ISSN: 2321-7782 (Online)

Special Issue, December 2013

International Journal of Advance Research in Computer Science and Management Studies

Research Paper

Available online at: www.ijarcsms.com

Application of Data Mining Techniques in Maximum Temperature Forecasting: A Comprehensive Literature Review

R. Nagalakshmi[1]
M.E.
Department of Computer Science and Engineering
Velammal Engineering College
Chennai – India

M. Usha[2]
Assistant Professor
Department of Computer Science and Engineering
Velammal Engineering College
Chennai – India

RM. A. N. Ramanathan[3]
Asst. Meteorologist Gr.II
Regional Meteorological Center
Chennai – India

***Abstract:** There are several applications for Machine Learning (ML), the most significant of which is data mining. It is used to discover the hidden patterns from large volume of raw data. Data mining can be classified into supervised learning (classification and prediction) and unsupervised learning (clustering, association rules etc.,). The ultimate goal of data mining is prediction and predictive data mining is the most common type of data mining. The chaotic nature of the atmosphere, Weather prediction is a difficult process and a demanding task for researchers. The prediction of atmospheric parameter is becoming more important to protect life and property, climate monitoring, drought detection, severe weather prediction, agriculture and production, planning in energy industry, aviation industry communication, pollution, economy and business decisions etc.. In this paper, a brief survey about various data mining techniques used for weather prediction and their accuracy is discussed. Also the various potential predictors used for obtaining better prediction accuracy are discussed.*

***Keywords:** Data Mining Techniques, Maximum Temperature, Forecasting.*

I. Introduction

Weather and climate affect human society in all dimensions. In agriculture, diurnal variation is important for crop protection. In water management, rain is the most important factor. Energy sources, like natural gas and electricity are greatly depends on weather conditions. Further maximum temperature likely to be realised on a day is an important weather parameter in aviation as high temperature affect the load factor of aircrafts. Weather forecasting entails predicting how the present state of the atmosphere will change. Creating forecasts is a complex process which is constantly being updated. Weather forecasts made for two and three days are usually good. Weather forecast made for 12 and 24 hours are typically quite accurate. The approaches used in modern weather forecasting are traditional synoptic weather forecasting, numerical weather prediction, statistical methods, and various short-range forecasting techniques. One statistical approach, the analog method, examines past weather records to find ones that come close to duplicating current conditions.

This approach is useful for predicting local weather when recorded cases are plentiful. Statistical models were successful in normal temperature and failed remarkably during the extreme temperature. Main drawbacks of statistical models are. Statistical models are not useful to study the highly nonlinear relationships between temperature and its predictors. So there is no ultimate end in finding the best predictors. Now casting, uses radar which cover 450km and mainly used for know the cyclone movements and geostationary satellites, ground level observations, radiosonde balloons monitor upper-air data over a particular location, windsock, weathervane, anemometer, thermometer Which are quickly forecast severe weather events, such as thunderstorms, tornadoes, hail storms. Maximum daily temperature is observed when the amount of incoming shortwave radiation equals to the amount of outgoing long wave during the mid to late afternoon at about 1400hr local time. Thus maximum temperature corresponding to, say 15 May is impact realized on the afternoon of 14 May and is recorded at 1730 IST. So it is appropriate to develop a 24 hour forecast scheme to forecast at 1730 the maximum temperature of the next day. The forecast lead time will only be 21 hour, as the maximum temperature is generally reached at about 1400. The success of a forecasting scheme depends to a large extent on the correct choice of predictors besides the application of suitable methodology [07]. As maximum temperature is a parameter observed daily, which is influenced by the observations of previous days so it becomes as possible predictors. The following are considered as predictors For 24 hr forecast.

- Dew point temperature.
- Outgoing long wave radiation is considered as possible predictor, because it balances the heat of atmosphere during the night.
- Precipitation in the form of hail, storm, snow, rain drops (drizzle below .02 inch).
- Minimum temperature recorded prior to the time of issue of forecast also is likely to be related to the maximum temperature.
- Advection of temperature is an important mechanism in the variation of temperature.

