IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-4, Special Issue-5, Jul 2016 E-ISSN: 2347-2693

Comparative Analysis of Data Mining Techniques for Weather Prediction

Abhishek Jaswal[1], Yashwant Singh[2], Pradeep Kumar Singh[3]

Department of Computer Science &Engineering
JaypeeUniversity of InformationTechnology, Solan, India

Available online at: www.ijcseonline.org

Abstract- Weather forecasting is critically important application in meteorology and has been one of the most scientifically and technologically challenging problem through out the globe since past and still only approximations are being made to accurate prediction of weather eventslike cloudburst,hailingetc. Machine learning algorithms are implemented in data mining process to extract hidden patterns and useful information from huge weather databases, essential to get prepare for the worst of the ambience.

This paper propounds the comparative analysis of various data mining techniques applied by different researchers in different domains. This survey also reviews the available literatures of algorithms applied by different researchers to exploit various data mining techniques for detecting and predicting weather events. For weather prediction Decision Tree, Artificial Neural Network and SVM techniques gives better results with high prediction accuracy than other data mining techniques for multidimensional weather data sets.

Keywords-Data mining; Decision Tree; Hailstorm; Machine Learning; Multilayer Perceptron; Support Vector Machine

I. INTRODUCTION

Weather is one of the most functional and powerful constraint, which affects our lives in many ways. We modify ourselves with respect to the climate from our eating, dressing like habits to other planning activities such as agriculture, vegetation and tourism. Since the adverse effects of weather may cause a great loss to our lives, belongings and properties. During the last decade availability of climate data has increased tremendously due to enhanced technology. And it is important to find effectual and accurate tools to examine and extract hidden useful information from this huge amount of data, which can play a very crucial role in perceiving the climate volatility in future. As various sectors like agriculture, vegetation, water resources and tourism are dependent on climate.

The word KDD (knowledge discovery in databases) was conceivedfirstly at the KDD workshop in order to emphasis upon the concept that Knowledge is the final destination of a data driven discovery[1]. Knowledge discovery is a series of systematic steps from collecting the raw data ill the desired useful information.where as data transformation, pattern discovery, pattern evaluation, and knowledge presentation.The technique which one may use to get useful information from the available data depends directly to the domain of application,type of data and especially need of the customer/user.

Data mining is an application of machine learning, as Data Mining concept uses the machine learning algorithms in many different ways to tackle the problem of finding out hidden useful information from huge *Data Mining*[2][5][17]is a specific particular step in this information discovery process.The ultimate final goal is to get high level knowledge from low level data in the context of large data sets.

II. BACKGROUND STUDY

Weather prediction is the application of science and technology to predict the ambience of a given location. A lot of research is in progress in this area till now over the world. Still we get encounter with a number of worst weather situations like flood,cloudburst etc. in Indiaregularly after interval of one or two years regularly.Thus an accurate predictionmethod is crucial for scientists, agriculturists, farmers, tourism industryand disaster management like communities such that natural happenings like floods, cloudburst, hailstorms etc. events can be known in advance toplan and be preparedforevading such happenings.

A. Data Mining[2][18]is the process of discovering interesting patterns from massive amounts of data. As a *knowledge discovery process*, it typically involves data cleaning, data integration, data selection,

amount of raw meteorological data.

B. Machine Learning controls how computers can learn andboost their performance based on data. Main aim of a research aspect since a long time is to make computers such intelligent that they can make independent intelligent decisions in any situation without human intervention. They should implicitly learn to observe, identify complex patterns and make intelligent decisions based on the dataon its own.

1) Supervised Learning: based upon the learning from available data and constructs amodel whichclassifies the new tuples to a particular class. Some examples of supervised learning techniques are SVM,decision tree[19] etc.

2) Unsupervised Learning: implies clustering because initially no predefined classes are there in the data set. Clusters are built from the tuples which holds some similarity and after that user can map these clusters to a particular class. Commonly we use clustering to identify the classes within the data. Unsupervised model built cannot tell us about the semantic meaning of the clusters identified, because training data is unlabeled.eg. KNN.

3) Prediction: is the most important attribute, quality of data mining which uses set of class labeleddata tuples to design a model that can assigna particular classto the unlabelednew instancesbased upon the learning from training data.The way we exploit the machine learning algorithms depends directly upon the kind of data we are going to use for learning phase and the type of information we expects to be mined. As some time we need to classify instances,cluster unlabeled data,find relation among independent and dependent variables etc. to make predictions.

