www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 12 Dec. 2016, Page No. 19784-19788

A Review of Weather Forecasting Using Data Mining Techniques

Ms.P.Shivaranjani[1], Dr.K.Karthikeyan[2]

1) Ms.P.Shivaranjani, Research Scholar,
Dr.SNS Rajalaksmi College of Arts & Science, Coimbatore.
Emai-Id: shiva230482suresh@gmail.com
2) Dr.K.Karthikeyan Assisstant Professor,
Department of Computer Science,
Government Arts & Science College, Karambakudi.
Email-Id : ithodsns@gmail.com

Abstract :

The vast dramatically changes occurs day by day in certain fields due to the development of advanced technology and nature one such among them is rainfall. The rainfall is the fragment of the agriculture and unable to understand the monsoon condition, predicating the crop yield and the soil fertility. Data mining is the techniques used to extract the knowledge from the set of data.This paper provides a survey of different data mining techniques being used in weather prediction or forecasting which helps the farmer for yield worthy productive and nourish the soil fertility such as artificial feed-forward neural networks (ANNs), fuzzy inference system, decision tree method, time series analysis, learning vector Quantization (LVQ) and biclustering technique.

Keywords: Data mining, agriculture, soil Fertility, crop yield, ANNs, FIS, LVQ, biclustering

1.INTRODUCTION:

The backbone of Indian economy is Agriculture. Now a day's weather or rainfall is the stimulating problems around the world. Rainfall prediction is nothing but weather forecasting. Weather forecasting application is an art of science and technology use to the state of atmosphere for a location. The weather forecaster's work 24/7, 365 days of the year, using supercomputers it is easy to predict the weather for hours, days, weeks, seasons and even years ahead.

Weather forecasting is an area of meteorology that is carried by collecting dynamic data related to current state of weather like fog, rainfall, temperature, wind etc. We continually update our knowledge of the current state of the atmosphere by

- Satellites measure radiation from Earth's plane and the impression.
- Balloons and aircraft measure the bit of the air that they passing through.
- Buoys and land stations measure the lower part of the atmosphere.
- Radar systems measure the indication of emission from rain drops and snowflakes

The data collected from various states are to distorted intoa numerical representation of the recent atmospheric conditions. This process is known as assimilation. Small changes in atmospheric conditions lead to very different weather patterns, so it's vital that the current state of the atmosphere is represented as accurately as possible.

The climate variations need to be addressed and an analysis is to be made in order to help the farmers to maximize the crop productivity [1].

2.LITERATURE REVIEW:

Kit Yan Chan [4], presents a comparison of two sub sampling nonparametric methods for designing algorithms to forecast time series from the cumulative monthly rainfall. Both approaches are based on artificial feed-forward neural networks (ANNs).

Jesada, Kok and Chung [5] proposed fuzzy inference system for monthly rainfall prediction in the northeast region of Thailand. The predicted show of the proposed model was compare to be conservative Box-Jenkins and artificial neural networks model. Accordingly, the experimental results show the modular FIS is superior another method to predict accurately. The predicted mechanism can be interpreted through fuzzy rules. Auto-regression, Seasonal auto regressive integrated moving average and ANN modular FIS provide better results. The experimental results give together accurate results and human-understandable prediction mechanism.

Narasimha, Prudhvi and Naidu [6] proposed decision tree method using SLIQ to implement the precipitation model. It is observed that decision tree method achieves closer agreement between actual and estimated rainfall. SLIQ method gives high accuracy rate when compared to other prediction model like fuzzy logic, NN etc. The use of Gini index for rainfall analysis is quite apt because of the irregularities present in the statistical data of precipitation. It gives accuracy of 72.3% and completely based on historical data. The decision tree constructed and the classification rule are generated.

