ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

Weather Prediction Using J48, EM And K-Means Clustering Algorithms

P. Kalaiselvi, D. Geetha

M.Phil Scholar, Dept. of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, India

Head of the Department, Dept. of Computer Application [UG], Sree Saraswathi Thyagaraja College, Pollachi, India

ABSRACT: Data mining is the computer assisted process of digging through and analyzing immense sets of data and then extracting the relevant data. Data mining tools predicts behaviors and future mode, allowing businesses to make proper and good decisions. It can answer questions that traditionally were very time strong to resolve. Therefore they can be used to predict meteorological data. That is called as weather prediction. Weather forecasting is an important application in meteorology and has been one of the most scientifically and technologically challenging problems around the world. Predicting the weather is essential to help preparing for the best and the worst climate. We need to be on alert to the adverse weather conditions by adapting some precautions and using prediction mechanisms for early warning of hazardous weather phenomena. Many weather predictions like rainfall prediction, thunderstorm prediction, predicting cloud conditions are major challenges for atmospheric research. This dissertation uses the Data Mining techniques for weather predictions and studies the benefit of using it. Decision tree J48, EM(Expectation Maximization) and k-means clustering algorithm has been used in this research work to identify the variation in the weather conditions in terms of Temperature, Sunny, Rainfall, Overcast and Wind Fall.

KEYWORDS: Decision J48; KM ; K-Means Clustering.

I.INTRODUCTION

Data mining is the process of collecting, searching through, and analyzing an immense amount of data in a database, as to discover patterns or relationships. It is a powerful new technology with great power to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and allowing business to make proactive, knowledge-driven decisions[1]. Data mining is a term from computer science. Sometimes it is also called knowledge discovery in databases (KDD).

Data mining is about finding relevant information in a great amount of data. The information obtained from data mining is hopefully both new and useful[11]. It help users to analyze data from many different dimensions or angles, categorize it, and integrate the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in immense relational databases.

Data Mining techniques are used in Weather prediction process[10]. Weather is most effective environmental constraints in every phase of our life. Weather forecasting is used in many fields like Agriculture **,** Food security disasters and science . In ancient years we don't know weather conditions. So, we faced many problems on industry and agriculture field and food management process. But, now we have many ways to find weather conditions. That is the main reason for applying data mining techniques to find the weather conditions.

II. RELEATED WORK

Arti R. Naik and S.K. Pathan reviewed "Weather classification and forecasting using feed forward neural network"[1] , It is used Artificial neural networks in weather prediction and classification in the forecasting, predicted values are obtained simply by observation. After the numerical methods were developed, the images from satellite were used to retrieve data. The weather data is updated for every 24 hours and records the values till next 5 days. In this paper weather forecasting is done by using Back-propagation Feed forward Neural Network. The authors collected the data by using wireless sensors like anemometer sensor, thermo hydro sensor during 2012. The noise in the data is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

removed in pre-processing stage and in the back propagation algorithm the error is back-propagated and respectively the weights are modified so as to reduce the error.

Prasanth Rao jillella.s.s reviewed "Artificial Neural Networks for weather prediction"[2], It's inspired by biological Nervous systems. It's mainly based on two ways: Exploration techniques and ANN. Another way, K-mean clustering Groups the data into single clusters. In the periods [2001 – 2005]. ANN Produced a result reliable and the result are reached 95% from dataset. But, it's limited by unknown condition.

M.viswambari reviewed "Data mining techniques to weather prediction"[3], Classification and back propagation. Backward propagation of errors applied to this prediction. It's a common method of artificial neural networks and the network learns from many inputs and desired output. To have a better result it focused on observed parameters information only.

Rajesh Kumar reviewed "Decision Tree For The Weather Forecasting"[4], The application based on Decision trees and machine learning algorithm are applied to this prediction. The measurements are fog, rain, thunder, humidity and pressure. Classification methods, decision tree based methods, rule based, memory based reasoning, neural network and support vector machine methods applied in this process. In this prediction, main goal is to create a model that's going to predict the value of target parameters based on input parameters and huge amount of data can be used to predict.

