Middle-East Journal of Scientific Research 23 (1): 35-40, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.01.21747

Hourly Based Climate Prediction Using Data Mining Techniquesby Comprising Entity Demean Algorithm

Zaheer Ullah Khan, Ashfaq Ahmad and Maqsood Hayat

Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan

Abstract: Forecasting is to see ahead the occurrence of the underline actual event generation. Forecasting is a science and technology application in order to predict the atmospheric condition. In the last few decades, Weather Forecasting is one of the most crucial, high valued most significantly and technologically challenging problem of the world. We often see and hear in the News that a great Mass and Property is being damaged and destroyed due to the abnormal occurrence of the weather uneaten change. Most of the works have been carried out on weather forecasting for different period of prediction pattern. Some models carry predictions of weather forecasting from real time to Annual period. In this paper, we develop a system that carries out climate prediction using previous state of weather having attribute (Date, Time,Hour,Temperature,Dew Point (DP),Relative Humidity, Wind Speed (WS), Wind Direction (WD), Wind Chill (WC) and StnPressure(SP) etc.Various data mining techniques are employed for prediction of weatherforecasting including K-Nearest Neighbor, Decision Trees and naive Bayes. Decision Tree has achieved quite promising performance among using algorithms.

Key words: K-Nearest Neighbor (KNN) · Decision Tree · Ensemble Decision Tree · Naïve Bayes

INTRODUCTION

Hourly based Weatherprediction is the world recent developmentresearch trend of metrology. Hourly Weather forecasting is one of the challenging problems due to its chaotic nature.Weather and climate affect human society in so many aspects [1-18], it also entails making destruction and damage to the mass and property. Weather forecast made for 12 to 24 hours is quite accurate [7] in contrast; making forecasting hourly basis is a challenging task due to the freneticand unpredictable nature of the weather [4].

The Data mining so called Knowledge Discovery in Databases (KDD) is the field of discovering inferential and potentially useful information from large amounts of data [1]. The statistical model or mathematical models both are static in nature. On the other hand, Data Mining can extract interesting and useful informative pattern from hidden pattern without prior hypothesis. This also depends upon the technique employed and the underline data used.Some of the well-known techniques used in Data Mining are K-Nearest Neighbors (KNN) and Decision Tree, Artificial Neural Network (ANN), Genetic Algorithm, Rule Inductionetc.KNN is awell-known clusteringalgorithm that is based on Euclidian Distance computation [17], it ismost widely used in scientific and industrial research and application. The Decision Tree [DT] is a hierarchical model for supervised learning, where each internal node, branch, leaf node represents a test on an attribute, an outcome of the test and class label respectively [19].

Many researchers and Organizations have developed different tools for weather prediction on scale from Real Time [5] to annual climate prediction [18]. Most of the weather prediction models were statistical [8]. With the advancement of technology and computation speed most of the organizations and institutions now using data mining technique for weather prediction e.g. Artificial Neural Network [16], Genetic Algorithm [13], KNN [18], DT [19]. In this paper, we applied both the Data Mining algorithm i.e. KNN and DT over weather dataset (of type Numeric) and a performance metric i.e.

Corresponding Author: Zaheer Ullah Khan, Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan.

MATERIAL AND METHODS

Datasets: Data Mining is a popular technology in the field of Data Warehousing and Very Large Data Bases (VLDB)[18]. Data mining is a collection of employed processes that analyze data from a diverse perspective and represents it into useful information. In order to obtained unbiased and high quality dataset, it is passed from various procedures.

There are many organizations such as Government agencies and some educationalinstitutions thatprovide access to weather data. NCDC (National Climate Data Center) andCanadian climatological datacenter provide a large weather data range from surface to Radar and Satellite imagery. Relevant data are extracted. The selected attribute of the meteorological datasetincluding description of attributes (extracted features) along with their type and brief descriptionare shown in Table 1.

Data Acquisition and Preprocessing: The weather dataset has been collected from the Canadian climatological web site that gathers weather information over a vast geographical location for about all the states. In this study, we have used the weather / Climate of Alberta. The hourly weather data has been collected over two year 2012 and 2013* current.