Mark, Bobby, Yung and Beth [7] proposed time series analysis is used as prediction algorithm. Two

components rainfall/evaporation and crop management. Decision support system for Agriculture management using prediction algorithm aimed to develop a system that will determine the trend of rainfall and evaporation using time series analysis as its prediction algorithm, to develop web-based application that displays graphs and tables according to the result of the prediction algorithm, and to utilize a classification of crops that aids farmers as basis for recommendation according to the predicted amount of rainfall per quarter. The system is found useful in terms of efficiency, reliability. It shows interface the quarter of the year labeled Q1, Q2,Q3, Q4, prediction of average amount of rainfall and evaporation, the trends, and the seasonal effects in its provided field in the table.

Jethangir and Onaiza [8] proposed BP and learning vector Quantization (LVQ) is used for monsoon rainfall prediction. 45 years monsoon rainfall data is used to train Neural Network and evaluate the performance of these models over a test period of 5 years from 2005-2009. The results were compared with multiple linear regressions and statistical downscaling models, but the results reveals neural network has better performance in terms of accuracy, and also in terms of greater lead time and fever required resources. LVQ is used for classification. LVQ overcomes the problem that we might face in BP of having output 1 for more than one output neurons. This may raise potential problems. LVQ takes less training time than BP. However, in our case of monsoon rainfall prediction almost a year in advance, training time difference that was in seconds is insignificant.

Dong li, XuShu, Meng and Yang [9] proposed a time series analysis method which is decomposed into trend items, cycle items, respectively extracted by establishment of various forecasting model and statistics method is used to predict the month precipitation in crop growth period in the area of Chahayang from 1956 to 2008, in order to seek the rule of month precipitation change in crop growth period in this area. It provides data to evaluate the efficiency water resource utilization, and provide reliable basis for local department to manage and plan.

Kesheng and Lingzhi [10] presented a novel modular type support vector machine to simulate rainfall prediction. V-SVM regression model,which introduced a new parameter "V" which can control the number of support vectors and training errors without defining ϵ a prior. To be more precise, the author posed that „V" is an upper bound on the fraction of margin errors and lower bound of the fraction of support vectors. First of all, a bagging sampling technique is used to generate unlike training sets. Secondly, changed kernel function of SVM with different parameters, i.e, base models, is then trained to formulate different regression based on the poles apart training sets. Thirdly, the partial least square (PLS) technology is used to select decide the appropriate number of SVR combination member. Finally, a V-SVM can be produced by learning from all base models. V-SVM produced greater forecasting accuracy and improving prediction quality V-SVM is to solve nonlinear regression problems.

S. Kannan and S. Ghosh [11] contributed in the direction of developing methodology for predicting state of rainfall at local or regional scale for a river basin from large scale climatological data. A model based on K- mean clustering technique joined with decision tree algorithm, CART, is used for the generation of rainfall states from large scale atmospheric variables in a river basin. Daily rainfall state is derived from the past daily multi-site rainfall data by using K-mean clustering. Various cluster validity measures are applied to observed rainfall data to get the optimum number of clusters. CART is use to train the data of daily rainfall condition of the river basin for 33 years. The methodology is tested for the Mahanadi River in India. The change usual in the river basin owed to overall warming is set by the comparisons of the number of days falling under different rainfall states for the observed period and the future predicted. CART algorithm proved to be good inpredicting the daily rainfall state in a river basin using statistical downscaling.

Z. Jan *et al. [12*] developed new accurate and sophisticated systems for Seasonal to inter annual climate prediction using data mining technique, K-Nearest Neighbor (KNN). It uses numeric past data to predict the climate of a specific region, city or country months in advance. Dataset uses 10 years of notable data with has 17 attributes, i.e. mean temperature, Max Temp, Min Temp, Wind Speed, Max Wind Speed, Min Wind Speed, Dew Point, Sea Level, Snow Depth, Fog, gust, SST, SLP, etc., with 40000 records for 10 cities. The dataset uses data cleansing to pact with noisy and missing values. It is stored in MS ACCESS format. It can predict a huge set of values at the same time with elevated level of accuracy. The predict result of KNN is easier to understand.

