Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 5, May 2014, pg.348 – 358

RESEARCH ARTICLE

A Framework for an Outlier Pattern Detection in Weather Forecasting

Miss. Kavita Thawkar[1]

Department of Computer Science& Engineering.
G.H.Raisony Academy College of Engineering& Technology, Nagpur, India.
kavithawkar@gmail.com

Prof. Snehal Golait[2]

Department of Computer Science& Engineering.
Priyadarshani College of Engineering& Technology, Nagpur, India.
snehal.golait@gmail.com

Prof. Rushi Longadge[3]

Department of Computer Science & Engineering.
G.H.Raisony Academy College of Engineering& Technology, Nagpur, India.
rushilongadge@gmail.com

Abstract:
Data Mining is the process of discovering new patterns from large data sets. Meteorological data mining is a form of data mining which concerned with finding rare patterns inside largely available weather data. To detect rare Weather pattern is difficult challenge because these rare events are characterized by low occurrence and uncertainty. In this paper, we proposed an Adaptive Markov Chain Algorithm Model which uses an open number of states of Markov Chain to accommodate the dynamic temporality of data. The data is collected with the Tropical Atmosphere Ocean (TAO) array which was developed by the international Tropical Ocean Global Atmosphere (TOGA) program. Data Variables including latitude, longitude, zonal wind, meridional wind, humidity, air temperature and sea surface temperature are considered for identifying climate change patterns in this paper.
By adding the Markov property as a global restriction, the granular size of the clusters is determined for optimal performance.
Our climate change pattern detection algorithm is proven to be of potential use for climatic and meteorological research as well as research focusing on temporal trends in weather and the consequent changes.
Keywords:-weather forecasting, data mining, pattern detection, rare event detection, Adaptive Markov Chain model.

I. Introduction

Data mining has developed as one of the major research domain in the recent decades in order to extract hidden and useful knowledge. The extraction of hidden predictive information from large databases, is a powerful new technology with most likely to analyze important

information in data warehouses. Thus, data mining consists of more than collecting and analyzing data, it also includes evaluate and predictions. Anomaly pattern detection, as a data mining task it refers to disclosing patterns that do not conform to expected behaviors in databases [1]. Weather-related data mining is a form of data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved can be transformed into usable knowledge. Weather is one of the meteorological data that is rich by important knowledge [2].

Weather forecasting is a vital application in meteorology and has been one of the most scientifically and technically challenging problems around the world in the last century. This paper focused on various meteorological variables like latitude, longitude, zonal winds, meridional winds, humidity, air temp, sea surface temp. But some other climate variables, such as precipitation and sea level pressure should affect the classification of weather patterns too. Data mining have been employed successfully to build a very important applications in the field of meteorology like predicting rare events like hurricanes, storms and river flood prediction [2][15]. These applications can maintain public safety and benefit.

Climate change is one of the greatest natural resource challenges the world faces. To understand the impacts of climate change on people, landscapes, fisheries, and wildlife, the USGS has conducted research for over hundreds years [2]. This reliable science is needed to help representatives, land managers, and the people make informed and balanced decisions on complex questions about climate change that impacts to the Nation. The proposed budget increase in next year's would support priority research in three programs, which are the National Climate Change and Wildlife Science Center and the Department of the Interior Climate Science Centers, Climate Research and Development, and Biological Carbon Sequestration. In 2013, Interior required bureaus to incorporate climate adaptation into policies, programs, planning, and operations [2]. To accomplish this, improved understanding is needed on what assets are most exposed to climate change, ultimately helping bureaus prioritize their management needs.

Paper primarily concentrates on some aspects increasing the efficiency of data by observing previous data [1]. The efficiency is ensured by the incrementality of the modeling processes. By adding the Markov property as a universal restriction, the granular size of the clusters is determined for optimal performance. The global modeling result is presented by a synopsis, which provides a base of data mining tasks. A novel method is presented for updating the synopsis to reflect current behavior and detecting the developing trends of time evolving data streams.