Temperature advection refers to the change in temperature caused by movement of air by the wind. Forecasting temperatures using advection involves looking at the wind direction at your forecasting site and the temperatures upstream (in the direction from which the wind is blowing). If they are warmer, that means warmer air is being transported towards your station and the temperature should rise. Put in another way, if there is warm advection occurring at a given station, expect the temperatures to increase. In contrast, if cold advection is occurring at a given station, expect the temperatures to drop. The negative sign indicates that this is cold temperature advection, moving colder air into a region of warmer air. To compute temperature advection, uwind(east to west), vwind(north to south) and temperature upstream dataset. Many researchers have tried to use data mining technologies in areas of meteorology and weather, climate prediction. The most commonly used techniques in data mining are: Artificial neural networks, genetic algorithms, Fuzzy logic, Rule induction, Nearest Neighbour method, Memory–Based Reasoning, Logistic Regression Discriminate Analysis and Decision Trees, support vector machine, ensemble neural network. This work provides a brief overview of data mining techniques applied to weather prediction.

The remainder of the paper is organized as follows. In Section 2 the related work for solving maximum temperature prediction using data mining techniques is presented. Section 3 concludes the paper with fewer discussions.

II. Related Work

2.1 Mohsen Hayati and Zahra Mohebi (2007)

Mohsen & Zahra [04] make use of Multi Layer Perceptron(MLP) for one day ahead prediction of temperature of Kermanshah city located in west of Iran. They have used ten years (1996-2006) meteorological data. For accuracy of prediction they split data into four seasons and then for each seasons one MLP network was constructed. The development period consists of 65 % of data and testing period consists of 35 % of data. Two random days in each season are selected as unseen data which have not been used in training. Mean Absolute Error (MAE) is used to measure the performance. Tan-sig is used as activation function at each hidden layers & pure-linear function is used at the output layer. The optimum structure of each season is listed out in Table 1[04].

Table 1 Optimum structure of each season [04]

MLP	Spring	Summ er	Fall	Winter
Number of Hidden Neurons	4	4	6	4
Number of Iterations	2000	2000	2000	2000
Activation function for hidden layer	tan-sig	tan-sig	tan-sig	tan-sig
Activation function for Output layer	Pure linear	Pure linear	Pure linear	Pure linear
Minimum Error	0.001	0.0148	0.0336	0.0019
Maximum Error	0.3569	1.6417	0.7896	0.5679

Mean Absolute Error vary between 0 and 1.7

2.2 PAL et al. (2002)

A neural network based forecasting model for maximum and minimum temperature for the DUM DUM station in west Bengal was proposed by PAL et al. [05]. They have used 2285 records as a training data (1989 - 1995) and 270 records for test the model. In the training data first 25 fields of a record act as on input data set for input node of the network and the remaining 2 fields of the record as on target data set for the output node in the output layer of the network. It takes previous two consecutive days information as input for predicting the maximum and minimum temperature. The record structure contains the following field.[05]

(I – 2) $_{th}$ day information collected

1. Mean sea level pressure at 17.30 hr
2. Mean sea level pressure at 08.30 hr
3. Vapor pressure at 17.30 hr
4. Vapor pressure at 08.30 hr
5. Relative humidity at 17.30 hr
6. Relative humidity at 08.30 hr
7. Maximum temperature at 17.30 hr
8. Minimum temperature at 08.30 hr
9. Rainfall
10. Direct radiation
11. Diffuse radiation

(I – 1) $_{th}$ day information collected

12. Mean sea level pressure at 17.30 hr
13. Mean sea level pressure at 08.30 hr
14. Vapor pressure at 17.30 hr
15. Vapor pressure at 08.30 hr
16. Relative humidity at 17.30 hr
17. Relative humidity at 08.30 hr

18. Maximum temperature at 17.30 hr
19. Minimum temperature at 08.30 hr
20. Rainfall
21. Direct radiation
22. Diffuse radiation

I_{th} day information collected

23. Day
24. Month
25. Year
26. Maximum temperature at 17.30 hr
27. Minimum temperature at 08.30 hr

They used the gradient decent learning algorithm and found that 20 or less neurons in the hidden layer is optimum and it gave an error rate within ± 2^o C for 80 % of test cases.