III. LITERATURE REVIEW

As weather data is high dimensional data carrying multiple interdependent attributes like wind speed, wind direction,humidity,temperature etc.And to discover the desired knowledge or useful information from data a number of data mining approaches are available which are implemented by various researchers and deployed in practical scenario to predict the future events so that it necessary measures can be taken to prevent severe damages caused by weather events like hailing,cloudburst etc. to protect human lives and crops.

Data mining algorithms can be classified in two aspects i.e. supervised and unsupervised and what systematic steps are followed in order to reach the final destination to be able to predict or deploy the technique.

The work of various researchers has been analyzed, with respect to technique wise which they applied in predicting weather related phenomenon.

1. DECISION TREE

A supervised classification algorithm [19] which have different varieties in internal node splitting algorithm like ID3, CART, C4.5, Gini Index have been implemented by authors in predicting temperature, rainfall, evaporation, wind speed and weather events.[3,8,9,11]. As it is supervised algorithm,previous years or days data is used for training the classifier and then used for prediction purpose of unlabeled data.

2. SUPPORT VECTOR MACHINE

SVM again a classification algorithm used for differentiating or classifying the data/tuples in to different classes based upon the maximum margin concept between the support vectors. Many kernel functions are used for transforming the high dimensional data into different plane where it gets easy to partitions the data into two classes. Used for predicting max temperature and weather events[21,22,23] coupled with regression technique aiming to enhance the performance or predicting power of the algorithm.

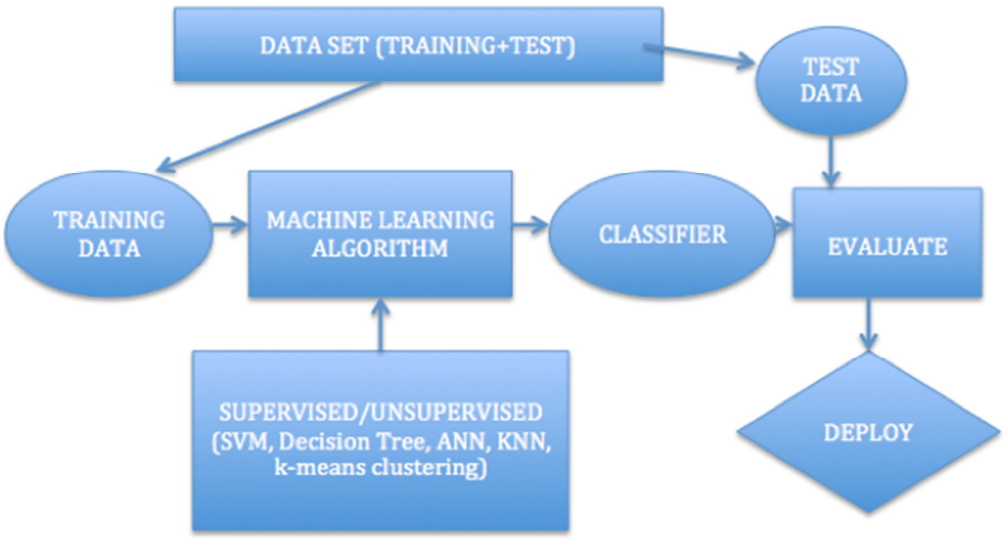

Fig 1. Steps followed from training to deploying of a classifier.

3. K-NEAREST NEIGHBOR

KNN as name implies consider the K nearest numbers of already classified tuples in order to assign the label to newer one. Nearest measured by Euclidean distance. Other distance measures like Manhattan, Minkowski can also be used, but as errors are normally distributed thus solution by least square i.e. Euclidean distance gives appropriate results. Temperature, humidity, weather events have been predicted individually or by mixing with clustering technique.[4,6]

4. ARTIFICIAL NEURAL NETWORK

ANN having some input layers, hidden layers and output layers is used for predicting the weather events by using the previous data(supervised) in order to adjust its weights associated with each of the neurons/nodes in hidden layers nodes,and running the model till the errors gets minimized to the already classified instances,termed as back propagation neural network and as Multi Layer Perceptron in *WEKA*[20].used for temperature,rainfall,wind speed prediction[5,7,8,13]

5. CLUSTERING

Unsupervised technique used for grouping the data or instances based upon some similarity measures.As we are not aware in the start the label for the tuples,so data is grouped by some similar feature and given a class name,and thus followed by adding the more tuples having same characteristics. In practical way can be understood as when a new species of animal or reptiles is found which have different characteristics to all others already known to human kind ,a new name is given to that one and all other found in future having similar features are assigned to this new category. Used for temperature,humidity,cloudburst prediction[6,10,16]. Also exploited by mixing with KNN too.