As mentioned previously, many research activities have been carried out to improve the understanding of climate change patterns by means of different techniques and made considerable contributions [7]-[8]. However, the introduction of data mining techniques into this research field has been limited. Therefore, this paper aims to develop a detection method based on data mining techniques for detecting and classifying weather patterns through a case study on weather data. Actually weather forecasting was and is; a big arrangement of time, funds, talent and absolute use of technology. It involves computing complex mathematical calculations that was not entertained previously by the old age computer. As a result the models were made simple and provide some valuable information. Today's weather forecasting models are more accurate as compared to newly advanced technology. For the achievement of accuracy in the forecasted result is the model. The complexity of the model become more and more complex due to the addition of more factors that inclined the weather i.e. Temperature, humidity, wind, rain etc. [9].

In any data mining model, the raw data is the first required input. In this part, we will collect and build a dataset about climate related attributes of ocean over a history of years. The dataset may include attributes related to wind, temperature, etc. that will be gathered by Tropical atmospheric ocean domain experts. In the second stage and in order to build a climate and weather classifier, we will train the model through using actual data. After training and evaluating the model, it will be used for detecting weather. Climate modeling may include studying the following attributes: Historical weather records, daily rainfall and max and min temperatures, raining and temperature parameters relative to time, location, and height (e.g. spatiotemporal rainfall Distribution, etc), relative humidity, soil moisture, air temperature, soil temperature, etc. In order to apply data mining techniques on climate forecasting, several preprocessing techniques should be utilized to improve accuracy and eliminate outliers. Some of the included steps may include in this :a) Data scaling is a very important step before the models can be formulated and developed. All the input variables were standardized by subtracting mean value and divided by the standard deviation [10]. This would generate a set of standard normal random variables with mean '0' and standard deviation '1'.b) Preparation of inputs for verification of models. c) Some available data mining techniques may be applied depending on suitability and accuracy. Examples include Neural Networks, KNN, Genetic programming, etc. Many numerical techniques of stream flow prediction have been widely used in water resources managements. There is flow release models of climatic variations designed as physically-based models [11].

II. Literature Survey

Many outlier detection algorithms have been proposed [17]. These algorithms can be classified into multi-dimensional space based methods and graph based methods. Multi-dimensional outlier detection methods use distance, depth, or density functions in the multi-dimensional space to check if a point is different from the majority of the data.

There are some approaches to detect rare events have been based on the classification schemes created by examining past normal behavior. So the analysis of such various methods has been discussed. Yu Meng, Margaret H. Dunham, Marchetti and J. Huang et all says that the use of EMMs for unsupervised rare event detection in a spatiotemporal environment is high, adopting representative granules in the data space as states in a dynamic Markov chain. It increases the pattern identification rate and the rare patterns can also be identified. But the algorithm does not produce high recognition rate in all the data [3]. Prasanta Gogoi, D. K. Bhattacharyya, B Borah et all presented a comprehensive survey of well-known distance-based, density-based and other techniques for outlier detection and compare them. Distance based methods for outlier detection is based on the calculation of distances among objects in the data with clear geometric interpretation. Development of an effective outlier detection technique is of mixed-type. Evolving network traffic data, especially in the presence of noise, is a challenging task [5]. Yang Zhang, Nirvana Meratnia, and Paul Having a et al introduces the key characteristics and brief description of current outlier detection techniques using the proposed taxonomy framework and provide an evaluation for each technique [4]. Outlier detection techniques are required to maintain a high detection rate while keeping the false alarm rate low. The detection rate represents the percentage of anomalous data that are correctly considered as outliers, and the false alarm rate, also known as false positive rate (FPR), represents the percentage of normal data that are incorrectly considered as outliers.

In the previous work, the outliers are usually facts. In the real world, there are many cases in which the outliers display in other spatial forms such as line or region. Such outliers exist in weather and climate data. For these two dimensional outlier detection, the determination of the region and their neighbors would be essential. The features should be rather similar, for the points bounded in a region, while for the outside points surrounding the region; the feature would be distinctly different. We cannot detect the points within outlier region as outliers using traditional point outlier detection.