Soo-YeonJi et al. [13] predicted the hourly rainfall in any geographical regions time efficiently. The chance of rain is first determined. Then only if there is any chance of rainfall, the hourly rainfall prediction is performed. Although quite a lot methodology have been introduced to predict hourly prediction, most of them have performance limitations because of the existence of broad range of variation in data and limited amount of data. CART and C4.5 are used to offer outcomes, which may provide hidden and important patterns with transparent reasons. About 18 variables were used from weather station. For justification purpose, 10 fold cross validation method is performed. CART gives slightly better performance than C4.5. Considering the chances, only a small number of instances are left for prediction which makes it hard to predict.

The soil testing laboratories are provided with suitable technical literature on various aspects of soil testing, including testing methods and formulations of fertilizer recommendations [15]. It helps farmers to decide the extent of fertilizer and farm yard manure to be applied at various stages of the growth cycle of the crop.

A Mucherino et al. [16] apply a supervised biclustering technique to a dataset of wine fermentations with the aim of selecting and discovering the type that are responsible for the problematic fermentations and also exploit the selected features for predicting the quality of new fermentations. Taste sensors are used to obtain data from the fermentation process to be classified using ANNs [17]. Similarly, sensors are used to smell milk, which is classified using SVMs [18].

3. DATA MINING IN WEATHER FORECASTING:

DOI: 10.18535/ijecs/v5i12.77

Data Mining deals with what kind ofpatterns can bemined. Based on the kind of data to be mined, there are two kinds of functions involved in Data Mining such as Descriptive model and Predictive model. The Descriptivemodel identifies patterns or relationships in data and dealswith general properties of data in the database. The predictive model is the process of finding a model whichdescribes the data classes or concepts, the purpose being tobe able to use this model to predict the class of objects whose class label is unknown[1].

Data mining techniques are mainly separated into two groups, viz. classification and clustering techniques. Classification techniques are designed for classifying unknown samples using information provided by a set ofclassified samples. This set is usually referred to as atraining set, because, it is generally used, to train theclassification technique i.e. how to perform itsclassification. If a training set is not available, there is nopreviousknowledge about the data to classify.

Clustering technique is used to group the element that is particular area occupied by rainfall regions and the rainfall is predicted in a particular region. The different classification techniques for discovering knowledge are Rule Based Classifiers, Artificial Neural Network(ANN),Bayesian Networks (BN), Decision Tree (DT), Nearest Neighbour(NN), Rough Sets, Fuzzy Logic, Support Vector Machine (SVM),Genetic Algorithms.[2]

The different clustering methods are Hierarchical Methods(HM), Partitioning Methods (PM), Density-based Methods(DBM), Grid-based Methods , Model-based Clustering Methods(MBCM) and Soft-computing Methods [fuzzy, neural network based], Squared Error—Based Clustering (Vector Quantization), network data and Clustering graph [3]

***3.1 k*-nearest neighbor :**

The *k*-nearest neighbor (*k*-NN) method is one of the data mining techniques considered to be among the top 10 techniques for data mining [19].It tries to classify an unknown model based on the known classification of its neighbors. Suppose that a set of samples with known classification is available, the so-called *training set*. Intuitively, each model should be classified likewise to its close samples.

Therefore, if the classification of a example is unknown, then it could be predicted by considering the classification of its nearest neighbor samples. Given an strange sample and a training set, all the distances between the unknown sample and all the sample in the guidance set can be computed.

The distance with the smallest value corresponds to the sample in the training set closest to the unknown sample. Therefore, the unknown sample may be classified based on the classification of this nearest neighbor.

3.2 Artificial Neural Networks:

ANNs can be used as data mining techniques for classification. They are inspired by biological systems, and particularly by research on the human brain. ANNs are developed and planned in such a way that they are able to study and take a broad view from data and experience.

In general, ANNs are used for modelling functions having an unknown mathematical expression. The multilayer perceptron has the neurons organized in layers, one input layer, single or multiple unseen layers and one output layer. In some applications there are only one or just two hidden layers, but it is more convenient to have more than two layers in some other applications.

