Data Mining Technique To Analyse The Temperature Data In Machine Learning Tool

S.Yoganica[1], A.Mohamed Yasin[2], Dr.S.Manimozhi[3]
[1, 3] Dept of computer science and application
[2]Assistant professor, Dept of computer science and application
[1, 2, 3] Periyar maniammai institute of science and technology Vallam, Thanjavur, Tamil Nadu, India

Abstract- *Data Mining is the process of discovering new patterns from large data sets, this technology which is employed in inferring useful knowledge that can be put to use from a vast amount of data, various data mining techniques such as Classification, Prediction, Clustering and Outlier analysis can be used for the purpose. Weather is one of the meteorological data that is rich by important knowledge. Meteorological 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. We know sometimes Climate affects the human society in all the possible ways. Knowledge of weather data or climate data in a region is essential for business, society, agriculture and energy applications. The main aim of this paper is to overview on Data mining Process for weather data and to study on weather data using data mining technique like classification technique. By using this technique we can acquire Weather data and can find the hidden patterns inside the large dataset so as to transfer the retrieved information into usable knowledge for classification and prediction of climate condition. They discussed how to use a data mining technique to analyze the Rain like Weather data. They compare the classification technique Like Naïve bayes, Decision Tree they produce 98% accuracy random forest.*

I. INTRODUCTION

Weather data has Synoptic data or climate data are the two classifications. Climate data is the official data record, usually provided after some quality control is performed on it. Synoptic data is the real- time data provided for use in aviation safety and forecast modelling. We know the Climate and weather affects the human society in all the possible ways. For example: Crop production in agriculture, the most important factor for water resources i.e. Rain, an element of weather, and the proportion of these elements increases or decreases due to change in climate. The effect of frost on the growth and quality of crops is leading potentially to total harvest failure. Energy sources, e.g. natural gas and electricity are depends on weather conditions. Hence changes Climate or weather condition is risky for human society as in all the possible ways.

The increasing availability of climate data during the last decades (observational records, radar and satellite maps, proxy data, etc.) makes it important to find effective and accurate tools to analyze and extract hidden knowledge from this huge data. Meteorological 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 . Useful knowledge can play important role in understanding the climate variability and climate prediction. In turn, this understanding can be used to support many important sectors that are affected by climate like agriculture, vegetation, water resources and tourism.

The paper describes how to use a data mining technique and how to develop a system that uses numeric historical data to forecast the climate of a specific region or city. In this paper we try to extract useful knowledge from weather data by using Clustering technique i.e. k-Means Partitioning Method and we discussed Data mining Process for Weather Data Analysis System.

Time series data mining is one of the hot research topics in the domain of knowledge discovery .The data with time series approach is collected over a specific period of time such as daily,weekly, monthly, quarterly or yearly .This data can be used for predictions in different domains such as finance, stock market and climate change etc. Data mining techniques are used to extract the hidden knowledge from time series data for future use .Weather prediction with time series data is beneficial but quite challenging task. It comes with an array of complexities which needs Tobe tackled for optimal results .The statistical weather data has a wide variety of fields which are called features such as humidity, pressure, wind speed, pollutants, concentrations etc. Data mining techniques can predict the weather on the basis of hidden patterns among these features .Rainfall prediction is an important aspect of climate forecasting. Accurate and timely rainfall prediction is crucial for the planning and management of water resources,

flood warnings, construction activities and flight operations etc.. This study used 5data mining techniques for rainfall prediction in Lahore, capital of Punjab province, Pakistan. In Lahore, development and construction activit ie are increasing exponently, so timely rainfall prediction is crucial for better assessment of future requirements and planning. The used data mining techniques include: Support Vector Machine, Naïve Bayes, k Nearest Neighbor, Decision Tree and Multilayer Perceptron. These algorithms belong to supervised data mining class where pre-classified data is required first for training purpose. During training, these algorithms make rules of classification for input dataset (test data).In this research, dataset is obtained from kaggle .com website and 1 year data collect, which contains several weather related attributes such as Temperature, Atmospheric pressure,Relative humidityetc.For rainfall prediction, a classification frameworks used in which the dataset gone through cleaning and normalization process before classification. Cleaning is performed to deal with the missing values and the purpose of normalization is to keep the attribute values in a certain limits. These pre-processing activities are crucial for the smooth classification process as well as for good results. Prediction performance of used data mining techniques is evaluated in terms of precision, recall and f measure, which are the important metrics of information retrieval. Finally the results are shown in tables and graphs.
Many researchers have been working to achieve high accuracy in rainfall prediction using data mining techniques; some of the selected studies are discussed here. Researchers in performed a comparative analysis of multiple classifiers for rainfall prediction in Malaysia. Classifiers include Naïve Bayes, Support Vector Machine, Decision Tree, Neural Network and Random Forest. Dataset was obtained from multiple stations of Selangor, Malaysia. Pre-processing tasks were applied before classification to deal with the noise and missing values. According to results, RandomForest performed better as with small training data it correctly classified large amount of instances. In ,researchers presented Clusterwise Linear Regression (CLR)method, which is the combination of clustering andregressiontechniques. The proposed technique is used to predict monthly rainfall in Victoria, Australia,byusing input data of 8 geographically diverse weather stations.To analyze the performance of proposed CLR, results were compared with other techniques such as: CLRusing the maximum likelihood framework by the expectation maximization algorithm, multiple linear .