III. Proposed Method

The proposed method is an Adaptive Markov chain algorithm. It is a discrete-time process for which the future behavior, given the past and the present, only depends on the present and not on the past. The centroid of the cluster is adaptive to the data that belongs to the cluster. This is the reason why we name it adaptive Markov chain model.

An Adaptive Markov model:
We proposed an adaptive Markov chain algorithm. Markov chains are an especially powerful and widely used tool for analyzing a variety of stochastic (probabilistic) systems over time.

A Markov chain, studied at the discrete time points 0, 1, 2, ... , is characterized by a set of states' S and the transition probabilities Pij between the states. Here, Pij is the probability that the Markov chain is at the next time point in state j, given that it is at the present time point at state i. The matrix P with elements Pij is called the transition probability matrix of the Markov chain. The definition of the Pij implies that the row sums of P are equal to 1. Under the conditions that all states of the Markov chain communicate with each other (i.e., it is possible to go from each state, possibly in more than one step, to every other state), the Markov chain is not periodic, the Markov chain does not point away to infinity, the probability Pi (n) that the system is in state i at time point n converges to a limit πi as n tends to infinity. These limiting probabilities, or equilibrium probabilities, can be computed from a set of so-called balance equations. The balance equations balance the probability of leaving and entering a state in equilibrium. This leads to the equations

$$\pi_i \sum_{j \neq i} p_{ij} = \sum_{j \neq i} \pi_j p_{ji}, \qquad i \in S$$

Or

$$\pi_i = \sum_{j \in S} \pi_j p_{ji}, \qquad i \in S.$$

In vector-matrix notation this becomes, with π the row vector with elements πi,
π=πP. (1)

Together with the normalization

$$\sum_{i \in S} \pi_i = 1,$$

The solution of the set of equations (1) is unique.

An Adaptive Markov model is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. An adaptive Markov chain pattern detection (AMCPD) method has a flexible structure because it freely allowing the pattern grows and the transition probabilities of the Markov chain are adjusted adaptively. So it is named as an Adaptive Markov Model, it can be considered the simplest dynamic Bayesian network. As shown in Fig. 1, a node in Markov chain represents a weather pattern [1]. In fact a node also represents a group or a cluster of weather data which belong to the same weather pattern. For instance, the first input weather data, which is daily weather information of a particular day, is placed as a center of a cluster.

The main description of the proposed adaptive Markov chain model data mining method can be described by using the following parameters:
a. X, input data which is a set of observation symbols and in this paper represents daily summary data {xk(1), xk(2), ...,x(n)}(n =7 in this paper), and each X is associated with at least one state.
b. S, a set of states and each has three attributes < *i, mik , Cik* >
c. P, an N × N matrix represents state transition probabilities. Each element, p, is the probability of a transition from state Si to Sj.

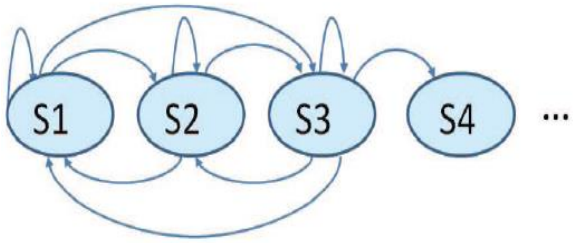

Fig 1: Markov Chain Model

The main process of the proposed adaptive Markov chain model for data mining method is as follows:

1) Initialize the system by creating an initial state, *Sφ*. This state provides a starting point for model construction. No transitions are ever made to *Sφ*. After initialization, *X*, and *P* are empty.
2) Designate the input *Xk* to the system and the current state as *SC.*
3) Calculate the similarity of *Xk* to the existing clusters put it in a cluster according to the similarity obtained. Readjust all of the transition probabilities and update the whole system. If *Xk* is not similar to any existing cluster, a new state will be created, expand matrix *P* to account for the new states, and calculate the new transition probabilities, and the system will be updated.

4) Go to step 2, repeating until there are no more input data.

When more than one input are placed into this cluster, the cluster will include more weather data, and the centroid of the cluster will change over time. In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an Adaptive Markov Model gives some information about the sequence of states.