The data selected werehaving outlier and missing values and noisy. A couple of statistical and Data mining techniques (Mean, Bin Mean) were applied for cleaning the data.Standard pre-processing procedures are adopted to bring the data in consistent and appropriate format for mining purposes.Table 2 shows the detail calculated statistics of the data.

Proposed Model: In this paper, we have proposed a new framework for the development of weather forecasting system, using DT. In our proposed model the data is preprocessed in order to convert data into consistent and proper format for data mining purpose.

$$\text{Humidex} = (0.5555) \times \left(\left[6.11 \times 10^{[5417.7530 \times ((\frac{1}{273.16}) - (1/\text{Dewpoint})]}\right] - 10.0\right) \tag{1}$$

$$\text{Wind Chill} = 13.12 + 0.6215 * T_{air} - 11.37 * {v_{10m}}^{0.16} + 0.3965 * T_{air} * {v_{10m}}^{0.16} \tag{2}$$

All the data of having 72 major class labels were brought over the scale of unit variance and zero mean.First, the numerical features are extracted, then the preprocessing, normalization and cleaning processes areperformed. After that the numerical feature space is provided to classifier to predict weather climate. The figure1illustrated the framework of proposed model for weather prediction system.

Waikato Environment for Knowledge Analysis (Weka): All the algorithms and Data Mining techniques were employed using very well-known open source package Weka. Weka is a software environment that integrates several machine learning tools with in a common framework and a uniform GUI. Classification and summarization are the main data mining task supported by the weka system [20-27]. It is developed in java and comprised of many power full algorithms [20].

Classification

KNN: KNN is aclassification technique that is much more resembles with the cluster algorithm[12]. KNN is widely and extensively used for supervised classification, estimation and predictionproblems [18, 23, 24]. Unlike SVM global classification model, k nearest neighbor or KNN classifier determines the decision boundary locally [27].

The decision of classification depends upon the nature of the collected data used;KNN is using majority voting technique in case of classification whereas in case of regression problems it uses average response technique.The KNN uses Euclidian distances formula to calculate the distance between the instances. Calculated distance values are sorted in ascending order $d_i = d_i+1$, i = 1, 2, 3...k, where kis the number of samples.

The value havingleast calculated distance with the novel sample then that novel instance calculated value is classified on that class. Unlike supervised learning algorithm the KNNdoesn't learn an explicit mapping function f from the underlined data. KNNuses the training data ata test time to make prediction. Selection of method to compute distances depends upon the data type of the features.If the feature ((xi ? RD) is real values then Euclidean distance is used.

$$d(x, y) = \sqrt{\sum_{m=1}^{D} (x_{im} - x_{jm})^2} == \sqrt{\|x_i\|^2 + \|x_j\|^2 - 2x_i t x_j} \tag{3}$$

Table 1: Attributes of Underline Meteorological Dataset

Sn	Attributes Name	Type	Description
1	Date/Time	Date Time	Denotes long string of Date Time Current*
2	Year	Numeric	Represent Year
3	Month	Numeric	Represent Month
4	Day	Numeric	Represent Day
5	Time	String	Represent only Time of the String
6	Data Quality	String	Represent Partner Data quality measure
7	Temp (°C)	Numeric	Represent Temperature in Centigrade.
8	Temp Flag	String	Temperature Flag
9	Dew Point Temp (°C)	Numeric	Temperature of Dew point recorded
10	Rel Hum (%)	Numeric	Relative Humidity index Measure in Percentage
11	Wind Dir (10s deg)	Numeric	Direction of wind
12	Wind Spd (km/h)	Numeric	Speed of the Wind Measure in Kilometer per hour
13	Visibility (km)	Numeric	Visibility sight in square kilometer
14	Stn Press (kPa)	Numeric	Atmospheric pressure.
15	Hmdx	Numeric	Humidity Index
16	Wind Chill	Numeric	Wind Chill, Sense that feels how cold or warm the air is