II. LITERATURE SURVEY

ANALYSIS OF TEMPERATURE AND HUMIDITY DATA FOR FUTURE VALUE PREDICTION

Knowledge of climate data in a region is essential for business, society, agriculture, pollution and energy applications. In research and development, it forces the researchers to pay an extra attention towards this type of matter. As there is a spectacular achievement in this field over the past few years, among all the other seasonal climatic attributes, the main factor used by the researcher is the Sea Surface Temperature (SST) to develop the systems for temperature and humidity prediction.

Data mining is one such technology which is employed in inferring useful knowledge that can be put to use from a vast amount of data, various data mining techniques such as Classification, Prediction, Clustering and Outlier analysis can be used for the purpose. The main aim of this paper is to acquire temperature and humidity data and use an efficient data mining technique to find the hidden patterns inside the large dataset so as to transfer the retrieved information into usable knowledge for classification and prediction of climate condition.

Synoptic data or climate data are the two classifications of weather data. Synoptic data is the real-time data provided for use in aviation safety and forecast modeling. Climate data is the official data record, usually provided after some quality control is performed on it .

Climate and weather affects the human society in all the possible ways. Crop production in agriculture, the most important factor for water resources i.e. Rain, an element of weather, and the proportion of these elements increases or decreases due to change in climate . Energy sources, e.g. natural gas and electricity are greatly depends on weather conditions. Climate is not fixed, the fluctuation in the climate can be seen from year to year, e.g. rain/dry; cold/warm seasons significantly influence society as in all the possible ways. Depending upon the techniques used Data Mining can be divided into three basic types, i.e. Association Rules Mining,Cluster analysis and Classification/Prediction . The paper describes how to use a data mining technique, “k-Nearest Neighbor (KNN)”, how to develop a system that uses numeric historical data to forecast the climate of a specific region or city. The main aim of this paper is to acquire temperature and humidity data and use k-Nearest Neighbor algorithm to find hidden patterns inside a large data so as to transfer the retrieved information into usable knowledge for classification and prediction of temperature and humidity.

IMPLEMENTATION OF DATA MINING TECHNIQUES FOR METEOROLOGICAL DATA ANALYSIS

Meteorological 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. In this paper we try to extract useful knowledge from weather daily historical data collected locally at Gaza Strip city. The data include nine years period [1977-1985]. After data preprocessing, we apply outlier analysis, clustering, prediction, classification and association rules mining techniques. For each mining technique, we present the extracted knowledge and describe its importance in meteorological field.

III. METHODOLOGY

EXISTING METHOD

DATA MINING IN METEOROLOGY

Meteorology is the interdisciplinary scientific study of the atmosphere. It observes the changes in temperature, air pressure, and moisture and wind direction. Usually, temperature, pressure, wind measurements and humidity are the variables that are measured by a thermometer, barometer, anemometer, and hygrometer, respectively. There are many methods of collecting data and Radar, Lidar, satellites are some of them. Weather forecasts are made by collecting quantitative data about the current state of the atmosphere. The main issue arise in this prediction is, it involves high dimensional characters. To overcome this issue, it is necessary to first analyze and simplify the data before proceeding with other analysis. Some data mining techniques are appropriate in this context.

CLUSTER ANALYSIS

Clustering analyses data objects without consulting a known class label. The unsupervised learning technique of clustering is a useful method for ascertaining trends and patterns in data, when there are no pre-defined classes. There are two main types of clustering, hierarchical and partition. In hierarchical clustering, each data point is initially in its own cluster and then clusters are successively joined to create a clustering structure. This is known as the agglomerative method. In partition clustering, the number of clusters must be known a priori. The partitioning is done by minimizing a measure of dissimilarity within each cluster and maximizing the dissimilarity between different clusters.

