International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-3 Issue-11 September-2013

Data Mining with Meteorological Data

A. R. Chaudhari[1], D. P. Rana[2], R. G. Mehta[3]

Abstract

Prediction of the future values by analysing meteorological data is one of the important parts which can be helpful to the society as well as to the economy. Estimates of these values at a specific time of day, from daytime and daily profiles, are needed for a number of environmental, ecological, agricultural and technical applications, ranging from natural hazards assessments, crop growth forecasting to design of solar energy systems. Work has been done in this constrain since years. This paper is discussing the application of different data mining techniques applied in various ways to predict or to associate or to classify or to cluster the pattern of meteorological data.

Keywords

Data Mining, Environment, Meteorological Data, Prediction

1. Introduction

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data [1]. In contrast to standard statistical methods, data mining techniques search for interesting information without demanding a priori hypotheses, the kind of patterns that can be discovered depend upon the data mining task employed. There are two types [1] of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data and predictive data mining tasks that attempt to do predictions based on inference on available data.

We were looking for some challenging dataset area from which we can mine useful knowledge. So, we have selected meteorological data which can be huge in size with the time information.

A. R. Chaudhari, Computer Engineering Department, Sardar Vallabhbhai National Institute of Technology, Surat, India.
D. P. Rana, Computer Engineering Department, Sardar Vallabhbhai National Institute of Technology, Surat, India.
R. G. Mehta, Computer Engineering Department, Sardar Vallabhbhai National Institute of Technology, Surat, India.

Knowledge of meteorological data in a site is essential for number of applications like Rainfall analysis and prediction, temperature prediction, pollution and energy application studies and development. For example, in advanced energy system designs the profile of any meteorological parameter is a prerequisite for systems operating management on daily and/or hourly basis and simulations of long-term performance of energy plants require detailed and accurate meteorological data as input. Especially temperature data is used to determine the number of application parameter estimation. An element of weather and the proportion of these elements increases or decreases due to change in climate temperature [2], [3].

Meteorologists and weather forecasters make predictions for weather mainly on numerical and statistical models. The simulation conducted often requires intensive computations, involving complex differential equations and computational algorithms. The accuracy is bound by constraints, such as the adoption of incomplete boundary conditions, model assumptions and numerical instabilities, etc. [2].

Since the data of everyday weather condition are so huge the weather parameter relations cannot be found easily only by directly observation which can be found by data mining techniques. There are number of data mining techniques and methods are available and they are applied to many application areas [2]. Some applications which use a dynamic prediction based approach include ophthalmic oncology, vehicle fault diagnosis, pre-fetching, fault prediction models, fault restoration prediction models, financial distress prediction models, chemical reactivity predictions, real time vehicle tracking, forecasting, anomaly detection, churn prediction, clinical predictions, etc.[4]. Thus, here we focused on survey of different data mining techniques utilized to find the hidden relationship between various dataset variables values and to identify correlations between meteorological data.

The organization of rest of this paper is as follow: In Section 2 we are presenting the brief overview of the data mining techniques. In Section 3 we are discussing the literature survey on meteorological

data prediction. Section 4 is concluding the research work and showing the future works.

2. Data Mining Techniques

Data mining techniques are used to specify the kind of patterns to be found in data mining tasks that characterize the general properties of the data in the database or perform inference on the current data to make predictions. As in some cases, users may have no idea regarding what kinds of patterns in their data may be interesting and hence may like to search for several different kinds of patterns in parallel. A data mining system can mine multiple kinds of patterns and also at various abstraction levels to accommodate different user expectations or applications. Also, data mining systems should allow users to specify hints to guide or focus the search for interesting patterns. Because some patterns may not hold for all of the data in the database, a measure of certainty or trustworthiness is usually associated with each discovered pattern [5].

1. Classification

Classification is the method to map each item of the selected data into one of a predefined set of classes [1], [6]. Given the set of predefined classes, a number of attributes and a "learning (or training) set", the classification methods can automatically predict the class of other unclassified data of the learning set. The accuracy of the classified data is first evaluated on the training data first, and then it will be applied to the real world dataset.

The application of classification is to classify a car loan applicant as a good or a poor credit risk. To solve, this type of problem there is a need of car loan application model to determine whether an applicant is a good credit risk at this time rather than in some future time period. The classification technique is the suitable technique here, as it will be modeled on history data of the loan persons.

