WEATHER PREDICTION USING CLASSIFICATION

Mohammad Abrar [*1, 2], Alex Tze Hiang Sim [3], Dilawar Shah[4], Shah Khusro[5], Abdusalam[6]

[1] Faculty of Computing, Dept. of Computer Science, Universiti Teknologi Malaysia, Malaysia

[2] Assistant Professor, Dept. of Computer Science, Kardan University, Kabul, Afghanistan

(Email: muhammadabrar78@gmail.com)

[3]Senior Lecturer, Faculty of Computing, Dept. Of Computer Science, Universiti Teknologi Malaysia, Malaysia

(Email: Alex.sim@utm.my)

[4]Assistant Professor, Dept. Of Computer Science, Bacha Khan University, Charsadda, Pakistan

(Email: dilawar_shah@yahoo.com)

[5]Assistant Professor, Dept. of Computer Science, University of Peshawar, Pakistan

(Email: khusro@upesh.edu.pk)

[6]Assistant Professor, Dept. of Computer Science, Abdul Wali Khan University, Mardan, Pakistan

(Email: abdusalam@awkum.edu.pk)

*Corresponding Author

ABSTRACT: *Weather is the single most events that affect the human life in every dimension, ranging from food to fly while on the other hand it is the most disastrous phenomena. Therefore, prediction of weather phenomena is of major interest for human society to avoid or minimize the destruction of weather hazards. Weather prediction is complex due to noise and missing values dataset. Numerous efforts were made to make weather prediction as accurate as possible, but still the complexities of noise are affecting accuracy. In this article, we propose a comprehensive prediction technique that utilizes the modified version of kNN classification. MkNN is more robust towards the noise and produces more accurate results. The proposed technique is comprehensive, as it is capable to predict almost 17 different weather attributes. The data used in this research is collected from National Climatic Data Center and the analyses were carried out by comparing the WP-MkNN with state-of-the-art classification and prediction models. The extensive experimental results showed that the proposed technique is more accurate even in noisy dataset*

Keywords

Climate Prediction, Weather forecasting, data Mining, MkNN

1 INTRODUCTION

Technological improvements in the computational power are still not sufficient to handle the weather prediction efficiently [1, 2]. Although the current abilities of computer systems helped the meteorologist to implement more advance model that requires high computation and improves the prediction capabilities; the accuracy and timely prediction of weather phenomena is still a major issue. Further, the global climate changes and incident of some disastrous weather events increased the importance of timely and accurate weather prediction. In 2011, the hazardous weather events caused more than $50 billion loss to US and 675 deaths [3]. Pakistan faced deadly flood of its history which caused 1985 human loss and $9.8 billion financial loss. Russia gone through the worst ever drought and the deadliest heat wave in human history. Australia and Columbia faced a record natural disaster due to heavy rain[4]. Another Red signal is issued by the World Economic Forum, who reported that the 21st century climate change will be one of the greatest global challenges for human [2].

Despite the fact that none can control the natural disasters but accurate and timely prediction model can help to minimize loss of human lives and other financial cost, which became the motivating factors for the development of weather prediction model. Improved model can predict the event accurately and as early as possible. To achieve this goal, across the globe, many national and international organizations are making efforts for better climate prediction; including: World Meteorological Organization (WMO), International Research Institute (IRI), World Climate Application and Services Program (WCASP). These are the major contributors, dedicated for integrating high quality research to predict environmental changes, protect life and property, and provide reliable, timely, and scientific information to the decision makers [5, 6].

Current weather prediction models can be divided, broadly, into two categories, numerical models and statistical models. Both are working on the same basic principle i.e. take input data and apply either complex numerical equations or heavy statistical inferences for prediction. The inputs for these models are called initial conditions that consist of climatic observations. Special purpose devices collect these observations; these devices can be sensors, radar, aircraft, air balloons, ships etc.

The current weather prediction models provide reasonable results but they have the following limitations;

(a) They use very complex calculation for which the fastest supercomputers may take long time to produce the results[7].

(b) A smallest error in initial condition can lead to erroneous results and mostly these smallest errors doubled every five days which make long run prediction nearly impossible [8].

