ISSN:0975-9646

Analysis of Temperature and Humidity Data for Future value prediction

Badhiye S. S.#1, Wakode B. V.#2, Chatur P. N.#3

#1. M. Tech Computer Science and Engineering Department,

#2. Asst. Professor Information Technology Department,

#3. Head Computer Science and Engineering Department,

Government College of Engineering, Amravati, Maharashtra, India.

Abstract - **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.**

Keywords: **Data Mining, Data Mining Techniques, Sea Surface Temperature**

I. INTRODUCTION

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 [1].

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 [7]. 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 [7]. 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.

II. LITERATURE SURVEY

Prediction of the future values by analyzing Temperature and humidity data is one of the important parts which can be helpful to the society as well as to the economy. Work has been done in this constrain since years. Different techniques have been applied to predict the temperature and humidity and other parameters of weather. Some of the work in this area is as follows:

In data mining, the unsupervised learning technique of clustering is a useful method for ascertaining trends and patterns in data. Most general clustering techniques do not take into consideration the time-order of data. *Tasha R. Inniss* used a mathematical programming and statistical techniques and methodologies to develop a seasonal clustering technique for determining clusters of time series data, and applied this technique to weather and aviation data to determine probabilistic distributions of arrival capacity scenarios, which can be used for efficient traffic flow management. The seasonal clustering technique is modeled as a set partitioning integer programming problem and resulting clustering’s are evaluated using the mean square ratio criterion [2]. The resulting seasonal distributions, which have satisfied the mean square ratio criterion, can be used for the required inputs (distributions of airport arrival capacity scenarios) into stochastic ground holding models. In combination, the results would give the optimal number of flights to ground in a ground delay program to aid more efficient traffic flow management [2].

S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias investigate the efficiency of data mining techniques in estimating minimum, maximum and mean temperature values. Using temperature data from the city of Patras in Greece, a Regression algorithm is applied for the number of results. The performance of these algorithms has been evaluated using standard statistical indicators, such as Correlation Coefficient, Root Mean Squared Error, etc. [1]

Godfrey C. Onwubolu1, Petr Buryan, Sitaram Garimella, Visagaperuman Ramachandran,Viti Buadromo and Ajith Abraham, presented the data mining activity that was employed in weather data prediction or forecasting. The approach employed is the enhanced Group Method of Data

Handling (e-GMDH). The weather data used for the research include daily temperature, daily pressure and monthly rainfall [3]. The results of e-GMDH were compared with those of PNN and its variant, e-PNN. E-GMDH outperformed PNN an its variant in most modeling and prediction problem. They showed that end users of data mining should endeavor to follow the methodologies of data mining since suspicious data points or outliers in a vast amount of data could give unrealistic results which may affect knowledge inference.

S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias proposed a hybrid data mining technique that can be used to predict more accurately the mean daily temperature values [4], it was found that the regression algorithms could enable experts to predict temperature values with satisfying accuracy using as input the temperatures of the previous years. The hybrid data mining technique produce the most accurate results.

Simple temperature prediction methods mining in the past weather data records produced accurate prediction for development of intelligent control solutions. The problem was closely related to the prediction of the actual weather conditions within the immediate environment of the greenhouse, an intelligent greenhouse collects its own climate data, with time weather records from weather station localized strictly by the greenhouse were mined to the algorithm, increasing the prediction accuracy. *Peter Eredics* demonstrates the limited performance of uninformed, simple methods for temperature forecasts, and introduces more accurate solutions using information from the problem domain [5].

A. Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers.

Example: Use in finding Fraud usage of credit cards. Outlier Analysis may uncover Fraud usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the location and type of purchase or the purchase frequency.

B. Clustering

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 [10]. 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.