Elia Georgiana reviewed "Decision Tree Based Weather Prediction"[5], It's based on a decision support tool and CART decision tree algorithm is used for prediction. CART- It effort with nominal variables and applied with weka tool. This prediction analyzes meteorological data registered during the last years [2002-2005]. CART classification tree is used to predict the measurements year, month average pressure, clouds quantity, relative humidity, precipitations and average temperature values in Hong Kong. It produced in a good accuracy result in prediction.

M. A. Kalyankar and S. J. Alaspurkar reviewed " Data Mining Technique to Analyse the Metrological Data"[6] , It is used data mining techniques to gain weather data and find the hidden patterns inside the immense dataset so as to transfer the retrieved information into usable knowledge for classification and prediction of weather condition. Data mining process is applied to extract knowledge from Gaza city weather dataset. This knowledge helps to obtain useful predictions and support the decision making process. Dynamic data mining methods are required to build, that can learn dynamically to match the nature of rapidly changeable weather nature and sudden events.

III.EXISTING SYSTEM

The existing system focused on applications of data mining in weather forecasting. Weather forecasting has been one of the most critical problems around the world because it consists of multidimensional and nonlinear data. Recently, climate changes causes much trouble in rainfall forecasting. This process applying five years previous data from Jan 2010 -Jan 2014 for Nagpur station[12]. Generally this algorithms are used for the prediction on only available datasets by applying the Frequent Pattern Growth Algorithm for deleting the incorrect data. Generally temperature, humidity, wind speed are mainly important for the rainfall prediction on the percentage of these parameters , rainfall can be predicted. The applications that use the FPG , utilized only the basic algorithm, while many other improvements are available.

IV.PROPOSED SYSTEM

The proposed system focused on weather forecasting entails predicting how the present state of the atmosphere will change. Weather forecasting has been one of the most scientifically and technologically critical problems around the world in the last century. To make an correct prediction is one of the major challenges facing meteorologist all over the world. The proposed work is used to find forecasting Temperature, Rainfall and Overcast and wind speed. The data is used from wounder ground weather website between 2015 and 2016.

The methodology involves Preprocessing, Classifying, Clustering, Associative, Attribute selection and finally visualization. In classification, Applying Decision tree algorithm for classifying weather parameters such as temperature, rainfall, humidity and wind speed. In Clustering algorithms, Applying EM (Expectation Maximization) and K-means clustering to group the overall average on weather conditions like sunny, rainy, overcast and overall attributes maximum and minimum values based on cendroid. Finally, the results displayed in visualize form and the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

results show how these parameters have impact the weather observed in these months over the study period. The results obtained were evaluated with the test data set prepared along with the training data.

V.WEKA TOOL

WEKA - Waikato Environment for Knowledge Analysis[7] **.** It is a machine learning algorithms for data mining tasks .The algorithms can either be applied directly to a dataset or called from your own java code. weka contains tools for data Pre-processing, Classification, Clustering, Association rule and Visualization. It also well suited for developing new machine learning schemes. It is also well-suited for developing new machine learning schemes. Each entry in a dataset is an instance of the java class and each instance consists of a number of attributes.

VI. ALGORITHMS IMPLEMENTATION

A. DECISION TREE J48 ALGORITHM:

It is similar to the tree structure having root node, intermediate nodes and leaf node. Each node in the tree contain a decision and that decision leads to our result. Decision tree is a decision support tool that helps a tree-like graph or model of decisions and their possible consequences, including chance event results, resource costs, and utility. It is one way to display an algorithm. [7] The decision tree J48 is the implementation of algorithm ID3 (Iterative Dichomiser 3) developed by the WEKA project team. By applying a decision tree like J48 Algorithm on dataset would allow to predict the target variables of a new dataset record.

It is like a tree structure. The goal is to create a model that predicts the value of a target parameter based on many input parameter[9]. A tree can be made to learn by splitting the Source data set into subsets based on an attribute value test.

Following steps in J48 Algorithm process:

Step1: Divide the input space of a weather dataset into manually exclusive areas.

Step 2: Assign the label of having an attributes.

Step 3: Describe the value of each data points.

Step 4: Applying the rule.