These limitations and the availability of huge volume of climatic data draw the attention of researcher to utilize Data Mining techniques in weather prediction.

Data Mining is a proved technique for identifying unknown pattern and accurate prediction in complex and large datasets[9, 10]. It explores hidden patterns, relationships and interdependencies that other traditional techniques might overlook. Therefore, using data mining, decision makers and

business executives can learn the novel and interesting information from the historic data. The discovered knowledge is equally useful for a wide and diverse range of applications, such as market basket analysis, customer purchasing pattern, fraud detection, health care, weather forecasting, and telecommunication.

In this paper we propose an enhanced weather prediction model that utilized the concept of Data Mining into weather prediction. Our specific contributions in this paper are listed below.

a) An extensive preprocessing method is presented that brings the diverse data into a uniform format and converts data to numeric form, where it is necessary.

b) The model is capable to predict any attribute available in the dataset while other data mining based algorithm produces results for specific attributes.

c) The proposed model has the capability to predict weather beyond two weeks, which also make it a climate prediction tool.

d) A detailed evaluation against other prediction data mining tools is performed that provide a fair comparison to show the effectiveness of the new model.

e) The new model is computationally efficient and makes it suitable for small devices such as android environment.

The rest of the paper is organized as the next section reviewed the literature, which is followed by the description and working of the new model in section 3. Section 4 analyzes and discusses the result and finally section 5 concludes the paper with future directions.

2 LITERATURE REVIEW

In this section we provide critical analysis of the existing prediction models. The analysis is divided into two subsections where one explains the models used in traditional weather prediction i.e. numerical and statistical models and second subsection deals with the data mining attempts made in weather prediction.

2.1 Traditional weather prediction Models

Generally, the existing models use complex statistical equations and inferences that are extremely expansive in term of computing and time resources. Most of the models are either running on super computers [13] or mainframe computer while other are using cluster computations [14]. The second issue with these models is that the results produced by these systems are only understandable by the experts. Another issue is related to their layered approach. The process of one model is based on the results of some other complex model and so on. Here a short description of the major, in use, climate prediction systems is presented.

In [15] Hansen et al presented a global atmospheric model – I with good computational efficiency. This model is capable of long-range prediction. The updated version is called model – II which is the result of several modifications to model – I.

European Community HAMburg 5 (ECHAM5) was developed at Max Planck Institute for Metrology, Germany. It is an improvement of ECHAM4. In contrast to the previous version, several changes incorporated in this model. These modifications were made in the representation of land surface process and so on. As compared to its previous version, ECHAM5 is more powerful and flexible. The system was tested on various platforms [16].

The National Centers for Environmental Prediction's Global Forecast System (GFS) is a global forecast model. GFS forecasts produce every six hours at 00, 06, 12 and 18 GMT. The performance of the model improved time to time [17].

Community Climate Model (CCM) developed at the National Centers for Atmospheric Research (NCAR) in the United States [18]. Currently its different variants i.e. CCM3, CCM3.2, CCM5, CCM6.2 and CCM6.6 are operational.

2.2 Data Mining and weather prediction

Jan et al. [12] proposed CP-kNN which is new technique for climate prediction on seasonal to inter-annual time scale. The system defers from all the above system as it works on the historical numerical data and uses DM techniques KNN (K Nearest Neighbor) for prediction. It is capable to predict the climate of a region well in advance with reasonable accuracy. The technique can forecast up to seventeen climatic attributes e.g. minimum temperature, maximum temperature, wind, and so on. The predicted results of the system are easier to understand even for nonprofessional individuals.

Another attempt was made by Peter and Frantisek [19] to predict fog for UAE and Slovakia Airports. They preprocessed data and used a combination of decision tree and neural networks for prediction. Their prediction was limited to 1 hour in advance.

A comparative analysis was conducted in [1] for two different classification algorithms C5.0 and ANN over meteorological dataset for year 2000 to year 2009. Using C5.0 they identified monthly pattern and ANN was used to identify relationship between different climatic attributes.