2.3 Srinivasan and Hashim (1965)

Srinivasan et al. [06] studied the persistence of maximum temperature tendencies during April, May and June for the period of 1945 – 1964 (20 years) daily maximum temperature record of Madras. The rate of spell of days having the same type of maximum temperature tendency $s_x = Kp^x$. s_x representing the spells of x days duration. Kp^x provides the probability of occurrence of any length of spell of raise or fall day temperatures. The following parameters need to be considered to prepare the maximum temperature prediction[06]

1. Wind components at 00 GMT of date at Madras at 3000 ft along the direction Anantapur – Madras.
2. Wind components at 00 GMT of date at 3000 ft at Anantapur along the same direction.
3. The previous day maximum temperature of Madras.

$T_x = f(T_{x-1}, WA, WM)$[06]

T_{x-1} = Refers to the previous day maximum temperature

W is the wind parameter at 00 GMT of date at the two stations.

The test data has taken for the month of May for years 1955, 1956, 1957 and 1965. So out of 115 days considered for testing the result, 78 % of the days had the maximum temperature between 34^oC to 40^oC. When the temperature ranged between 34 to 36 ^oC, 36 to 38^oC and 38 to 40^oC the forecast were correct 62, 68, 77 % respectively. But the temperature was less than 34^oC the results are not satisfactory. On a study of the test data it was seen that on rainy days the temperature are very much below the normal May values. So author concludes, consider precipitation parameter to avoid larger magnitude of errors when actual temperature is less than 35^oC when predicting maximum temperature.

2.4 RAJ (1998)

Forecasting schemes based on statistical technique has developed to predict daily summer March – May maximum temperature of Madras by Raj [07]. A set of optimal number of predictors were chosen from a large number of parameters by employing step wise forward screening method. Separate forecasting schemes for Madras city and Air port with the lead time of 24 and 9 hr were developed from the date of 12 years (1971 – 1982) and tested with 4 years (1983 – 1985 and 1995). Wind at upper level was not considered to predict with lead time of 24hr. but essential to consider for with lead time of 9hr prediction. Finally Maximum temperature of the previous day, normal daily maximum temperature, temperature advection index and morning zonal wind at Madras at 900 hPa level were among the predictors selected. The schemes yielded good results providing 77 – 87 % correct forecasts with skill scores of 0.29 – 0.57. Possible Predictors for maximum temperature of Madras and Madras Airport[07]

1. Maximum Temperature (I - 2) D
2. Maximum Temperature (I - 1) D
3. Daily normal Temperature
4. Minimum Temperature (I – 2) D
5. Temperature 17:30 (I - 2) D
6. Dew Point Temperature (I - 2) D
7. Cloud Index 17:30 (I - 2) D
8. Maximum Temperature Advection Index
9. Minimum Temperature (I – 1) D
10. Zonal Wind – 0.9 km 05:30 (I - 1) D
11. Temperature– 0.85 km 05:30 (I - 1) D

The possible predictor 1 – 8 is available for 24 h forecast and 9 – 11 are additional potential predictor for 9 h forecast. Finally for both Madras and Madras Airport for the 24 hr scheme prediction, maximum temperature of (I - 1) D, the temperature of the previous day emerging as the leading predictor. The multiple Correlation Coefficient obtained was 0.9 explaining nearly 80.8 % of the total variation. For 9 hr scheme Zonal Wind emerged as an important predictor. Table 2 shows the statistics of summer maximum temperature of Madras and Madras Airport.