Table 2: Statistics of the underline dataset attributes

		Min	Max	Mean	Std. Deviation	Skewness	Kurtosis
1	Temp (°C)	-17.6	35.1	9.443	10.113	-0.04	-0.83
2	Dew Point Temp (°C)	-21.5	24.3	4.991	9.5662	-0.16	-0.84
3	Rel Hum (%)	16	101	75.9	16.941	-0.58	-0.37
4	Wind Dir (10s deg)	0	36	19.4	9.858	-0.49	-0.99
5	Wind Spd (km/h)	0	69	15.79	9.06	1.04	1.472
6	Visibility (km)	0	80	17.02	8.14	0.321	3.525
7	Stn Press (kPa)	95.6	100.9	98.76	0.7005	-0.41	0.856
8	Hmdx	25	44	30.29	3.982	0.896	0.447
9	Wind Chill	-28	-1	-9.65	4.598	-0.87	0.522

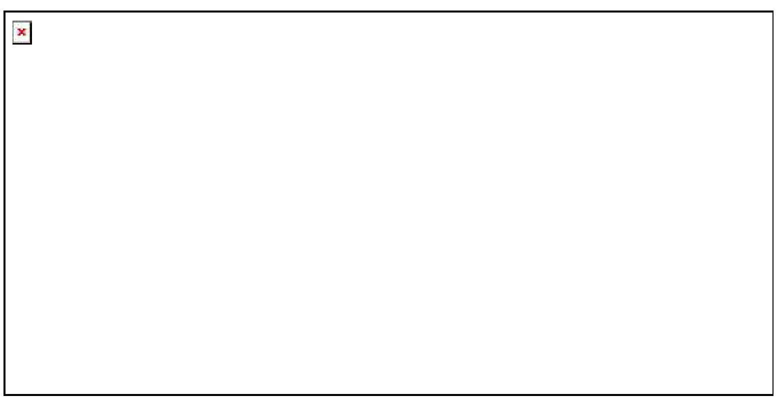

Fig. 1: Framework of Proposed Weather prediction Model

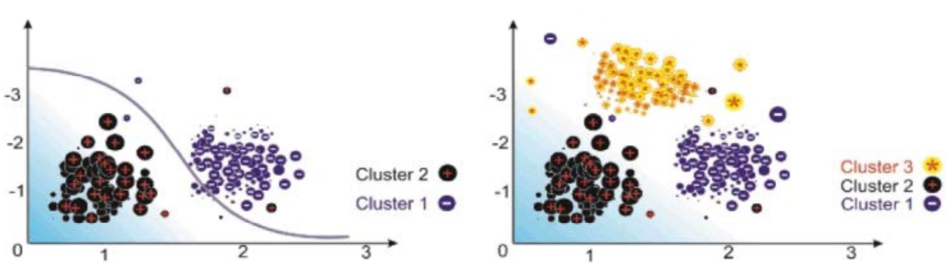

Fig. 2: KNN Cluster Separation

$$||Xi||^2 = \sqrt{\sum_{m=1}^{D} x_{im}^2} \quad (4)$$

Equation 4 is norm of x_i it is also length off vector x. The choice for the K-Neighborhood size is data dependent or heuristic based. For small value of,K KNN comes with small regions for each class and leads to non-smooth decision boundaries and overfit. While in case oflarge value of K it creates fever smother regions, which often resulting in underfit. So the value of Kshould be chosen between maximum and minimum.