Data Mining was also applied to predict the rainfall using reflective data [20]. They used five different classification algorithms i.e. neural networks, CART, support vector machine, kNN and random forest. They developed three different models (model – I, Model – II and Model – III). The authors first evaluated the five algorithms using model – I where NN outperform the other. Then the model – II and III were evaluated using NN. Their analysis concluded that Tipping bucket gauge data can improve the rainfall prediction.

K-Nearest Neighbor –kNN [21] is a well-known and widely used classification technique. It has some prominent features like its simplicity, robustness against noise in the training data and its effectiveness in case of large training data. Despite its merits, kNN suffers from following drawbacks.

- It requires high computational cost to compute the distance of each query against all training data.
- Large memory is required for the implementation of the algorithm.
- High dimensional datasets reduce accuracy.
- In distance-based learning, it is very difficult to determine the suitable distance measurement technique.

Parvin [11] enhanced the kNN algorithm and came up with Modified k-Nearest Neighbor (MkNN). MkNN partially overcomes the problems of low accuracy of the kNN. MkNN first preprocess the training set by computing the validity of each sample in the training data and eliminate those instances which fail the validity test; it also gives

weight-age to every valid instance. This reduces the computation to only valid records. Once this computation is accomplished, classification then applies the weighted kNN on the test data.

The literature is evident of the successful efforts of Data Mining in weather prediction; still there is a need for more generalized model that is not limited to the prediction of single attribute but capable to deal with multiple attributes at the same time. The WP-MkNN (Weather Prediction using Modified k NN) is an attempt to provide a solution to this issue. In next section the proposed technique with detailed explanation is presented

3 WEATHER PREDICTION USING MODIFIED *k* NN

WP-MkNN consists of two major steps; namely: a) preprocessing of dataset and b) prediction. The Preprocessing steps further consist of cleaning dataset, formatting and conversion of attributes types. Prediction step consists of applying MkNN algorithm on dataset to predict the weather attributes well in advance. The data for the model evaluation was downloaded from National Oceanographic and Atmospheric Administration (NOAA) website.

3.1 Dataset description

Since 1900, the global weather and climate data is available at National Climatic Data Center (NCDC) [22]; a sub body of NOAA. Data were downloaded for four stations (Islamabad Airport, Karachi Airport, Lahore Airport and Quetta Airport). The range of dataset is 1980 to 2010 comprising of 30 years for each station. The original datasets consist of 13 climatic attributes where last attribute (FRSHTT) represent binary values for Fog, Rain, Snow, Hial, Thunder and Tornado making total of 19 attributes. The climate data is commonly erroneous due to faulty devices, communication problems and human error. Therefore a thorough preprocessing is required before applying WP-MkNN.

NCDC handles missing values separately for different attributes e.g. if observation is not available for Sea Level Pressure (SLP), it is filled with 9999.9, but for GUST and snow depth (SNDP) the value is 999.9 while for precipitation it is 99.99. Similarly, FRSHTT stores string data where each character represents the binary state of one climatic attribute. Further, sometime station did not report the maximum and minimum temperature. Therefore, they are calculated by the quality control software and marked with '*' to show the difference between reported and calculated values. This '*' converts the type of attribute to string from float.

Another issue is related to unnecessary attributes. There are some attributes that are not needed for our analysis. These attributes represent the number of observation (counts) used in each value. For example, if a daily average temperature is calculated using 21 observation then the temp will show the average value and its count will show 21.

3.2 Dataset Cleansing

In order to properly handle the cleansing process, a program was written in java to overcome the above mentioned issues. All the missing values are converted to '?'. The FRSHTT attribute is divided into six binary attributes where each represents FOG, RAIN, SNOW, HAIL, THUNDER and TORNODO with 0 and 1 numeric values. The minimum and maximum temperature were also converted to floating point variable and * was removed from all the data. Finally, the unnecessary attributes were removed from datasets. These steps were repeated for all four datasets.

3.3 Modified k NN

MkNN consists of two steps. First it computes the validity of every instance in training data and select only valid data instances; then it applies the weighted kNN to compute the distance between other instances.