IV. COMPARATIVE ANALYSIS

A comparison of all the data mining techniques applied by different authors is done by examining the type of data upon which a technique is applied, algorithm, size of data,attributes,accuracy,percentage of error etc. measures.After doing above analysis,we observed that as weather data comprises of several different attributes which are not independent of others attributes for example humidity and rainfall. Thus those techniques are useful in mining the information from raw data, which take into account weightage of every attribute equally, in predicting the class attribute or future event.

Table 1:Comparison of Mining Techniques reviewed in Literature For Weather Prediction

Authors	Application	Techniques	Algorithm	Attributes	Data Set size
E.G.Petre[3]	Weather prediction	Decision Tree	ID3, C4.5	Climate,humidity,storm,temp	30 instances
Zahoor Jan et al[4]	Interannual climate prediction	Lazy learning	KNN	Windspeed,dew,sealevel,snowdepth,rain	40000 instances
Sarah n kohail[5]	Daily temperature prediction	Clustering, regression,ANN,association rules	K-means clustering,MLP ANN	humidity,wind speed, temp	8 years data
S.Badhiye et al[6]	Humidity, Temp prediction	Lazy learning,clustering	KNN,K-means clustering	Temp,Humidity	-
F Oliya, AB Adeyemo[8]	Weather prediction	Decision Tree,ANN	C4.5,CART,TLFN	Rainfall,temp,wind speed	36k instances
S Yeon et al.[9]	Hourly rainfall prediction	Decision Tree	C4.5,CART	Temp,winddirection,speed,humidity,pressure,gust	26k instances
M. A. Kalyankar, S. J. Alaspurkar[16]	Meteorological data analysis	Clustering	K-means clustering	Temp,humidity,windspeed,rain	8k
SoumadipGhosh et al[15]	Weatherpredictio n	ANN	BPN with Hopfield network	Temp,windspeed,humidity	15k
Y.Radhikaet al[21]	Weather prediction(Max temp)	SVM	Non linear Regression	Temperature	Previous N days temperature
R.Usha Rani et al[22]	weather forecasting	SVM	ESVR	Temp,watervapors,pressure,windspeed,directionetc	
Narasimha et al.[11]	Precipitation (rainfall)predictio n	Decision Tree	SLIQ with Gini Index	Humidity,temp,pressure,dew,wind speed	365-4000 instances
T.Rao et al[23]	Weather	SVM	-	-	5 years

	forecasting(Max Temp prediction)				data
K Pabreja[10]	cloud burst detection	clustering	K-means clustering	temp,humidity	-

Table 2: Advantages and Disadvantages of related work

S.No.	Authors	Advantages	Disadvantages
1	E.G.Petre[3]	verified performance	Can't handle continous range data directly
2	Zahoor Jan et al[4]	accurate results with large attributes	unable to adapt global changes
3	Sarah n kohail[5]	ANN gives better Co-relationcoefficient between actual and predicted.	-
4	S.Badhiye et al[6]	satisfactory for multi modal classes	No prediction in remote areas
5	F Oliya, AB Adeyemo[8]	feasible,accurate model is selected for prediction	accuracy varies with size of training data
6	S Yeon et al.[9]	High prediction accuracy	small data set left for testing prediction
7	M. A. Kalyankar[16]	Good prediction accuracy	Dynamic miningmethods required
8	SoumadipGhosh et al[15]	able to determine non linear relationship between attributes	satisfactory results
9	Y.Radhikaet al[21]	Suitable choice of n(previous days temp) as training give better results	N has to be found by experimentation
10	R.Usha Rani et al[22]	improved results with more kernels and using SOM.	Efficiencyof model varies wrtno.of kernel functions.
11	Narasimha et al.[11]	Better accuracy than J48,ANN	Gini consider only binary splits
12	T.Rao et al[23]	SVM gives better results as to MLP	Parameter selection has significant effect on performance of SVM
13	K Pabreja[10]	supplement with NWP models	accuracy degrades for long term prediction

V. CONCLUSION

This study of available literature concentrates on the various methodologies and techniques available and applied by various authors around the world in the area of weather forecasting. This paper presents the various algorithms implemented by various researchers in predicting various weather phenomena like cloudburst,rainfall,temperature,thunderstorms etc.Comparative analysis of all techniques implies that Decision Tree,SVM,ANN are best in case of weather predictive scenario with appropriate parameters selection. Their performance can be enhanced by qualitative selection of parameters so that future weather events, disasters can be predicted well in time and safety measures be taken.