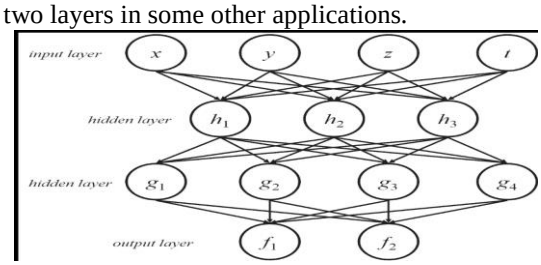

Figure 3.1 Example of a multilayer perceptron.

The input data are provided to the network through the input layer, which sends this information to the hidden layers. The data are processed by the hidden layers and the output layer. Each neuron receives output signals from the neurons in the previous layer and sends its output to the neurons in the successive layer. The final layer, the output one, receives the inputs from the neurons in the last hidden layer, and its neurons provide the output values.

3.3 Decision trees:

It is commonly used in data mining to examine the data and to induce the tree and its rules that will be use to make predictions. A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification And Regression Trees), Quest, and C5.0 [20]. A decision tree is a tree in which each division node represents a choice between a number of alternatives, and every leaf node represents a decision.

Depending on the algorithm, each node may have two or extra branches. For example, CART generate trees with only two branches at each node. Such a tree is called a binary tree. When

more than two branches are allowed this is called a multiway tree [21].

3.4 Support vector machines (SVMs) :

It is supervised learning methods used for classification [22,23, 23]. In their basic form, SVMs are used for classifying sets of samples into two displace classes, which are separated by a hyperplane defined in a suitable space. Note that, as consequence, a single SVM can only discriminate between two unlike classifications. However, as we will discuss later, there are strategies that allow one to extend SVMs for classification problems with more than two classes [23, 24]. The hyperplane used for separating the two classes can be defined on the basis of the information contained in a training set.

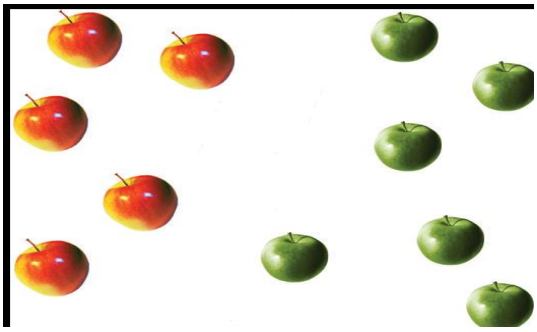

Fig. 3.2 Apples with a short or long stem on a Cartesian system.

Let us suppose that a general rule for classifying these apples is needed, i.e., a classifier is wanted that is able to decide if a given apple has a short or a long stem. A classifier could simply follow the rule: the apple has a short stem if it is in an area defined by the apples having a short stem, and it has instead a long stem if it is in the area defined by the apples having a long stem. Apples with a known classification can be used for defining the two areas of the Cartesian system related to these two different types of apples. Such apples define the training set, which can be used for learning how to classify apples whose length of the stem is unknown.

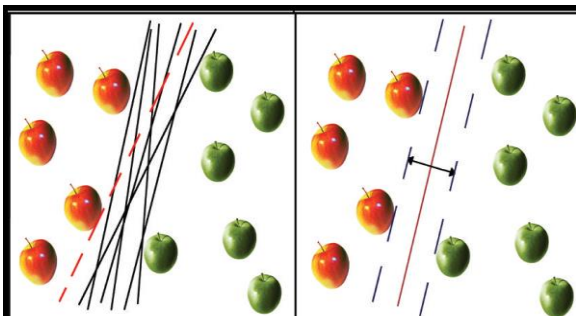

Fig. 3.3 (a) Examples of linear classifiers for the apples; (b) the classifier obtained by applying a SVM.

Once one of these lines has been defined, the classifier can work as follows. If an unknown apple is found to be in the area defined by the apples having a short stem, then it is considered to have a short stem, otherwise it has a long stem. Note that each line drawn in Figure 3.3 (a) classifies the apples of the training set correctly.

4. CONCLUSION:

From the various papers reviewed the supervised and unsupervised machine learning algorithms can be used to perform the weather prediction and yield of crop can be increased by using different data mining techniques can used for prediction of rainfall for daily ,monthly and yearly with various parameter and thus it provides better result.