Local rate of temperature $\frac{\partial T}{\partial t}$ is given by[07]

$$\frac{\partial T}{\partial t} = \frac{1}{C_p}\frac{\partial H}{\partial t} - \left(u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y}\right) - (\gamma_d - \gamma)$$

Non adiabatic heating and cooling rate is $\frac{\partial H}{\partial t}$

Table 2 Statistics of summer maximum temperature (^{o}C) of Madras and Madras AP based on data of 1971 – 1982[07]

Place	**Season**	**Mean**	**SD**	**Range**
Madras	March	32.1	1.3	29.2 – 37.5
	April	34.2	1.5	31.5 – 39.6
	May	36.9	2.8	29.8 – 43.6
	March - May	34.4	2.8	29.2 – 43.6
Madras A.P	March	33.1	1.6	29.9 – 38.6
	April	35.7	1.8	31.8 – 39.5
	May	38.0	2.6	29.7 – 44.3
	March - May	35.6	2.9	29.7 – 44.3

As seen from the Table 2 the mean increases for both the stations with the March of the season. The temperature of Madras AP is more than that of Madras, the difference which was only 1^{o}C in March increases to 2.1^{o}C in May. For both the stations more variation occurred in May. An average of 5.4 days per season experienced maximum temperature of above 40^{o}C in Madras and 9.2 days for Madras Airport.

2.5 DE (2009)

De [08] included three layered feed forward neural network to forecast average summer monsoon temperature over India (1901 – 2003). In this 75% data act as a training data and 25% of data take as a test data. The temperature of June, July & August has predicted with the help of January to May temperature. Maximum and minimum temperature is greatly predicted in the month of August with prediction error was below 5%. The learning rate γ (neu) was 0.9 for three models are generated for both Max. & Min. Temperature prediction. Initial weights were from -0.5 to 0.5. After 500 epochs the ANN has been found it produced a forecast with small prediction error.

Error: - For the month June the predicted Max. Temp & Absolute prediction

Error [08]

- June- Actual temperature from 32 to 37^{o}C. Predicted temperature varies - 34 to 35^{o}C. Predicted Errors -Up to 0- 3%.
- July- Actual temperature Varies-from 29.5 to 33.5^{o}C Predicted temperature varies - From 31 to 32^{o}C Predicted Errors- Up to 0- 3%
- August- Actual temperature Varies-from 29.5 to 31.5^{o}C Predicted temperature varies From 29.5 to 31.5oC Predicted Errors- Up to 0-2%

The study establishes that the third model is the best predictive model compared than other 2 models.

2.6 IMRAN MAQSOOD et al. (2004)

This study presented applicability of an ensemble of ANN and learning paradigms for weather prediction in southern Saskatchewan, Canada. Imran Maqsood et al. [09] proposed ensemble model performance is differences with multi layer perceptron network, Elman Recurrent Neural Network (ERNN), Radial Basis Function Network (RBFN), Hopfield Model (HFM) Predictive Models and Regression Technique. The data of temperature, wind speed and relative humidity are used to train and test the different model with each model; 24 hr ahead forecast are made for winter, spring, summer and fall season. Moreover the performance and reliability of the seven models are then assessed by a number of statistical measures Training data set of December 1, 2000 – February 25, 2001; March 1 – May 5, 2001; January 1 – August 6, 2001 and September 1 – November 9, 2001 were used for winter, summer, spring and fall season respectively. The hourly data set of February 26, May 6, August 7 and November 10 were selected to test the trained model. All of the weather data set normalized into values between -1 and +1 by dividing the difference between actual and minimum value by the difference between maximum and minimum value. Figure 1 shows the architecture of ensemble method.

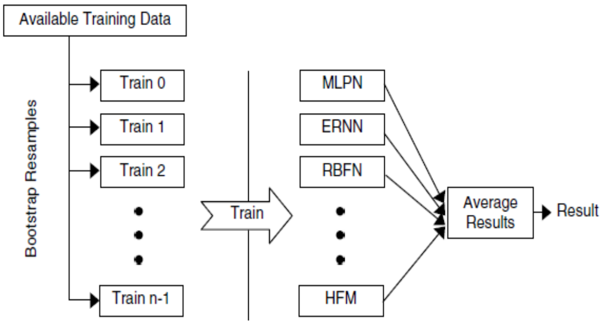

Figure 1. Architecture of Ensemble method [09]