3.4 Computing the Validity of Training Data.

MkNN first validates every sample and this computation takes place for all samples in the training data. The validation process is explained below.

If we want to validate a sample Attribute (A) in the training dataset, its H nearest neighbors is considered in the validation. The computation is represented in equation **Error! Reference source not found.**

$$V(A)=\frac{1}{H}\sum S(lbl(A),lbl(N_i(A))) \qquad \text{Eq (I).}$$

Where H is the number of neighbors for attribute *A* and *lbl(A)* returns the true class labels of the sample *A*, while $N_i(A)$ is the i^{th} nearest neighbor of Attribute *A*. Similarly, function *S(a,b)* , shown in equation (2), calculates the similarity of point *A* and the i^{th} nearest neighbor.

$$S(a,b)=\begin{cases}1 & a=b\\ 0 & a\neq b\end{cases} \qquad \text{Eq (II).}$$

In WP-MkNN, *S(a,b)* is used to identify the missing values and remove it from the training dataset to reduce the computational and space overhead.

3.5 Applying Weighted kNN.

Instead of using a simple majority vote like kNN, MkNN algorithm uses weighted votes for prediction. In this process each of the K sample is given a weighted vote which is usually equal to some decreasing function of its distance from the given unknown sample. A general formula is given

in equation 0

$$w(i)=Validity(i)\times\frac{1}{d_e+0.5} \qquad \text{Eq (III).}$$

Where *w(i)* is the weight and *Validity (i)* is the validity of the i^{th} sample in the training data. If the validity is false, then its corresponding weight becomes zero. Thus, if a value is not validated, it will not play any role in the prediction. Therefore MkNN increases the overall accuracy of the prediction process. This process also ignores outlier, a major strength of MkNN against traditional kNN.

3.6 Steps for Applying MkNN Algorithm

This section explains the process of MkNN and its application to weather prediction. After the validation process, next step is to apply the MkNN. The algorithm is explained stepwise as below.

Input: Cleans Dataset for particular region/City, Prediction Dates, Prediction Attribute

Output: Prediction for specific range and specified attribute

Step 1:

Select all data from noisy data source, and verify each instance.

While(i ≠ Ø)

If(verified(i)) Then weight(i) = 1

Else Weight(i) = 0

End While

Step 2:

PRED_DATE = sequence_to_be_predicted

BASE_SEQ = (PRED_DATE) – (NO_OF_DAYS)

Step 3:

While days ≠ Ø

Selected_days[ ] = day (if Validated)

End While

Calculate Distance(Selected_days[])

SORT(Selected_days[], Distance)

Step 4:

Find nearest neighbor using value of K

Repeat step 3 and step 4 till all days to be predicted are traversed

Step 5:

Stop when all data is examined

Algorithm 1: Pseudo code for WP – MkNN

Step1: Select all data from noisy data source, and verify each.

While(i ≠ Ø)

If(verified(i)) Then weight(i) = 1

Else Weight(i) = 0

End While

Step1 traverses the entire database and verifies the validity of each parameter, if parameter value is found noisy, zero weight will be given to that record and that record will not participate in prediction process.

Step2:

PRED_DATE = sequence to be predicted

BASE_SEQ = (PRED_DATE) – (NO_OF_DAYS)

The algorithm divides the whole data into equal chunks called sequences where every sequence is equal to the prediction time span i.e. if prediction is for 1 Month, the 12 year dataset will be divided into monthly chunks. It is required for the distance calculation in dataset.

Step 3:

While days ≠ Ø

Selected_days[ ] = DAY(day) of MONTH(month) (if Validated)

End While

Calculate Distance(Selected_days[])

SORT(Selected_days[], Distance)

This step performs the key operation of the algorithm. It selects the similar record from the whole dataset i.e. if we need to predict weather for 1st week of Jan 2003 then this step will select all records of 1st week of January from the whole dataset. Further it calculates its distance and finally it sorts the results according to distance.

Step4:

Find the *K* nearest neighbor and calculate mean

Last step extracts *K* nearest neighbors from the array, and takes its mean as predicted value for a specific day.