R1 : IF (Outlook = Sunny) AND
(Windy = FALSE) THEN Play = Yes
R2 : IF(Outlook = Sunny) AND
(Windy = TRUE) THEN Play = No
R3 : IF(Outlook = Overcast) THEN
Play = Yes
R4: IF (Outlook = Rainy) AND
(Humidity = High) THEN Play = No

R5 : IF (Outlook = Rainy) AND
(humidity = Normal) THEN
Play = Yes

A decision node outlook has two branches like Sunny, Overcast , Rainy and the leaf node Play.

Step 5: Splitting criterion to calculate the attributes is the best to split that portion of the training data that reaches a particular node. This algorithm is used to classify the weather attributes like Sunny, Rainy and Overcast.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

B . EM (EXPECTATION MAXIMIZATION) ALGORITHM:

An Expectation Maximization is an iterative method for finding maximum likelihood or Maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables[7]. There are two steps in this EM Algorithm process:
There are two step mainly involved in EM Algorithm process:
Step 1 : EM iteration alternates between performing an expectation step, whichmades a function for the expectation of the log-likelihood assess using the current estimate for the parameters. Then, calculate the expected value of the log likelihood function.

$Q(\Theta \mid \Theta^{(t)}) = E_{z \mid x, \Theta^{(t)}} [\text{Log } L\ 9(\Theta; X, Z)]$

Here ,
Z = Sequence of values or Sum
Log L = Log likelihood
Θ = Unknown Parameters
X = Observed Data ($\Theta^{(t)}$)
Step 2: A maximization step, which computes parameters maximizing , the expected log-likelihood found on the E-step. These parameters estimates are than used to determine the distribution of the latent variables in the next E step. Find the parameters that maximizes this quantity:

$$\Theta^{(t+1)} = \arg \max_{\Theta} Q(\Theta \mid \Theta^{(t)})$$

By Applying these two steps for finding maximization values of weather attributes like sunny, rainy and overall temperature and humidity mean and standard deviation and windy and play.

C . K-MEANS CLUSTERING:

K-means clustering is a data mining algorithm used to cluster observations into groups of relevant observations without any prior knowledge of those relationships[8]. K-means is the most popularly used algorithm for clustering. User need to specify the number of clusters (k) in advance. Algorithm randomly selects k objects as cluster mean or center and K-means Basic version works with numeric data only. It is a prototype based clustering technique defining the prototype in terms of a centroid which is considered to be the mean of a group of points and is applicable to objects in a continuous n-dimensional space.

It is a method of vector quantization, really from signal processing, that is famous for cluster analysis in data mining and Cluster data using the k means algorithm. It Can use either the Euclidean distance . If the Euclidean distance is used, then centroid are computed as the component-wise median rather than mean.:
The K-Means clustering algorithm is a partition-based cluster analysis method. Following steps in this process:

Step 1:
Initialization step: initialize K centroids
Do
Assignment step: assign each data point to its centroid
Re-estimation step: Re compute centroid (cluster centers)
While (there are still changes in the centroid)
select k objects as initial cluster centers. The dataset is partitioned into K clusters and the data points are randomly set to the clusters resulting in clusters that have roughly the same number of data points.

Step 2 : For each data point calculate the Euclidean distance from the data point to each cluster.
The Euclidean distance is the straight-line distance between two pixels,

$$ED = \sqrt{(x1 - x2)^2 + (1 - y2)^2}$$

Where $(x1, y1)$ & $(x2, y2)$) are two data points.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

Step 3: If the data point is nearer to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster. Repeat the above step until a entire pass through all the data points' results in no data point transferring from one cluster to another. At this point the clusters are stable and the clustering process ends.