One step secant learning algorithm used for training the MLPN & ERNN network. The activation function log sigmoid and hyperbolic tangent sigmoid for hidden units and pure line for output units. One Hidden layer with 72 Neurons selected and No significant reduction in error beyond 45 Neurons. Training time 5 minutes to 30 minutes for both MLPN & ERNN. RBFN implemented with two hidden layer of 180 Neurons selected and Gaussian activation function chosen for hidden units. Training time for HFM and RBFN only few seconds. Table 3 shows the performance of MLPN, ERNN, RBFN and HFM for forecasting temperature. Performance comparison of WTA ensemble, WA ensemble and Statistical method for temperature forecasting listed out in Table 4. Ensemble methods are implemented by weighted average algorithm and winner take all approach. Winner take all approach had more accuracy than the above all.

Table 3 Performance of MLPN, ERNN, RBFN, and HFM for forecasting temperature[09]

Model	Temperature		
	MAD	RMSE	CC
WTA	0.1127	0.0049	0.9998
WA	0.6558	0.0098	0.9974
Statistical Method	5.3364	4.3568	0.0895

From the Table 3 empirical results indicates that HFM is relatively less accurate and RBFN is relatively more reliable for temperature forecasting.

Table 4 Performance comparison of WTA ensemble, WA ensemble and Statistical method for temperature forecasting [09]

Sea-son	Reliability	Temperature			
		MLPN	ERNN	RBFN	HFM
Summer	MAD	0.5259	0.4699	0.4641	2.3151
	RMSE	0.0199	0.0199	0.0050	0.0650
	CC	0.9937	0.9976	0.9996	0.9897
	Training Time	1020	1260	3	2

From the table 4 WTA approach of the ensemble of neural networks produced the most accurate forecast.

Equation used for predicting the temperature [09].

$$T_s = -0.000003t^2 + 0.0089t + 13.386$$

2.7 Sarah N.Kohail et al. (2011)

In this paper Sarah N.Kohail et al. [10] try to extract useful knowledge from weather historical data collected locally at Gaza strip city. The data include 9 year period (1977 – 1985). 70% of data considered as training data and 30% used for test. Two prediction methods used to predict the daily average temperature of Gaza city. Table 5[10] shows the prediction results of two prediction methods applied for Gaza city weather data.

Table 5 Prediction results of two prediction methods applied for Gaza city weather data[10]

Method	Correlation coefficient	RMSE
Least Median Square Linear Regression	0.924	1.691
Neural Networks Learning rate: 0.3 Momentum: 0.2	0.933	1.726

The first method is Artificial Neural Network with 8 input layer, 6 hidden layers and 1 output layer. The second method is least median squares linear regression used. ANN provides better correlation coefficient between actual and predicted temperature, and lower Root Mean Square Error (RMSE). Therefore ANN is better than least median squares linear regression.

2.8 Dilruba Sharmin et al.(2011)

In this paper, Temperature Predicting Neural Network (TPNN) has been developed by Dilruba Sharmin et al. [11] by using feed forward back propagation multilayer ANN in Visual C++. The model was trained and tested using nine years (1992-2000) meteorological data. Inputs of neural network were daily maximum temperature, minimum temperature, average temperature, rainfall, humidity, sunshine hours and wind speed of previous day and the output was the max or min temperature of the day.

The data for the year 1992-1999 were used in training phase while that for the year 2000 were used to test the model. The optimum structure of MLP is listed out in Table 6.[11]

Table 6 Optimum structure for MLP[11]

Parameter	Maximum Temperature	Minimum Temperature
Number of neurons	3	10
Learning rate	0.3	0.7
Momentum	0.5	0.2
Number of Iterations	1000	1000
Activation Function	Sigmoid	Sigmoid

The accuracy of the model was calculated and means relative percentage error for the TPNN model was 7.45826% for maximum temperature and 8.655804% for minimum temperature prediction.