Step5: The process stopped when all data is examined. Algorithm 1 summarizes the whole process.

4 EXPERIMENTS AND DISCUSSION

We compared WP – MkNN with five different classes of prediction algorithms i.e. linearRegression – a regression based model, SMOreg – Support Vector Machine variant, MultiLayerPerceptron – a flavor of Artificial Neural Networks, RepTree – a decision tree based technique and finally with CP-KNN – a climate prediction model based on kNN. The experiments carried out on Intel Core i3-2350M, 2.3GHz 64bit, 8GB main memory, running Windows 7. For evaluation purpose, we implemented WP – MkNN and CP – kNN in JAVA, while Weka 3.7.9 [23] was used to generate results of other algorithms. In order to make the comparison unbiased, all the techniques are applied on the same preprocessed datasets that were used for WP-MkNN.

There are two major comparison of WP – MkNN, one with the CP – kNN [12] and second with other prediction techniques mentioned above. CP – kNN and WP – MkNN share the same basic concept, such as prediction using k – nearest neighbors, the former is called climate prediction model. The difference between weather and climate is that weather refers to the atmospheric condition at a specific location and time, for example temperature of New York City at noon. Weather represents these conditions for shorter period of time i.e. less than 15 days and these conditions normally represented by numbers i.e. 57°F Temperature. Climate, on the other hand, is an averaged atmospheric condition of a location at longer period of time and these averages are preferably represented as linguistic or relative terms instead of numbers i.e. hot summer etc. The prediction range of CP – kNN is higher than 15 days therefore it is considered as Climate Model. WP – MkNN can also predict the weather at longer time span but as output produced by WP-MkNN are numbers of daily weather conditions instead of climatic average therefore we call it weather prediction model.

The comparison between CP – kNN and WP – MkNN is shown in Figure 2. The figure is evident that our propose model is performing better than its counterpart for both minimum and maximum temperature, irrespective of K values.

In general, the good prediction model reduces the error rate, therefore comparison based on the error measurement is

provided for all algorithms. The difference between predicted and actual values is shown in Table 1. Further, different Error measurement techniques are applied on Table 1 and the results are summarized in Table 2.

Table 1: MkNN Comparison with other prediction algorithms 1-Feb 2010 to 15 Feb 2010: this table shows the difference between actual and predicted value of algorithm for every predicted day

Prediction Date	Linear Regression	SMOReg	MultiLayer Perceptron	RepTree	Knn	MkNN
1-Feb	-1.007	0.259	2.641	2.018	-3.900	-2.600
2-Feb	-2.316	-1.382	-2.918	-0.869	-4.200	-2.600
3-Feb	-3.761	-2.875	-4.205	-3.031	-4.000	-2.400
4-Feb	-3.359	-2.464	-2.133	-1.857	4.200	-1.700
5-Feb	-0.773	0.029	-3.995	-1.917	3.900	-2.100
6-Feb	-0.103	0.981	-2.215	-0.441	-3.800	2.200
7-Feb	2.704	3.914	1.951	2.734	4.100	2.600
8-Feb	5.497	6.161	4.068	4.800	-3.900	-2.000
9-Feb	2.496	3.443	-1.735	3.491	-3.400	1.800
10-Feb	1.863	3.387	-9.397	5.201	-3.300	2.200
11-Feb	-2.583	-2.126	-3.936	-3.577	-3.100	-2.100
12-Feb	-5.619	-3.843	-12.401	-1.544	3.100	1.400
13-Feb	-1.104	-0.480	-11.375	0.766	3.400	1.700
14-Feb	-4.715	-4.123	-11.496	-3.300	3.100	1.500
15-Feb	3.589	5.072	-3.188	5.524	2.900	2.300

Table 2. Different Error measurements for all algorithms

	Linear Regression	SMOReg	MultiLayer Perceptron	RepTree	kNN	MkNN
MAE	2.703	5.177	2.766	2.738	3.620	**2.080**
RMSE	3.226	6.384	3.206	3.144	3.646	**2.114**
MAPE	24.943	49.202	23.435	27.661	5.679	**3.279**
MSE	10.409	40.751	10.282	9.884	13.294	**4.471**

5 CONCLUSION AND FUTURE WORKS

The results show that proposed model produces very good results as compare to other data mining prediction techniques. Although, we presented the results only for Minimum and Maximum temperature, but the model is capable to predict all available attributes. The efficiency and simplicity of the model make it suitable to utilize it at organization for short term energy recourses requirements and management.