K-means clustering:

K means method is one of the most popular and mostly used clustering techniques. The K means algorithm is very simple because the idea behind that certain partition of the data in K clusters. The centers of the cluster can be computed as the mean of the all sample belonging to a cluster. The center of the cluster can be considered as the representative of the cluster. The center is quite close to all samples in the cluster.

The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the clustering algorithm works is given as below:

1) The algorithm randomly selects k points as the initial cluster centers ("means").
2) Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3) Each cluster center is recomputed as the average of the points in that cluster.
4) Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

IV. Experimental Setup

1) Data collection:

The data is collected with the Tropical Atmosphere Ocean (TAO) array which was developed by the international Tropical Ocean Global Atmosphere (TOGA) program. The TAO array consists of nearly 70 moored buoys spanning the equatorial Pacific, measuring oceanographic and surface meteorological variables critical for improved detection, understanding and prediction of seasonal-to-interannual climate variations originating in the tropics, most notably those related to the El Nino/Southern Oscillation (ENSO) cycles. Each mooring measures air temperature, relative humidity, surface winds and sea surface temperatures down to a depth of 500 meters.

The data consists of the following variables: date, latitude, longitude, zonal winds (west<0, east>0), meridional winds (south<0, north>0), relative humidity, air temperature, sea surface temperature and subsurface temperatures down to a depth of 500 meters. Data has taken from the buoys from as early as 1980 for some locations.

Variable Characteristics:

The latitude and longitude in the data showed that the buoys moved around to different locations. The latitude values continued within a degree from the approximate location. However the longitude values were sometimes as far as five degrees off of the approximate location.
Looking at the wind data, both the zonal and meridional winds fluctuated between -10 m/s and 10 m/s. The plot of the two wind variables showed no linear relationship. Also, the plots of each wind variable against the other three meteorological data showed no linear relationships. The relative humidity values in the tropical Pacific were typically between 70% and 90%. Both the air temperature and the sea surface temperature fluctuated between 20 and 30 degrees Celsius. The plot of the two temperatures variables shows a positive linear relationship existing. The two temperatures when each plotted against time also have similar plot designs. Plots of the other meteorological variables against the temperature variables showed no linear relationship.

2) Preprocessing Techniques:

As mentioned in data information that there are some missing values because not all buys are able to measure measure current rainfall and solar radiation so these values are missing dependent on the individual buoy. So for filling those missing and reducing the noisy data we used normalization technique. The process of preprocessing has many steps, but can be summarized as the extraction, transformation and loading of the data. Min-Max Normalization transforms a value one to another which fits in the given range. It is given by the formula below (Figure 2):

$$V` = V\text{-}MinA / MaxA – MinA$$

Figure 2: Min-Max Normalization Formula

Where A stands for attribute value, V for Original value and V` for normalized value. The purpose of using normalization technique is that it reduces the number of values for a given continuous attribute by partitioning the range of the attribute into intervals. Interval labels replace actual attribute values. Euclidian distance and how it's can be used to evaluate similarities between samples of data is given by the following formula of Euclidean distance.

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Fig. 3**:** Euclidean distance formula

A simple data preprocessing technique could improve the effectiveness of analysis in orders of amount. So Euclidian distance is used to evaluate similarity between sample of data. It improves the effectiveness and the performance of the mining algorithms.

V. Result Analysis

The data is made train and apply preprocessing technique to remove the missing values. The preprocessed data is loaded successfully and it is shown in the following module (Fig.4). Now number of many missing values present in data are evaluated and generated in report format module (Fig.4).