2. Clustering

Clustering is the unsupervised classification of data into natural groups (called clusters) so that data points within a cluster are more similar to each other than to data points in other clusters [1], [7]. The term unsupervised stands for the fact that there is no a priori knowledge about the partition of the data. Clustering algorithms are based on some distance function that evaluates in which cluster an object should be assigned. There is also an evaluation function that evaluates how good the achieved clustering is. For example, minimizing the distance of each data point from the mean of the cluster to which it is assigned.

The application of clustering is to analyze e-commerce customer data to identify homogeneous subpopulations of customers. The clusters may represent individual target groups for marketing. A 2-D plot of customers can be used to locate customer in a city. Three clusters of data points are generated for the three cities.

3. Association Rules

Association rules are used to predict the relationship of a particular item in a data transaction on other items in the same transaction [1], [8]. An association rule is representing the relation in "X implies Y, with the confidence and the support factor provided by the user given limit", where X is called antecedent and Y is called consequent.

The main application of association rule mining is the market basket analysis which helps the retailer for the rack arrangement, cross-marketing, sale campaign analysis, etc. The Fig. 1 shows the association rule "Milk → Clothes" for Support= 2% and Confidence=60%. Here, a support of 2% means that 2% of all transactions under analysis show that Milk and Clothes are purchased together. A confidence of 60% means that 60% of the customers who purchased milk also bought the clothes. Association rules are considered interesting if they satisfy both a minimum support factor and a minimum confidence threshold.

Though, most of the available data-analysis methods are based on classification or clustering algorithms that try to categorize the data to the specific group or to establish groups of correlated data respectively.
Although such algorithms have been quite successful, they have some drawbacks like a data record has to be grouped in one and only one group and no relationship can be inferred between the different members of a group.

To overcome such problems, the potential impact of the association rule discovery technique is investigated. This is an unsupervised data mining technique that seeks descriptive rules in potentially very large datasets. This method should resolve the above drawbacks of existing grouping approaches for the following reasons. First, any data item can be assigned to any number of rules as long as its expression fulfills the assignment criteria, without limitation. Second, rules are orientated (If … then …)

and thus to a certain extent describe the direction of a relationship. Last but not least, by focusing on strong rules, the knowledge extractor does not have to browse and study a huge number of redundant rules.
The next section is discussing, how these various data mining techniques are applied to various fielded applications of meteorological data to derive the knowledge and to discover the relationships between the data.

3. Meteorological Data in Prediction

Some of the work in the area of prediction based on meteorological data is as follow:

In 2005, Liang et. al [9] derived the sequence of ecological events using temporal association rule mining. Red tide phenomena occurred during 1991 and 1992 in Dapeng bay, South China Sea was taken as an example to validate T-Apriori algorithm which generated frequent itemsets and corresponding temporal association rules and K-means clustering analysis used to map the quantitative association rule problem into the boolean association rules. Their experiment shows that T-Apriori algorithm can successfully extract temporal association rules that described the close relationship between environmental factors and ecological events.

In 2007, Huang et. al [10] analyzed historic salinity-temperature data to make predictions about future variations in the ocean salinity and temperature relations in the waters surrounding Taiwan. Traditional statistical models that assume data independence are not applicable as ocean data are often inter-related. Association rules mining can be used to find interesting salinity and temperature patterns. However, the traditional method ignores spatial and temporal information in the data. They proposed to use inter-dimensional association rules mining with fuzzy inference to discover salinity-temperature patterns with spatial-temporal relationships. They concluded their prediction of when, where and what event will occur with accuracy of 79%.

In 2007, the other authors S. Kotsiantis et. al [3] proposed a hybrid data mining technique that can be used to predict the mean daily temperature values. A number of experiments have been conducted with well-known regression algorithms using temperature data from the city of Patras in Greece. The performance of these algorithms has been evaluated using standard statistical indicators, such as Correlation Coefficient, Root Mean Squared Error. It was found that the regression algorithms could enable experts to predict the temperature values with satisfying accuracy using as input the temperatures of the previous years. The methods used in this work, has to be still validated by including temperature data with other meteorological parameters as well.

In 2011, N. Kohail et. al [11] tried 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, they apply basic algorithms of clustering, classification and association rules mining techniques. For each mining technique, they presented the extracted knowledge and describe its importance in meteorological field which can be used to obtain useful prediction and support the decision making for different sectors.