2.9 Y.Radhika and M.shashi(2009)

This paper presents an application of support vector Machines for weather prediction by Radhika et al.[12]. The weather data of University of Cambridge for a period of five years (2003-2007) were used to build the models and data between January and July of year 2008 is used to test the model. The results are compared with Multi layer perceptron trained with back-propagation algorithm and the SVM performs better than MLP. The Mean Square Error in the case of MLP varies from 8.07 to 10.2 based on the order whereas it is in the range of 7.07 to 7.56 in case of SVM. Radial basis function (RBF) kernel function used in SVM Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

III. Conclusion

In this study it is found that Radial Basis Function Network (RBFN) is performed well than MLPN, ERNN, HFM. Winner Take All Algorithm of Ensemble approach gave best results which take input from MLPN, ERNN, HFM, and RBFN. Support vector Machine gave very best accuracy than Multilayered Perceptron of Artificial Neural Network. So from the above study, the success of a forecasting scheme depends to a large extent on selecting the correct choice of predictors than the application of suitable methodology understood [6].By finding mean, standard deviation, Range, median (number of days the same temperature prevailed) on the dataset used will provide more knowledge which in turn can be used for analysis on the parameters needed for improving the prediction accuracy. So to further improve the performance of prediction is possible by applying statistical based feature selection techniques, and by eliminating extreme condition of day(hottest day got rain etc.,) through outlier analysis which reduces the error rate in accuracy. After going through all of above study & discussion we see that applying support vector machine for forecasting maximum temperature is most feasible rather than other predicting techniques which discussed above.

References

1. A.K. Banerji and A.B. Chowdhury, “Forecasting summer maximum temperature at Nagpur,” Indian J.Met.Geophys., vol.23, pp. 251, 1972.
2. V.K.Raghavendra,“Forecasting maximum temperature at Poona”, Indian J.Met.Geophys.,vol.7,pp.17,1956.
3. Meghali A.Kalyankar,S.J.Alaspurkar, “Data Mining Techniques to Anlayse the Metrological Data,” International Journal of Advanced Research in computer science and Software Engineering ,vol.3,no. 2, February 2013.

4. Mohsen Hayati, Zahra Mohebi, “Temperature Forecasting Based on Neural Network Approach,” World Applied Sciences Journal, vol. 2, no. 6, pp. 613 – 620, 2007.

5. S. Pal, J. Das, P. Sengupta and S.K. Banerjee, “Short term prediction of atmospheric temperature using neural networks,” Mausam, vol. 53, no. 4, pp. 471 – 480, October 2002.

6. T.R. Srinivasan and S.S. Hashim, “A Statistical study of daily maximum temperatures at Madras during the summer months,” Mausam, pp. 76 – 78, 1965.

7. Y.E.A. Raj, “On forecasting daily summer maximum temperature at Madras,” Mausam, vol. 49, no. 1, pp. 95 – 102, 1998.

8. S.S. De, “Artificial Neural Network Based Prediction of Maximum and Minimum Temperature,” Applied Physics Research, vol. 1, no. 2, pp. 37 – 43, 2009.

9. Imran Maqsood, Muhammad Riaz Khan and Ajith Abraham, “An ensemble of neural networks for weather forecasting,” Neural Comput & Applic, vol. 13, pp. 112 – 122, 2004.

10. Sarah N.Kohail, Alaa M.El- Halees, “Implementation of Data Mining Techniques for Meteorological Data Analysis (A case study for Gaza Strip),” International Journal of Information and communication Technology Research, vol. 1, no. 3, pp. 96 – 100, 2011.

11. Dilruba Sharmin,Farzana Hussain,Md.Shafiqur Rahman,Susmita Ghose,T.K.Yousufzai and Mahfuja Akter, “A Study on the parameters of back propagation artificial neural network temperature prediction model,” International Journal of Advanced Engineering Sciences and Technologies, vol. 2, no.1, pp. 099-103, 2011.

12. Y.Radhika ,M.Shashi,”Atmospheric Temperature Prediction using Support Vector Machines,” International Journal of Computer Theory and Engineering,vol.1,no.1,2009

13. Ilhami Colak a.*, Seref Sagiroglu b, Mehmet Demirtas a, Mehmet Yesilbudak c,” A data mining approach:Analyzing wind speed and insolation period data in Turkey for installation of wind and solar power plants,” Energy Conversion and Managemet,vol.65,pp.185-197,2013

© 2013, IJARCSMS All Rights Reserved