Naive Bayes: A Naïve Bayes Classifier is a simple probabilistic (Bayes Theorem) classifier, comprising of strong independent assumption of features.Naive Bayes classifier can be efficiently trained over a sample training data. The naive Bayes classifier assume the attribute list $X_1 \ldots X_n$ are all of conditionally independent of each other given Y. So the value of this assumption is that it simplifies the representation of P(X|Y) and the problem of estimating it from the underline training data.

$$P(X|Y)=P(X_1|Y)P(X_2|Y) \quad (5)$$

$$P(X1\ldots Xn|Y)=\prod_{i=1}^{n} P(Xi|Y) \quad (6)$$

$$P(Y=y_k|X_1\ldots X_n)=\frac{P(Y|yk)\prod_i P(Xi|Y=yk)}{\sum P(Y=yi)\prod_i P(Xi|Y=yj)} \quad (7)$$

Decision Tree: Decision Tree (DT) isdata mining methodology applied in many real world applications as a powerful solution to classification problem [27]. DT is a hierarchical model for supervised learning where the local region is identified in a sequence of recursive splits through decision nodes with test function [8, 12, 22, 25].

Entropy and information gain are used for evaluating worthy attributes. These highly informative features are placed higher up in the tree.Entropy measures the randomness / uncertainty in the data. e.g. Entropy for S examples with Cnumber of classes is given below

$$H(S) = -\sum_{c\in C} P_c \log_2 P_c \quad (8)$$

Entropy is the measure of the "Degree of Surprise". Small entropy means more certainty and high entropy means more uncertainty.

While the Information Gain (IG) measures the increase in certainty. For example if we have S_f be the total number of elements in S feature space of comprising F values, then IG can be give as follow

$$IG(S, F)= H(S) - \sum_{f\in F} \frac{|S_f|}{|S|} H(S_f) \quad (9)$$

Evaluation: The performance and efficiency of the propose model was evaluated using sub sampling test 10-fold cross validation. 10-fold cross validation is applied by most of the researchers to evaluate their computational model due to unique resultas well as less computational cost [21]. In the current study, we have adopted 10-fold cross validation test.

We have used the following performance measures including accuracy, sensitivity, false positive rate, MCC, AUC.

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \text{TP Rate} \quad (10)$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \text{1-FP Rate} \quad (11)$$

$$ACC = \frac{TP+TN}{TP+FP+TN+PN} \quad (12)$$

$$MCC = \frac{TPIN-FPFN}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}} \quad (13)$$

RESULT AND DISCUSSION

The Experiment is carried out over the weather dataset comprising 15000 records. The dataset having 16 prominent attributes is shown in the Table 1.

Various preprocessing steps are carried over the data. The noise, outlier, duplicate and null values areremoved. The data is normalized for further mining step.

First, we have examined the performance of classification algorithms on different folds. The success rates of classification algorithms on different folds are shown in Table 3. Among classification algorithms DT has achieved promising results compared to other used algorithms. The predicted outcomes of DT are 82.62% of accuracy, 0.826 of sensitivity, 0.029 FP rate and MCC of 0.992.

Table 3: Performance of different classifiers using 10-fold cross validation test.

	KNN	Naive Bayes	Decision Tree
Acc %	52.82%	43.43%	82.62%
Sen	0.524	0.434	0.826
FP Rate	0.101	0.393	0.029
MCC	0.426	0.863	0.992

Accuracy %, Sensitivity, FP Rate (False Positive Rate), MCC Mathew Correlation Coefficient

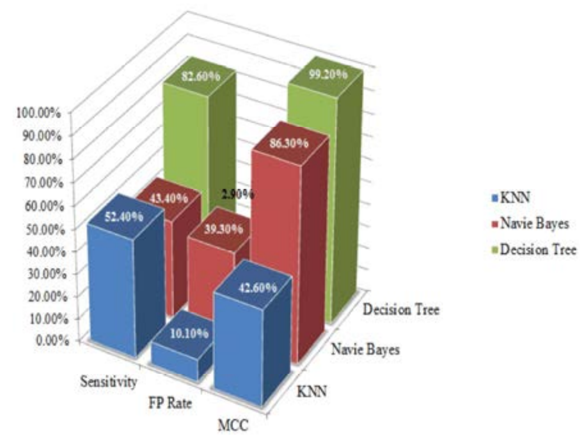

Fig..3: Different classifiers predictive measure over 10 fold of cross validation.