K-NEAREST NEIGHBORS ALGORITHM

Known as: Ibk algorithm, Nearest neighbors classifier, K-NN

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression: k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

PROPOSED SYSTEM

CLASSIFICATION:

Data set: The weather dataset collected by kaggle .com website .the data contains 366 instance and 25 attributes. the data is divided into two parts

1 st part: training data
2 nd part : Test data

Classification of patient Support Vector Machine is a supervised technique which is associated with learning algorithms.SVM technique is used for finding patterns for classification and quantitative predictions of one variable from the values of another. General format of SVM is predicting the output based on the trained data. The inputs are given the outputs are predicted within two options. New behaviors are then trained in same space such that output follows within two classes.

IV. MODULES

KDD Process

This term originated with research in the field of artificial intelligence, this process involves some stages in the analysis of data: Selection, processing, transformation (in case If neces- sary), data mining to extract patterns and finally interpretation and evaluation of the discovered structures . shows an illustration of the KDD process and its steps.

Classification - Real Data. When possessing real data the algorithm determined that the humidity of the soil depends directly on the variable illumination, when it has no value, the second with greater weight is the relative humidity. Therefore, based on these two factors, the classification rules are formed.

Prediction - Real Data. The Soil Moisture variable was predicted with a high degree of confidence, since the data obtained by the prediction process approached the real, with high precision.

Relative Humidity. Plants have to transpire water in order to transport nutrients and regulate their growth, this factor depends of the transpiration and the temperature that the greenhouse has. The percentage of relative humidity in which the plants have a correct development is of 55% to 70% .

Temperature. For their growth process and correct development, plants need a suitable temperature, otherwise, these processes stop. When this factor drops to zero degrees or less, the silver can suffer severe damage to their tissues, as it usually happens when they are in the open air during the night frosts.

In general the favorable effect of the greenhouse on the development of roots and cultivation is to maintain the adequate temperature of both air and soil .

V. SYSTEM ENVIRONMENTS

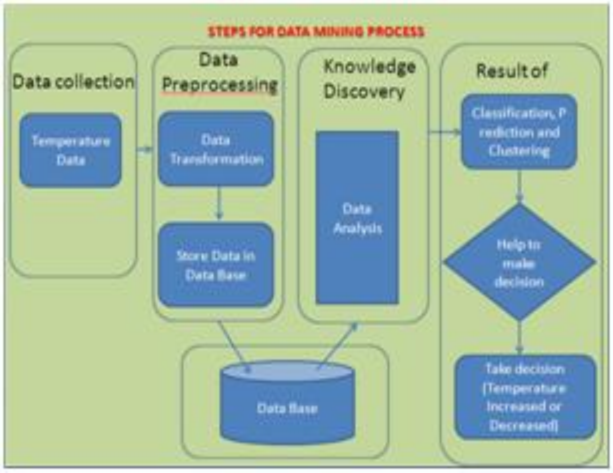
HARDWARE REQUIREMENTS

- Processor : Dual core processor 2.6.0 GHZ
- RAM : 1GB
- Hard disk : 160 GB
- Compact Disk : 650 Mb
- Keyboard :Standard keyboard
- Monitor : 15 inch color monitor

SYSTEM REQUIREMENT

- Frond end : weka
- Back end : MS EXCEL,SQL

Architecture:



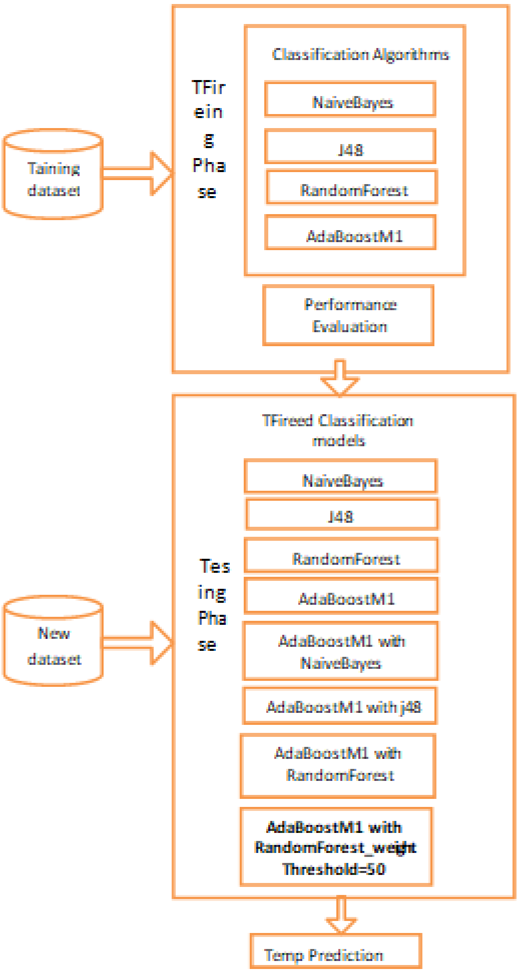
LEVEL 0

The Level 0 DFD shows how the system is divided into 'sub-systems' (processes), each of which deals with one or more of the data flows to or from an external agent, and which together provide all of the functionality of the system as a whole. It also identifies internal data stores that must be present in order for the system to do its job, and shows the flow of data between the various parts of the system.