In 2011, Sivaramakrishnan et. al [12] presented the method for prediction of daily rainfall. Meteorological data from 1961-2010 were used for analysis. For the atmospheric parameters temperature, dew point, wind speed, visibility and precipitation (rainfall) were considered for analysis. They filtered and discretized the raw data based on the best fit ranges and applied association mining on dataset using Apriori algorithm to find the hidden relationship between various atmospheric parameters to predict the rainfall. Finally the data has been validated using classifier approach where correctly classified instances and incorrectly classified instances were found out to justify the accuracy of the data prediction model.

In 2011, AlRoby et al. [13] analyzed wind speed behaviour for the data recorded between 2004 to November 2006 at Gaza with different data mining techniques. But, they found the classification using neural networks as the appropriate one with accuracy upto 68% which is higher than all the other techniques.

In 2012, Badhiye et. al [14] described how to use a data mining technique, "k-Nearest Neighbor (KNN)", to develop a system that uses numeric historical data to forecast the climate of a specific region or city. They proposed design of Temperature and Humidity Data Analysis System. The main aim of their research 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.

In 2012, K. Pabreja [15] used K-means clustering technique on real life case of cloudburst of Dhaka, Bangladesh to discover the formation of cloudburst. And achievement of this concluded the early signal to forecast the cloudburst for future time.

In 2012, Dadaser-Celik et al. [16] analyzed the usage of association rule for discovering the relationships between stream flow and climatic variables in the Kizilirmak River Basin in Turkey.

From the literature survey it is found that that most of the meteorological data based prediction techniques and methods are based on the statistical or widely used data mining techniques like Clustering, Classification, Regression analysis, Decision Tree etc. and upto some extent the Temporal Association Rule Mining which are shown and analyzed in Table 1. Nowadays, research is considering time as one of the important constraints. Most of the meteorological variables are related to each other and also vary with each other with respect to time. So if we use the previous historical data for prediction then we can predict the most accurate value for any weather parameters.

4. Conclusion and Future Work

Data mining techniques are now the important techniques utilized in all application area related to meteorological data for the prediction and decision making by discovering interesting rules or patterns or groups that indicate the relation between variables which are discussed in brief here. To understand the application of data mining techniques, different research works are discussed here, that summarizes that the meteorological data are so specific than the traditional data.

The future area of research is so wide in various ways: As the meteorological data are huge and time stamped base data, there is a need to modify the traditional data mining technique as per the specific application need to discover the useful knowledge with integration of temporal concept.

References

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Second Edition, 2012.

[2] F. Olaiya, "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", International Journal of Information Engineering and Electronic Business, Vol. 1, pp. 51–59, 2012.

[3] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias, "A Hybrid Data Mining Technique for Estimating Mean Daily Temperature Values", IJICT Vol. 1, Issue. 5, pp. 54–59, 2007.

[4] S. Liao, P. Chu, P. Hsiao, "Data Mining Techniques and Applications – A Decade Review from 2000 to 2011", Expert Systems with Applications, Vol. 39 (12), pp. 11303–11311, 2012.

[5] T. Silwattananusarn and K.Tuamsuk, "Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2(5), September 2012.

[6] A. Thombre, "Comparing Logistic Regression, Neural Networks, C5.0 and M5′ Classification Techniques", MLDM'12 Proceedings of the 8th international conferenceon Machine Learning and Data Mining in Pattern Recognition, ISBN: 978-3-642-31536-7, Springer-Verlag Berlin, Heidelberg, pp. 132-140, 2012.

[7] S. Koteeswaran, P. Visu and J. Janet, "A Review on Clustering and Outlier Analysis Techniques in Datamining", American Journal of Applied Sciences, ISSN 1546-9239, Vol. 9 (2), pp. 254-258, 2012.

[8] Ziauddin, S. Kammal, K. Z. Khan and M. I. Khan, "Research on Association Rule Mining", Advances in Computational Mathematics and its Applications (ACMA) Vol. 2(1), ISSN 2167-6356, World Science Publisher, United States, pp. 226-236, 2012.

[9] Z. Liang, T. Xinming and J. Wenliang, "Temporal Association Rule Mining Based On T-Apriori Algorithm and its Typical Application", Intl. Symposium on Spatial-Temporal Modeling Analysis, Vol. 5, Issue. 2, 2005.