The model can be extended to incorporate other attributes (such as solar radiations, water cycle, air pressure and earth rotation etc) to make the forecast more general and for larger area. It can also be extended to climate prediction in which it will predict the climate at seasonal to inter-annual time scale. The associativity and dependability among the attributes can also produce new pattern and a new way for prediction. The probability of global warming can also increase the new insight of the model.

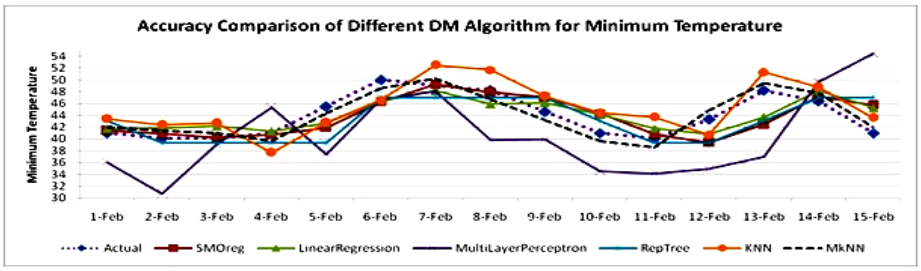

Figure 1: Comparison of Minimum temperature for kNN and MkNN

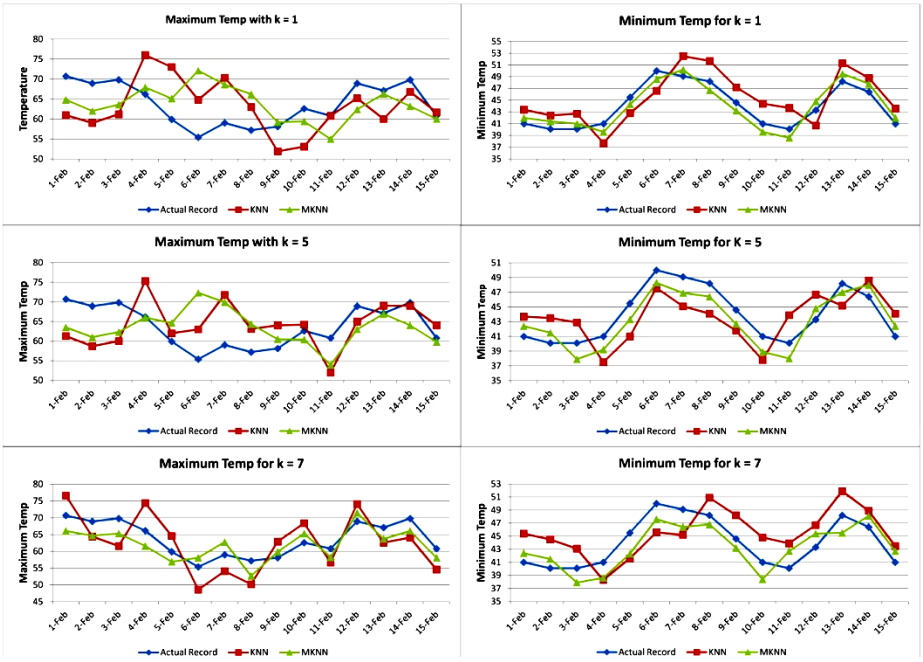

Figure 2 : Comparison of Minimum and Maximum temperature for different values of k