Table 4: Performance Comparison of various classification algorithms using K-fold cross validation tests

	Decision Tree				Naïve Bayes				KNN			
Fold	Sn	FP-rate	MCC	Acc%	Sn	FP rate	MCC	Acc%	Se	FP rate	MCC	Acc%
2	0.777	0.029	0.727	77.66	0.394	0.393	0.325	39	0.48	0.101	0.374	48.02
3	0.801	0.025	0.778	80.14	0.404	0.384	0.325	40.38	0.503	0.095	0.401	50.28
4	0.802	0.024	0.778	80.2	0.417	0.382	0.353	41.7	0.505	0.095	0.403	50.46
5	0.809	0.025	0.783	80.9	0.416	0.382	0.349	41.56	0.514	0.092	0.415	51.4
6	0.811	0.023	0.785	81.12	0.419	0.381	0.355	41.86	0.514	0.093	0.413	51.36
7	0.816	0.023	0.792	81.62	0.423	0.377	0.358	42.28	0.518	0.093	0.418	51.76
8	0.817	0.022	0.786	81.66	0.426	0.378	0.365	42.65	0.516	0.093	0.416	51.56
9	0.822	0.022	0.79	82.16	0.424	0.378	0.359	42.41	0.515	0.092	0.416	51.52
10	0.812	0.023	0.79	81.24	0.422	0.378	0.356	42.18	0.522	0.091	0.425	52.22
11	0.821	0.022	0.992	82.1	0.428	0.378	0.364	42.77	0.518	0.092	0.419	51.8

Sc. Sensitivity, Sp. Specificity, Acc % Accuracy %, MCC. Mathew Correlation Coefficient, FP Rate, False Positive Rate

CONCLUSION

The goal of this paper was to develop a robust computation model for weather prediction. We have downloaded the data for weather from a well-known climatological sit. We have incorporated the whole feature space for evaluating the predictive measures, worth of the attributes were calculated on information gain ration and entropy. In this paper we have achieved overall accuracy of 82.62%, sensitivity of 0.8260, FPRate of 0.029 and MCC of 0.992. The empirical results show that the proposed model can be very effective and a robust computation model in hourly based weather prediction.

REFERENCES

1. FolorunshoO laiya, 2012. Application of Data Mining Techniques in Weather Prediction and Climate Change Studies, I.J. Information Engineering and Electronic Business, Published Online February 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijieeb.2012.01.07, 1: 51-59.
2. Iza Sazanita Isa, Saodah Omar and Zuraidi Saad, 2010. Weather Forecasting Using Photovoltaic System and Neural Network, Second International Conference on Computational Intelligence, Communication Systems and Networks.
3. Harshani, R., K. Nagahamulla, Uditha R. Ratnayake and Asanga Ratnaweera, 2012. An Ensemble of Artifcial Neural Networks in Rainfall Forecasting, The International Conference on Advances in ICT for Emerging Regions - ICTer, pp: 176-181.
4. Jayanta Basak, Anant Sudarshan, Deepak Trivedi and M.S. Santhanam, 2004.Weather Data Mining Using Independent Component Analysis, Journal of Machine Learning Research, 5: 239-253.
5. Xiang Li and Beth Plale, Real-time Storm Detection and Weather Forecast Activation through Data Mining and Events Processing.
6. Godfrey C. Onwubolu, Petr Buryan and Sitaram Garimella, 2007.Self-Organizing Data Mining for Weather Forecasting, IADIS European Conference Data Ming.