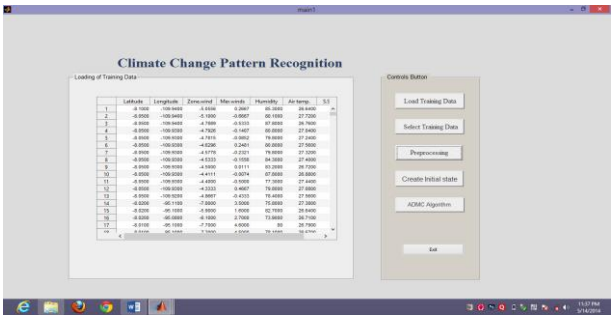

Fig.4: Training data load Module

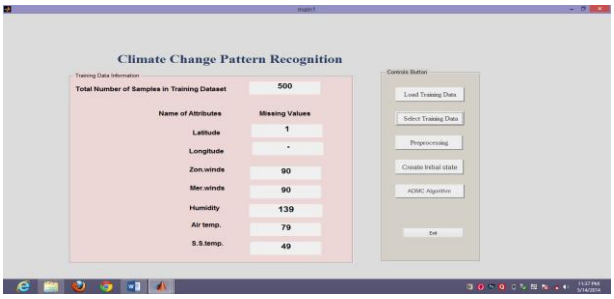

Fig.5: Missing values Report Module

The data that we take contain total 500 values but first 300 values use for clustering purpose and remaining values used as a new input data. The proposed algorithm Adaptive Markov Chain first calculates the cluster similarity with new data and then asssing new data values to cluster one. Algorithm updated the centroid of cluster and after this it again asssing new data values to cluster two and updated centroid of cluster two. This process continues, till the last one centroid is not updated. With the help of this the result generated and shows patterns are detected.

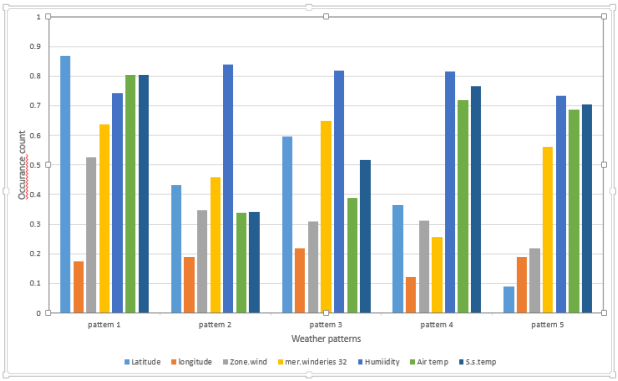

Fig. 6: Result Screen for Patterns detected

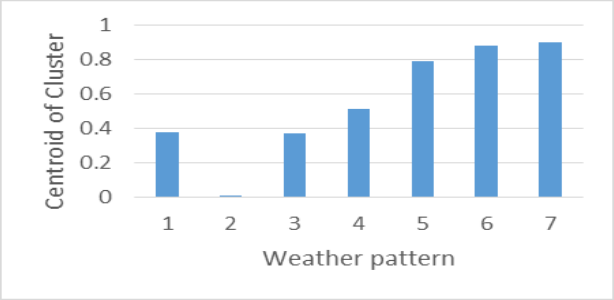

Fig. 7: Centroid of cluster

VI. Conclusion

By the result of proposed algorithm an Adaptive Markov chain Algorithm is able to detect patterns and their rare outlying changes. The ability to forecast the weather accurately is an increasingly important part of our economy and our society. To identify the patterns, various algorithm has been employed but we are doing considerably better at detecting it. Adaptive Markov chain Algorithm has a flexible structure because it allows to pattern grow and the transition probabilities of the Markov chain are adjusted adaptively. It improved the performance by detecting a rare events.

More weather data from different weather stations can be analyzed in future studies for further investigation on climate patterns.

References

[1]Zhaoxia WANG, Gary LEE, Hoong Maeng CHAN, Reuben LI, Xiuju FU and Rick GOH Pauline AW Poh Kimand Martin L. HIBBERD Hoong Chor CHIN "Disclose climate change patterns using an Adaptive Markov chain pattern detection method", DOI 10.1109/SOCIETY.2013.15, 2013 IEEE.

[2]UNEP "United Nations Environment Programme, Climate Change", Available from: http://www. Unep. Org climate change; cited 8 March2013.

[3] Y. Meng, M. Dunham, F. Marchetti, and J. Huang, "Rare event detection in a spatiotemporal environment", Proceedings of the Second IEEE International Conference on Granular Computing (GrC'06), pp. 10–12, 2006.