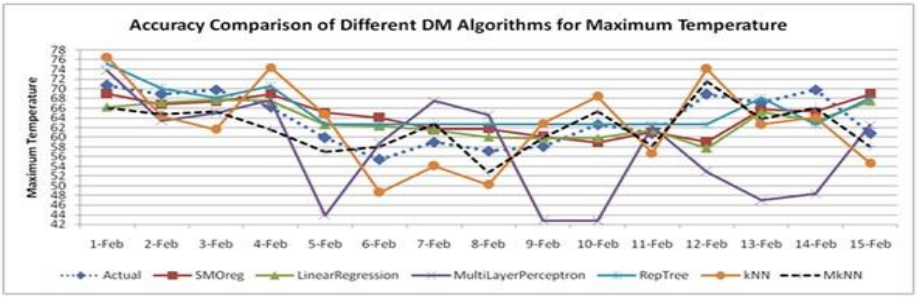

Figure 3: Comparison of Minimum temperature for kNN and MkNN

5 REFERENCES

1 Olaiya, F., and Adeyemo, A.B.: 'Application of Data Mining Techniques in Weather Prediction and Climate Change Studies', International Journal of Information Engineering and Electronic Business (IJIEEB), 2012, 4, (1), pp. 51

2 Lynch, P.: 'The origins of computer weather prediction and climate modeling', J. Comput. Phys., 2008, 227, (7), pp. 3431-3444

3 Underground, W.:'Severe Weather Headlines', http://www.wunderground.com/resources/severe/severe.asp?MR=1,2011, Accessed on 11 December, 2013

4 Masters, J.:'Dr. Jeff Masters' WunderBlog', http://www.wunderground.com/blog/JeffMasters/2010--2011-earths-most-extreme-weather-since-1816,2011, Accessed on December 11, 2013

5 Team, E.P.M.:'Ensemble Prediction Model', http://www.wpc.ncep.noaa.gov/ensembletraining/,2006 , Accessed on 11 December 2013

6 Cox, J.D.: 'Storm Watchers: The Turbulent History of Weather Prediction from Franklin's Kite to El Niño' (John Wiley & Sons, 2002. 2002)

7 Fayyad, U.M., Djorgovski, S.G., and Weir, N.: 'Automating the analysis and cataloging of sky

surveys', in Usama, M.F., Gregory, P.-S., Padhraic, S., and Ramasamy, U. (Eds.): 'Advances in knowledge discovery and data mining' (American Association for Artificial Intelligence, 1996), pp. 471-493

8 Han, J., Kamber, M., and Pei, J.: 'Data mining: concepts and techniques' (Morgan kaufmann, 2006. 2006)

9 Parvin, H., Alizadeh, H., and Minati, B.: 'A Modification on K-Nearest Neighbor Classifier', Global Journal of Computer Science and Technology, 2010, 10, (14)

10 Hansen, J., Russell, G., Rind, D., Stone, P., Lacis, A., Lebedeff, S., Ruedy, R., and Travis, L.: 'Efficient Three-Dimensional Global Models for Climate Studies: Models I and II', Monthly Weather Review, 1983, 111, (4), pp. 609-662

11 Yang, F., Pan, H., Moorthi, S., Lord, S., and Krueger, S.: 'Evaluation of National Centers for Environmental Prediction Global Forecast System at the Atmospheric Radiation Measurement Program Southern Great Plains Site'

12 Drake, J.B., Jones, P.W., and Carr, G.R.: 'Overview of the software design of the community climate system model', International Journal of High Performance Computing Applications, 2005, 19, (3), pp. 177-186

13 Jan, Z., Abrar, M., Bashir, S., and Mirza, A.M.: 'Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique': 'Wireless Networks, Information Processing and Systems' (Springer, 2009), pp. 40-51

14 Bednar, P., Babic, F., Albert, F., Paralic, J., and Bartok, J.: 'Design and implementation of local data mining model for short-term fog prediction at the airport', in Editor (Ed.)^(Eds.): 'Book Design and implementation of local data mining model for short-term fog prediction at the airport' (IEEE, 2011, edn.), pp. 349-353

15 Kusiak, A., Wei, X., Verma, A.P., and Roz, E.: 'Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach', 2013

16 NCDC, and NOAA:'Climate Data Online: Dataset Discovery', ftp://ftp.ncdc.noaa.gov/pub/data/gsod,2013, Accessed on 15 Nov 2013