5. REFERENCE:

[1] D Ramesh, B Vishnu Vardhan, "Crop Yield Prediction Using Weight BasedClustering Technique ", IJCEA, 2015.

[2]. Beniwal, S. &Arora, J. (2012), "Classification and feature selection techniques in data mining", International Journal of Engineering Research & Technology (IJERT), 1(6).

[3]Xu, R &Wunsch, D (2005), "Survey of clustering algorithms", Neural Networks, IEEE Transactions on, 16(3), 645-678.

[4] Kit Yan Chan,'' Neural-Network-Based Models for Short-Term Traffic Flow Forecasting Using a Hybrid Exponential Smoothing and Leven berg–Marquardt Algorithm" , IEEE trans on intelligent transportation system, VOL. 13, NO. 2, pp.644-646,JUNE 2012.

[5] International Conference on 31, 163-167, doi: 10.1109/ WISM, 2013.

[6] Narasimha Prasad, Prudhvi Kumar and Naidu MM, "An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree", 4th International Conference on Intelligent Systems, Modeling and Simulation, 2013.

[7] Mark Ian Animas, Yung-CheolByun, Ma.Beth Concepcion and Bobby D. Gerardo, "Decision Support System for Agricultural Management Using Prediction Algorithm", 2013.

[8] Jehangir Ashraf Awan and OnaizaMaqbool, "Application of Artificial Neural Networks for Monsoon Rainfall Prediction", Sixth International Conference on Emerging Technologies, 2010.

[9] Dong Li-li, XuShu-qin, Meng Fan-Xiang and Yang Xu, "Application of Time –series Model in the Chahayang Farm of Rainfall Prediction", 2010.

[10] Kesheng Lu and Lingzhi Wang, "A Novel Nonlinear Combination Model Based on Support Vector Machine for Rainfall Prediction", Fourth International Joint Conference on Computational Sciences and Optimization, 2011.

[11] S. Kannan , SubimalGhosh, "Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output", Springer-Verlag, July- 2010.

[12] Sarah N. Kohail, Alaa M. El-Halees, "Implementation of Data Mining Techniques for Meteorological Data Analysis", IJICT Journal Volume 1 No. 3, 2011.

[13] Simon S. Haykin, "Neural Networks: A Comprehensive Foundation", Second Edition, Prentice Hall International, 1999.

[14] Due R. A., "A Statistical Approach to Neural Networks for Pattern Recognition", 8th edition. New York: John Wiley and Sons publication, 2007.

[15] "Soil test", Wikipedia, February 2012

[16] A. Mucherino, A. Urtubia, Feature Selection for Datasets of Wine Fermentations, I3M Conference Proceedings, 10th International Conference on Modeling and Applied Simulation (MAS11), Rome, Italy, September 2011.

[17] Riul A Jr, de Sousa HC, Malmegrim RR, dos Santos DS Jr, Carvalho ACPLF, Fonseca FJ, Oliveira Jr ON, Mattoso LHC

Wine classification by taste sensors made from ultra-thin films and using Neural Networks. Sens Actuators B98:77–82, 2004.

[18] Brudzewski K, Osowski S, Markiewicz T Classification of milk by means of an electronic nose and SVM neural network. Sens Actuators B98:291–298, 2004.

[19] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10

Algorithms in Data Mining, Knowledge and Information Systems 14, 1–37, 2008.

[20] Introduction to Data Mining and Knowledge Discovery, Third Edition, Two Crowds Corporation, http://www.twocrows.com/intro-dm.pdf.

[21] Data mining Models and Algorithms, http://www.huaat.com/english/datamining/D_App.html.

[22]. C.J.C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*, Data Mining and Knowledge Discovery 2 (2), 955–974, 1998.

[23] C. Cortes and V. Vapnik, *Support Vector Networks*, Machine Learning **20**, 273–297, 1995.

[23] V.N. Vapnik, *Statistical Learning Theory*, JohnWiley & Sons, 1998.

[24] I. Steinwart, *Consistency of Support Vector Machines and Other Regularized Kernel Classifiers*, IEEE Transactions on Information Theory **51**, 128–142, 2005.