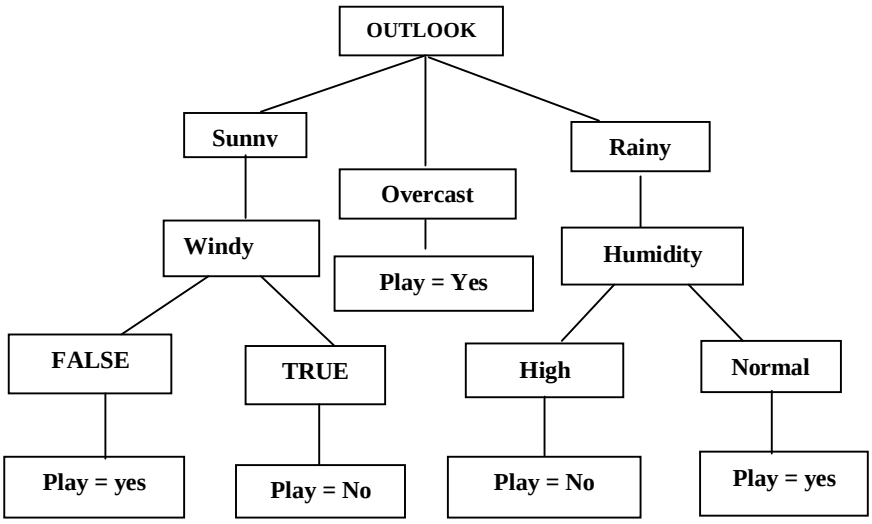
Step 4: Repeat this process until the criterion function converged and Square error criterion for clustering. The choice of initial partition can greatly affect the end clusters that result, in terms of inter-cluster and intra-cluster distances and cohesion.

K-Mean clustering algorithm is used for finding maximum and minimum weather condition detail in overall process. It is the most important flat clustering algorithm. Its objective is to reduce the average Squared Euclidean Distance of documents from their cluster centers is defined as mean or cendroid. There are two clusters like cluster instance 0 and cluster instance 1. These two clusters show final cendroid cluster display the maximum and minimum values of overall weather attributes.

VII. SIMULATION AND RESULT

The following Figure 1 explains the overview of weather prediction process. Weather prediction mainly based on weather conditions.

(Figure 1) outlook label have three seasons like Sunny, Overcast and Rainy. And also windy and humidity level also checked this process.



(Figure 1: overview of weather prediction process)

The following table explains the comparison of existing and proposed system. K-means algorithm produced a result 89.23%,EM produced 88.60% and DT produced 86.32% result on proposed system. Frequency Pattern Algorithm produced a result 70% on existing system (Table 1) .

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

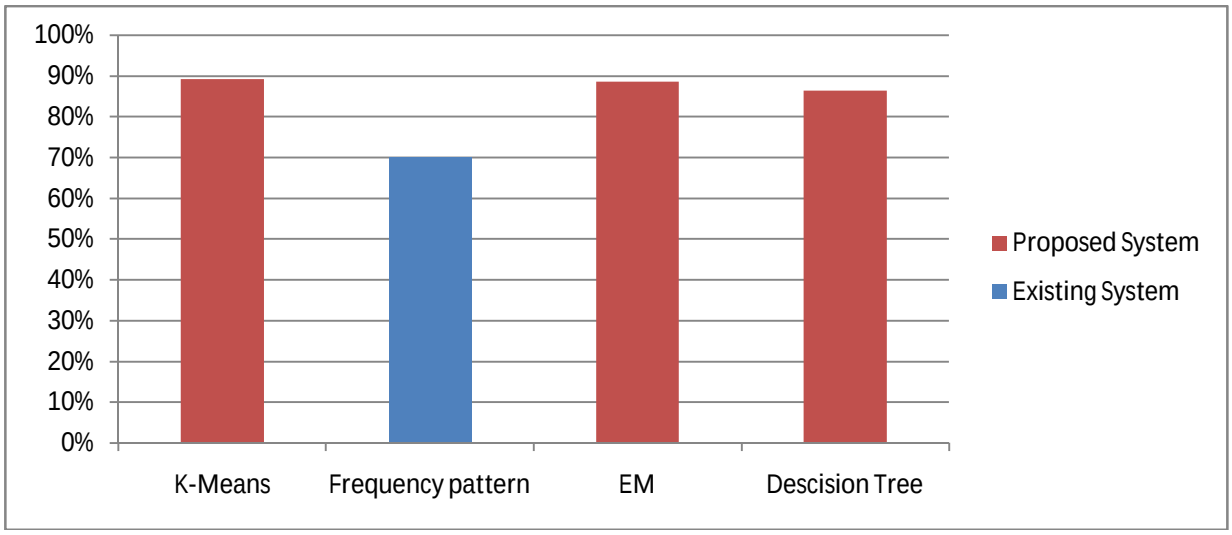
Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

Algorithms	Existing System	Proposed System
K-Means	0%	89.23%
Frequency pattern	70.00%	0%
EM	0%	88.60%
Decision Tree	0%	86.32%

(Table 1 : Comparision table of Existing and Proposed system):

The Following graph explains the analysis of existing and proposed system result analysis. The result shows the good result on proposed system compared than existing system (Figure 2).