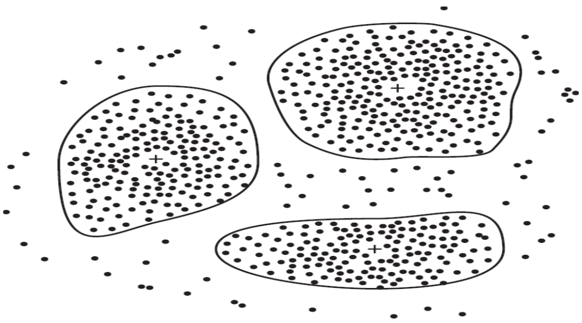

Fig. 1 Figure of Cluster Analysis

C. Classification and Prediction

Classification is the process of finding a model that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown [6].

Classification model can be represented in various forms such as

1) IF-THEN Rules
2) A decision tree
3) Neural network.
4) K-Nearest Neighbor etc.

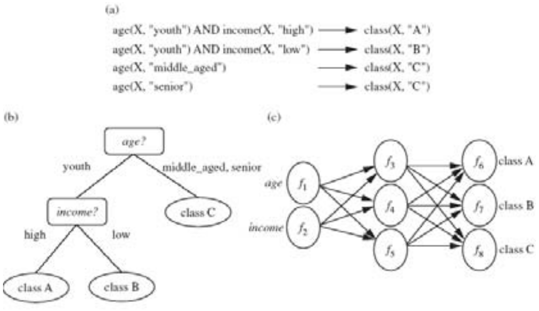
(a)
age(X, "youth") AND income(X, "high") → class(X, "A")
age(X, "youth") AND income(X, "low") → class(X, "B")
age(X, "middle_aged") → class(X, "C")
age(X, "senior") → class(X, "C")
(b)
age?
youth
middle_aged, senior
income?
class C
high
low
class A
class B
(c)
age
income
f1
f2
f3
f4
f5
f6
f7
f8
class A
class B
class C

Fig. 2 Classification models

III. MOTIVATION

It has become important to find an effective and accurate tool to analyze and extract hidden knowledge from climate data due to its increasing availability during the last decade.

Knowledge of climate data in a region is essential for business, society, agriculture, pollution and energy applications, research and development.

Temperature and humidity data is also used in the estimation of bio-meteorological parameters in a region. Data Mining is recently applied to show affect of climate variation in vegetation and thus, statistical Data Miner software STATISTICA 10 is used for the data mining purpose which is intelligent data miner software where various Artificial Intelligence algorithms are applied.

IV. PROPOSED PLAN

The objective is to be able to predict the values of temperature and humidity parameters of climate with higher accuracy, and prove the prediction ability of data mining technique in the same context.

There are different steps in which this paper will be implemented and various methodologies are used in each step as shown in the figure below:

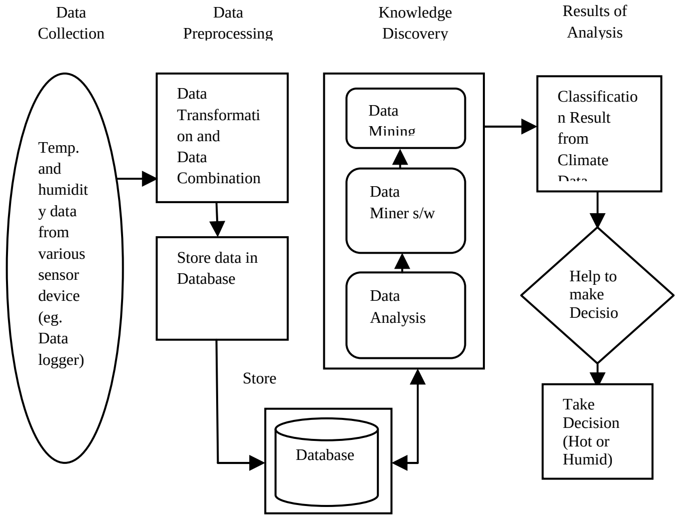
Data Collection
Data Preprocessing
Knowledge Discovery
Results of Analysis
Temp. and humidity data from various sensor device (eg. Data logger)
Data Transformation and Data Combination
Store data in Database
Store
Database
Data Mining
Data Miner s/w
Data Analysis
Classification Result from Climate Data
Help to make Decisio
Take Decision (Hot or Humid)

Fig. 3 Design of Temperature and Humidity Data Analysis System

A. Data Collection

This is the most important part while implementing any of the data mining technique and thus for this purpose we are using 10 channel midi-data logger system. This system provides temperature and humidity data in form of excel sheets.