7. Gurbrinder Kaur, 2012. Meteorological Data Mining Techniques: A Survey International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, 2(8).
8. Heckerman, D., 1996. Bayesian Networks for Knowledge Discovery. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy,. Menlo Park, CalifAAAI Press, pp: 273-306.
9. Bondalapati, Krishna D. Jeffrey M. Stein and Kathleen M. Baker, Neural Network Model to Predict Deoxynivalenol (DON) in Barley using Historic and Forecasted Weather Conditions.
10. Cofiño, A.S., J.M. Gutiérrez, B. Jakubiak and M. Melonek, 2003. Implementation Of Data Mining Techniques For Meteorological Applications, Realizing Teracomputing. W. Zwieflhofer and N.Kreitz, Editors World Scientific, pp: 215-240.
11. Mohsen Hayati and Zahra Mohebi, 2008. Application of Artificial Neural Networks for Temperature Forecasting, International Journal of Engineering and Applied Sciences, 4: 3.
12. Madhuri V. Joseph, LipsaSadath and VanajaRajan, 2013. Data Mining: A Comparative Study on Various Techniques and Methods, International Journal of Advanced Research in Computer Science and Software Engineering, Research Paper Available online at: www.ijarcsse.com ISSN: 2277 128X, 3(2).
13. Long Jin, Lin Jianling and Lin Kaiping, 2006, Precipitation Prediction Modeling Using Neural Network and Empirical Orthogonal Function Base on Numerical Weather Forecast Production, Proceedings of the 6thWorld Congresson Intelligent Control and Automation, Dalian,China.
14. SanthoshBaboo Dr. S. and I.KadarShereef, 2010. An Efficient Weather Forecasting System using Artificial Neural Network, International Journal of Environmental Science and Development, October ISSN: 2010-0264, 1(4).
15. Singh, Dr. Yashpal, Alok Singh Chauhan, Neural Networks in Data Mining, Journal of Theoretical and Applied Information Technology 2005 - 2009 JATIT. All rights reserved. www.jatit.org.
16. Devi, Ch. Jyosthna, B. Syam Prasad Reddy, K. Vagdhan Kumar, B.MusalaReddy and N.Raja, 2012. ANN Approach for Weather Prediction using Back Propagation, International Journal of Engineering Trends and Technology, 3(1).
17. Ghanbarzadeh, A., A.R. Noghrehabadi, E. Assareh and M.A. Behrang, 2009. Solar radiation forecasting based on meteorological data using artificial neural networks, IEEE International Conference on Industrial Automatics (INDIN 2009), pp: 227-231.
18. Zahoor, Jan, M. Abrar, Shariq Bashir and Anwar M. Mirza, 2008. Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique, Springer-Verlag Berlin Heidelberg.
19. Kumar, Rajesh, 2013. Decision Tree for the Weather Forecasting, International Journal of Computer Applications 76(2): 0975 -8887.
20. Inamdar, N.M., K.C. Ehrlich, M. EhrlichIannello, R.C. Frank and E. HallK, 2004. Data mining in bioinformatics using Weka. Bioinformatics, 20: 2479-81.
21. Chou, K.C. and H.B. Shen Cell-PLoc, 2008. A package of web-servers for predicting sub- Cellular localization of proteins in various organisms. Nat. Prot., 3: 153-62.
22. Hayat, M. and A. Khan, 2013. WRF-TMH: Predicting Transmembrane Helix by Fusing Composition index and physicochemical properties of Amino Acids, Journal of Amino Acid, 44(5): 1317-28.
23. Hayat, M. and A. Khan, Prediction of Membrane Protein Types by Using Dipeptide and Pseudo Amino Acid Composition based Composite Features, IET Communications, 6: 3257-3264.
24. Hayat, M. and A. Khan, 2011. Discriminating Outer Membrane Proteins with Fuzzy K-nearest Neighbor algorithms based on the general form of Chou's PseAAC, Protein and Peptide Letters, 11: 411-421.
25. Hayat, M., A. Khan and M. Yeasin, 2012. Prediction of Membrane Proteins Using Split Amino Acid Composition and Ensemble Classification, Journal of Amino Acids, 42: 2447-2460.
26. Shu Fang Wu, Jie Zhu and Yan Wang, 2012. Weather Forecasting using Naïve Bayesian, Advances in Intelligent and Soft Computing, 159: 337-341.
27. Mehmed Kantardzic, DATA MINING Concepts, Models, Methods and Algorithms, IEEE Press 445 Hoes Lane Piscataway, NJ 08854 IEEE Press Editorial Board.