[4] Z. Yang, N. Meratnia , and P. Havinga , "Outlier Detection Techniques for Wireless Sensor Networks: A Survey," IEEE Communications Surveys& Tutorials, vol. 12, no. 2, pp. 159–170, 2010.

[5] P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita , "A Survey of Outlier Detection Methods in Network Anomaly Identification," The Computer Journal, vol. 54, no. 4, pp. 570–588, Mar. 2011.

[6] Z. Wang, C. S. Chang, and Y. Zhang, "A feature based frequency domain analysis algorithm for fault detection of induction motors,"Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on, pp. 27–32 , 2011.

[7]Ayham Omary, Ahmad Wedyan, Ahmed Zghoul, Ahmad Banihani, and Izzat Alsmadi, "An Interactive Predictive System for Weather Forecasting",978-1-4673-1550-0©2012 IEEE.

[8] Bartok J., Habala O., Bednar P., Gazak M., and Hluch L., "Data mining and integration for predicting significant meteorological phenomena," Procedia Computer Science, p.37 – 46. 2010.

[9] Saima H. , J. Jaafar, S. Belhaouari, T.A. Jillani", Intelligent Methods for Weather Forecasting: A Review" ,©2011 IEEE.

[10] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys (CSUR), vol. 41, no. 3, p. 15, 2009.

[11] R. M. Li, Statistical Analysis of the Spatio Temp oral Variability of the Urban Heat Island in Singapore, Honours Thesis. Department of Geo grap hy, National University of Singapore. , 2009.

[12] W. Chow and M. Roth, "Temporal dynamics of the urban heat island of Singapore," International Journal of Climatology, vol. 26, no. 15,pp. 2243–2260, 2006.

[13] K. L. Ebi, N. D. Lewis, and C. Corvalan, "Climate Variability and Change and their Potential Health Effects in Small Island States: Information for Adaptation Planning in the Health Sector," Environmental Health Perspectives, vol. 114, no. 12, pp. 1957–1963, 2006.

[14] NEA, "National Environmental Agency, Weather Wise Singapore" , Available from: http: //www.app2. NEA.gov.sg/data/cms resource/2;[cited29 July 2011].

[15] NOAA, "El Nino and La Nina, Climate Prediction Center", Available from: http: // www. cpc.ncep.noaa.gov /products; [cited 5 July 2011].

[16] J. Harger, "Air-temperature variations and ENSO effects in Indonesia, the Philippines and El Salvador. ENSO patterns and changes from18661993" Atmospheric Environment, vol. 29, no. 16, pp. 1919–1942, Aug. 1995.

[17] M. Cui, J. Mo, and F. Qiao , "Einino phenomenon and extended associate pattern analysis," Journal of Hydrodynamics Series B-English Edition, vol. 16, no. 1, pp. 90–100, 2004.

[18] Duffy, J.J. and Franklin, M.A. (1975) 'A learning identification algorithm and its application to an environmental system', *IEEE Transactions on Systems, Man and Cybernetics*, Vol. SMC-5, No. 2, pp. 226-240.

[19] Bartok J, Habala O, Bednar P., Gazak M., and Hluch L, "Data mining and integration for predicting significant meteorological phenomena," Procedia Computer Science, p.37 – 46. 2010.

[20] Mohammadi K, Eslami H. R., Kahawita R. ,"Parameter estimation of an ARMA model for river flow forecasting using goal programming. "Journal of Hydrology, 331, 293–299. 2006.

[21] O. Boiman and M. Irani, “Detecting irregularities in images and in video”. IJCV, 74(1):17–31, 2007.

[22] Adesesan Barnbas Adeyemo, “Application of Data Mining Techniques in Weather Prediction and Climate Change Studies”, DOI: 10.5815/ijieeb. 2012. 01.07.

[23] Jeffrey S. Rosenthal et all “On adaptive Markov chain Monte Carlo algorithms”. Bernoulli 11(5), 2005, 815–828.