REFERENCES

[1] U. Fayyad, G. Shapiro, and P. Smyth. "From data mining to knowledge discovery in databases." AI magazine 17, no. 3 ,**1996**, pp. **37-54**.

[2] J. Han, M. Kamber and J. Pei, Data mining: concepts and techniques, Elsevier, 2011.

[3] E. Petre "A Decision Tree for Weather Prediction", Buletinul, Vol. LXI No. 1,**2009** pp.**77-82**.

[4] Z. Jan, M. Abrar, S. Bashir and A. Mirza,"Seasonal to inter-annual climate prediction using data mining KNN technique" In Wireless networks, information processing and systems,Springer Berlin Heidelberg,**2008**,pp. **40-51**.

[5] S. Kohail, A. Halees, "Implementation of Data Mining Techniques for Meteorological Data Analysis", IJICT Journal Volume 1 No. 3, **2011**.

[6] S. Badhiye, P.N. Chatur, B. Wakode,"Temperature and Humidity Data Analysis for Future Value Prediction using Clustering Technique: An Approach", International Journal of Emerging Technology and Advanced Engineering, **2250-2459**, Volume 2, Issue 1, January **2012**.

[7] S. Baboo and I Shereef, "An Efficient Temperature Prediction System using BPN Neural Network." International Journal of Environmental Science and Development 2,no.1,**2011**,pp.**49-54**.

[8] F. Olaiya, A. Adeyemo, "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", I.J. Information Engineering and Electronic Business,**2012**,pp. **51- 59**.

[9] S.Yeon, S. Sharma, B. Yu, D. Jeong, "Designing a Rule-Based Hourly Rainfall Prediction Model", IEEE IRI 2012, August **2012**.

[10] K. Pabreja,"Clustering technique to interpret Numerical Weather Prediction output products for forecast of Cloudburst", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 3 (1),**2996 – 2999**, **2012**.

[11] N. Prasad,P. Kumar and M. Naidu,"An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree",4th International Conference onIntelligent Systems Modelling& Simulation (ISMS), IEEE ,january**2013**, pp. **56-60.**

[12] A. Saxena, N. Verma and K. Tripathi, "A review study of weather forecasting using artificial neural network approach" International Journal of Engineering Research and Technology,Vol. 2. No. 11 ESRSA Publications, **2013**.

[13] A. Naik and S. Pathan,"Weather Classification and forecasting using Back Propagation Feed-forward Neural Network." International Journal of Scientific and Research Publicationsvol2 no.12,**2012**.

[14] GurbrinderKaur, "Meteorological Data Mining Techniques: A Survey",International Journal of Emerging Technology and Advanced Engineering, Volume-2, Issue-8, August **2012**,pp. **325-327**,

[15] S. Ghosh, A. Nag, D. Biswas, J.Singh, S. Biswas, D. Sarkar and P. Sarkar, "Weather data mining using artificial neural network. In Recent Advances in Intelligent Computational Systems" IEEE 2011 pp. **192-195**

[16] M. Kalyankar, S. Alaspurkar, " Data Mining Technique to Analyse the Metrological Data", International Journal of Advanced Research in Computer Science and Software Engineering 3(2), February – **2013**,pp. **114-118**.

[17] A. Ganguly and K. Steinhaeuser, "Data mining for climate change and impacts" IEEE International Conference onData Mining Workshops, **2008**,pp. **385-394**.

[18] N. Japkowicz and M. Shah, "Evaluating Learning Algorithms: A Classification Perspective" First Edition, Cambridge University Press, **2011**.

[19] L. Rokach and O. Maimon, "Data mining with decision trees: theory and applications" World scientific,**2014**.

[20] G. Holmes, A. Donkin and I.Witten, "Weka: A machine learning workbench",Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems,IEEE**1994**.

[21] Y. Radhika and M. Shashi, "Atmospheric temperature prediction using support vector machines", International Journal of Computer Theory and Engineering, **2009**,1(1), p.55.

[22] R. Rani and T. Rao, "An Enhanced Support Vector Regression Model for Weather Forecasting". IOSR Journal of Computer Engineering (IOSR-JCE) ,**2013**,pp. **2278-0661**.

[23] T.Rao, N. Rajasekhar and D. Rajinikanth, "An efficient approach for Weather forecasting using Support Vector Machines" In Proceedings International Conference Computer Technology Science(ICCTS),Vol. 47, pp. **208-212**.