VI. DATA FLOW DIAGRAM

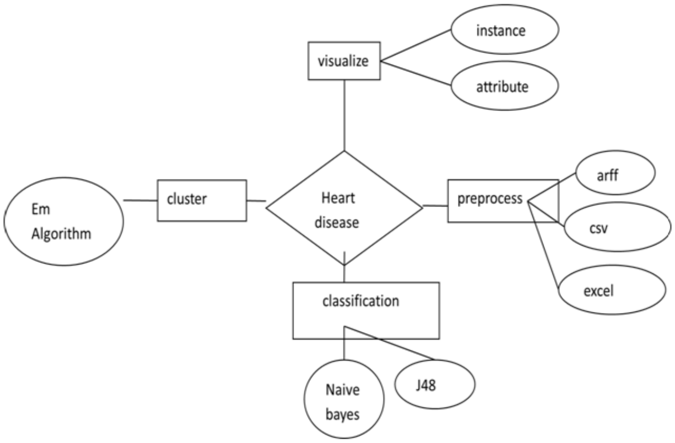
A Data Flow Diagram is a graphical representation of the "flow" of data through an information system, modeling its process aspects. A DFD is often used as a preliminary step to create an overview of the system which can later be elaborated. DFD can also be used for the visualization of data processing structured design. A DFD shows what kind of information will be input to and output from the system, where the data will come from and go to and where the data will be stored.

Proposed Architecture:



Level 0

ER DIAGRAM



C

ALGORITHM:

NAÏVE BAYES CLASSIFICATION

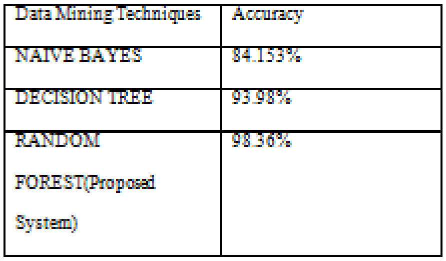
The Naive Bayes classification algorithm is an probabilistic classifier. It is based on probability models that incorporate strong independence conventions. Therefore they are considered as naive. To derive probability models by using Bayes' statement. Benefits of Naive Bayes: Super simple. If the NB provisional independence assumption essentially holds, a Naive Bayes classifier will join quicker than discriminative models like logistic regression, so need less tFireing data.

J48 CLASSIFIER

The C4.5 algorithm for constructing decision trees is applied in Weka as a classifier called J48. Classifiers, like filters, are prepared in a hierarchy: J48 has the full name weka.classifiers.trees.J48.

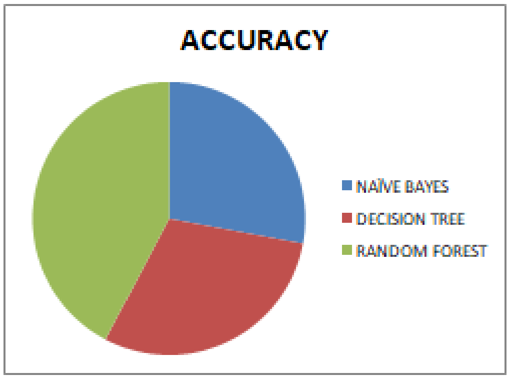
RESULTS AND DISCUSSION

Table: . Accuracy Calculations



Data Mining Techniques	Accuracy
NAIVE BAYES	84.153%
DECISION TREE	93.98%
RANDOM FOREST(Proposed System)	98.36%

This shows the accuracy of the result based on the data mining techniques.



VII. CONCLUSION

Classification and clustering are the methods used in data mining for analysing the data sets and divide them on the basis of some particular classification rules or the association between objects. Classification categorizes the data with the help of provided training data. On the other hand, clustering uses different similarity measures to categorize the data.The collection and archiving of Temperature data is important because it provides an economic benefit but the local/national economic needs are not as dependent on high data quality as is the weather risk market. Data mining tasks provide a very useful and accurate knowledge in a form of rules, models, and visual graphs. This knowledge can be used to obtain useful prediction and support the decision making for different sectors. The result of the this work compare the classifier technique will produce random forest algorithm is better .the accuracy rate 98%.

FUTURE WORK

Weather maps provide past, current, and future radar and satellite images for local cities and regions.