Data Loggers are based on digital processor. It is an electronic device that record data over the time in relation to location either with a built in instrument or sensor or via external instruments and sensors.

Data Logger can automatically collect data on a 24-hour basis, this is the primary and the most important benefit of using the data loggers.

B. Data Pre-Processing

The next important step in data mining is data preprocessing the challenge faced in knowledge discovery process in temperature and humidity data is poor data quality. Thus, data is to be pre-processed so as to remove the noisy and unwanted data. In this study, the weather data is used which consists of various parameters as temperature, humidity, rain, wind speed etc. but only temperature and humidity data is required for analysis in this proposed study, thus, pre-processing means removing the other unwanted parameters from the dataset.

C. Knowledge Discovery

For knowledge extraction various data mining techniques such as Outlier Analysis, Clustering, Prediction and Classification and Association rules can be applied in Statistical Data Miner Software such as Statistica 10.

The classification algorithm k-Nearest Neighbor is used which is based on Euclidean distance between two points, used to find out the closeness between unknown samples with the known classes. The unknown sample is then mapped to the most common class in its k-nearest neighbors.

D. Result of Analysis

The future values of temperature and humidity are predicted depending on the result of the classification algorithm.

V. CONCLUSION

k-Nearest Neighbor classification is an easy to understand and easy to implement classification technique. Despite its simplicity, it can perform well in many situations. In particular, a well known result by Cover and Hart [11] shows that the error of the nearest neighbor rule is bounded above by twice the Bayes error under certain reasonable assumptions. Also, the error of the general k-Nearest Neighbor method asymptotically approaches that of the Bayes error and can be used to approximate it. k-Nearest Neighbor is particularly well suited for multi-modal classes as well as applications in which an object can have many class labels. Thus, the proposed method aims at temperature and humidity prediction with accuracy nearer to 100%.

REFERENCES

[1] S. Kotsiantis and et. al., “Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values”, *World Academy of Science, Engineering and Technology* 2007 pp. 450-454

[2] Tasha R. Inniss “Seasonal clustering technique for time series data”, *European Journal of Operational Research* (175) 2006 pp. 376–384

[3] Godfrey C. Onwubolu1, Petr Buryan, Sitaram Garimella, Visagaperuman Ramachandran,Viti Buadromo and Ajith Abraham “Self-organizing data mining for weather Forecasting” *IADIS European Conference Data Mining* 2007 pp. 81-88

[4] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias, “A Hybrid Data Mining Technique for Estimating Mean Daily Temperature Values”, *IJICT Journal* Volume 1 (5) pp. 54-59

[5] Peter Eredics “Short-Term External Air Temperature Prediction for an Intelligent Greenhouse by Mining Climatic Time Series” *WISP 2009 6th IEEE International Symposium on Intelligent Signal Processing*, 26–28 August, 2009 Budapest, Hungary pp.317-322

[6] Thair Nu Phyu, “Survey of classification techniques in Data Mining”, *IMECS 2009* Volume 1 Hong Kong pp. 1-5

[7] Han J., Kamber M.: *Data Mining concepts and Techniques*, Elsevier Science and Technology, Amsterdam 2006

[8] Fix E., Hoges J. L.: *Discriminatory Analysis – Non parametric Discrimination : consistency properties*. USAF School of Aviation Medicine, Ranolph Field Texas (1951)

[9] Larose D. T.: *Discovering Knowledge in Data: An Introduction to Data Mining*, Wiley, Chichester 2005

[10] B. S. Everitt, *Cluster Analysis*, Halsted Press, John Wiley and Sons, New York, 1975

[11] Cover T, Hart P (1967) “Nearest neighbor pattern classification”. *IEEE Trans Inform Theory* Volume 13(1) pp. 21–27