(Figure 2 explains the result analysis of existing and proposed system)

VIII. CONCLUSION AND FUTURE ENHANCEMENT

Weather forecasting is an vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world. The proposed system analyzed the use of data mining techniques in forecasting weather. This can be carried out using data mining techniques like classification and clustering techniques and the algorithm like Decision tree Algorithm and clustering algorithms has been applied to the data collected in specific time.

Weka tool is applied with decision tree J48 algorithm for classifying weather parameters and Expectation Maximization algorithm for finding Maximum values of weather conditions. Finally applied k-means clustering

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

algorithm is applied for finding maximum and minimum weather condition details on the dataset 2015 and 2016 . Scatter plot matrix is used to display the overall result like Sunny, Rainy, Overcast and weather Temperature and Humidity ,Windy details. These algorithms which gave the best results on weather variables. The results shows that given enough details on data mining techniques can be used for weather forecasting.

In **future** work we used Neuro fuzzy models like Mamdani model and Takagi sugeno model with Mat lab can be apply for weather prediction process. It will help to find weather condition like best climate or worst climate.

REFERENCES

1. Arti R. Naik., Prof. S.K.Pathan ., " Weather classification and forecasting using feed forward neural network" , International Journal of Scientific and Research Publications, Vol 2, Issue 12, 2012 .
2. Prasanth Rao Jillella.S.S .,"Weather Forecasting Using Artificial Neural Networks And Data Mining Techniques" , International Journal Of Innovative Technology And Research Vol No:3, pp.2534 – 2539, 2015.
3. M. Viswambari., Dr. R. Anbu Selvi . , "Data Mining Techniques To Predict Weather: A Survey", International Journal Of Innovative Science, Engineering & Technology, Vol. 1 Issue 4, pp. 2348 – 7968, 2014 .
4. Ankita Joshi1 . , Bhagyashri Kamble2 ., Vaibhavi Joshi3 ., Komal Kajale4 ., Nutan Dhange5 , "Weather Forecasting And Climate Changing Using Data Mining Application", International Journal Of Advanced Research In Computer And Communication Engineering, Vol. 4, Issue 3, 2015.
5. Rajesh Kumar, "Decision Tree For The Weather Forecasting". International Journal Of Computer Applications (0975–8887) Volume76–No.2, 2013.
6. Meghali A. Kalyankar, S. J. Alaspurkar, " Data Mining Technique to Analyse the Metrological Data", International Journal of Advanced Research in Computer Science and Software Engineering 3(2), 114-118, February – 2013.
7. www.wikipedia.com
8. Fair bridge R.W. , "Data mining in course management systems: Moodle case study and tutorial/ Computers & Education", Vol. 51, No. 1, pp. 368-384, 2008 .
9. Rushing J.R ., Ramachandran U. , Nair S. ,Graves R., Welch, LinA., , "A Data Mining Tool kit for Scientists and Engineers"/ Computers & Geosciences, 31,pp607-618, 2005.
10. "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies" /"I.J. Information Engineering and Electronic Business", Vol.1, pp.51-59 ,2012.
11. Han,J.MichelinK, " Data Mining: Concepts and Techniques"/ , San Francisco, CA: Morgan Kaufmann publishers, ,2007 [8]The Many Users and Uses of Weather Data WFO Albuquerque Weather Data Awareness Week 2005.
12. Divyachuan ., jawaharthakur ., "Data mining techniques for weather prediction: A review"/ International journal on recent and innovation trends in computing and communication, volume: 2 issue: 8, 2014.

.