10: Parameter comparison between Predictions using J48 Decision Tree and SVM

Sl. No.	Parameters	J48 Decision Tree	SVM
4	Correctly classified Instances	17(54.8387%)	17(54.8387%)
5	Incorrectly Classified Instances	14 (45.1613%)	14 (45.1613%)
6	Kappa statistic	0.503	0.3288
7	Mean absolute error(MAE)	0.1563	0.2115
8	Root mean squared error(RMSE)	0.3192	0.3133
9	Relative absolute error(RAE)	85.4655%	115.6636%
10	Root relative squared error(RRSE)	101.4223%	99.5624%
11	Total Number of Instances	31	31

By analyzing table X, we see that J48 decision tree gives less mean absolute error and relative absolute error than SVM. Thereby, we can conclude that the J48 decision tree gives better prediction of weather than the prediction using Support Vector Machine.

11. COMPARISON OF MODELS BASED ON THRESHOLD CURVE

The models based on the numeric data of 365 days were analysed based upon the different class values like 'Rain', 'No Rain', 'Fog', 'Thunderstorm', 'Thunderstorm, Tornado', 'Rain, Thunderstorm' or 'Fog, Thunderstorm' etc. The threshold curve generated for Support Vector Machine and J48 decision tree were shown below.

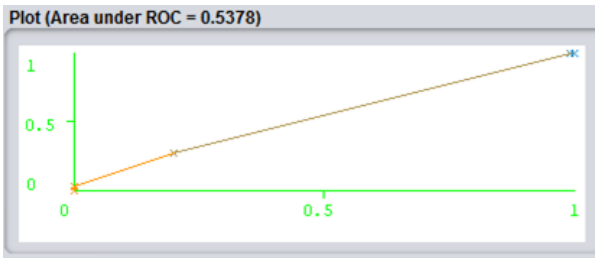

Fig. 3. Threshold Curve in SVM for class value Rain

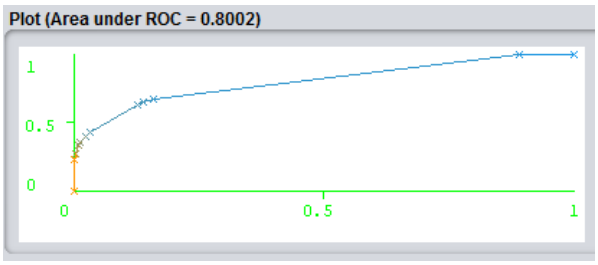

Fig. 4. Threshold Curve in J48 for class value Rain

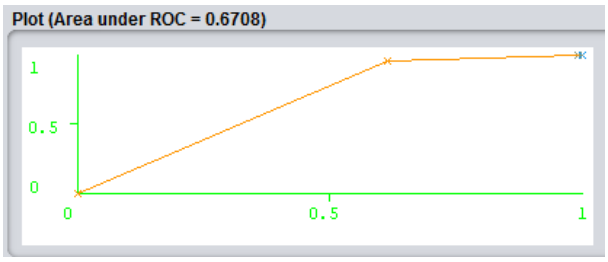

Fig. 5. Threshold Curve in SVM for class value No Rain

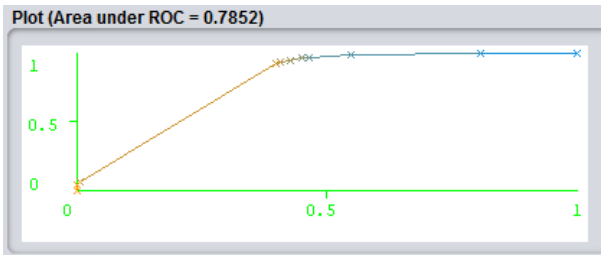

Fig. 6. Threshold Curve in J48 for class value No Rain

The comparison of threshold curves between SVM and J48 signifies that the J48 classifier gives better result in predicting weather events based upon the parameters taken.

12. CONCLUSION AND PROPOSED FUTURE WORK

It has been concluded that the text mining in predicting the rain fall has given promising results. It is found that the text mining gives better and quicker prediction in rain fall forecasting. This study clearly gives a picture that the text mining used in generating the J48 decision tree in weather forecasting gives result accurately as needed. When we use the support vector machine to predict the weather events and was compared with the statistical parameters obtained from predicting the same using J48 decision tree, it was remarked that later predicts better. As text mining is better than numerical data mining in case of J48 decision tree modelling, it is proposed to use text mining in J48 decision tree to obtain better and quicker weather prediction. The result obtained in the study emphasizes that text mining

and further application of NLP can be used for further weather forecasting experiments.

REFERENCES

[1] Hayati M. And Mohebi Z.(2007), "Temperature Forecasting Based on Neural Network Approach", World Applied Sciences Journal 2 (6):613-620.ISSN 1818-4952.

[2] Zan C. And Naing T. (2009), "Myanmar Rainfall forecasting using hidden Markov Model", IEEE International Advance Computing Conference.

[3] Chattopadhyay S. (2006), " Multilayered feed forward Artificial Neural Network model to predict the average summer – monsoon rain fall in India", IEEE, Volume : 11, pp.:125-130.

[4] Brian A. Smith, Ronald W.MClendon, Gerrit Hoogenboom, "Improving Air Temperature Prediction with Artificial Neural Networks", International Journal of Computational Intelligence , Vol.10, No.3, March, 2007.

[5] Oyediran O. F., Adeyemo A. B., "Performance Evaluation of Neural Network MLP and ANFIS models for Weather Forecasting Studies", African Journal of Computing & ICT, IEEE, Vol. 6, No. 1, March 2013.

[6] Dr. S. Santhosh Baboo, I. Kadar Shreef, "An efficient weather forecasting system using Artificial Neural Network", International Journal of Environmental Science & Development, W1.1, No.4, October – 2010, ISSN: 2010-0264.

[7] van der Heide A., Trivino G.(2009), "Automatic generated linguistic summaries of energy consumption data". In proceedings of 9th ISDA Conference, pp. 553-559.

[8] Eciolaza L., Pereira Farina M., Trivino G.(2012), "Automatic linguistic reporting in driving simulation environments", Applied Soft Computing.

[9] Radhika Y., Shashi M., "Atmosphereic Temperature Prediction using Support Vector machines", International Journal of Computer Theory and Engineering, Vol – 1, No -1., April, 2009.