[10] Y. P. Huang, L. J. Kao and F. E. Sandnes, "Predicting Ocean Salinity and Temperature Variations Using Data Mining and Fuzzy Inference", Intl. Journal of Fuzzy Systems, Vol. 9, Issue. 3, September 2007.

[11] S. N. Kohail, A. M. El-Halees, "Implementation of Data Mining Techniques for Meteorological Data Analysis", Intl. Journal of Information and Communication Technology Research (JICT), Vol. 1, Issue. 3, July 2011.

[12] T. R. Sivaramakrishnan and S. Meganathan, "Association Rule Mining and Classifier Approach for Quantitative Spot Rainfall Prediction", Journal of Theoretical and Applied Information Technology, Vol. 34, Issue. 2, 2011.

[13] M. F. AlRoby and A. M. ElHalees, "Data Mining Techniques for Wind Speed Analysis (A case study for Gaza Strip), Journal of Computer Engineering, ISSN: 20101619, Vol. 2 (1), 2011.
[14] S. S. Badhiye, B. V. Wakode, P. N. Chatur, " Analysis of Temperature and Humidity Data for Future value prediction", Intl. Journal of Computer Science and Information Technologies (IJCSIT), Vol. 3, Issue. 1, pp. 3012 – 3014, 2012.
[15] K. Pabreja, "Clustering technique to interpret Numerical Weather Prediction output products for forecast of Cloudburst", International Journal of Computer Science nd Information Technologies (IJCSIT), ISSN: 0975-9646, Vol. 3 (1), 2996 – 2999, 2012.
[16] F. Dadaser-Celik, M. Celik and A.S. Dokuz, "Associations between Stream Flow and Climatic Variables at Kizilirmak River Basin in Turkey", Global NEST Journal, Vol. 14 (3), pp. 354-361, 2012.

A. R. Chaudhari is PG Student at Computer Engineering Department, S. V. National Institute of Technology, Surat, Gujarat-395007, India.

D. P. Rana is Assistant Professor at Computer Engineering Department, S. V. National Institute of Technology, Surat, Gujarat-395007, India. She obtained her M. Tech.(R) degrees from S. V. National Institute of Technology with specializations in Computer and is currently pursuing her PhD degree. Her research interest is in the field of security in web applications, database management system, data mining and web data mining. She has published 14 papers in the area of data mining. She is a life member of ISTE and CSI.

R. G. Mehta is Associate Professor at Computer Engineering Department, S. V. National Institute of Technology, Surat, Gujarat-395007, India. She obtained her M. Tech.(R) degrees from S. V. National Institute of Technology with specializations in Computer and is currently pursuing her PhD degree. Her research interest is in the field of database management system, data mining, classification and clustering. She has published 28 papers in the area of data mining. She is a life member of ISTE and CSI.

Table 1: Analysis of Research Work

Author	Work	Technique	Remarks
Liang et. Al [9]	To derive the relationship between environmental and ecological factors	Temporal Association Rule Mining	Derived the red tide phenomena occurrence in Dapeng bay
Huang et. al [10]	Analyzed historic salinity-temperature data to make predictions about future temperature variations	Inter-transaction association rules mining with spatial information	Rule contains the salinity-temperature variations- used to predict when, where and what events
Kotsiantis et. al [3]	Proposed hybrid data mining technique used to predict mean, minimum and maximum daily temperature	Regression algorithm for continuous values	Method enabled experts to predict temperature for Patra city, tested in other regions with different climatic profile
Kohail et. al [11]	Tried to extract useful knowledge from daily weather historical data collected locally at Gaza city	All data mining techniques	Describe extracted knowledge importance in the meteorological field, used for prediction and decision making
Sivarama krishnan et. al [12]	Presented the method for prediction of daily rainfall	Association rule mining	Achieved the classification of instances as rain or no rain
AlRoby et al. [13]	Analyzed wind speed behavior	All data mining techniques	Classification is appropriate with much accuracy
Badhiye et. al [14]	Proposed design of temperature and humidity data analysis system	k-Nearest Neighbor (KNN)	Able to predict the values of temperature and humidity parameters of climate with higher accuracy
K. Pabreja [15]	To check the occurrence of cloudburst using relative humidity	K-means clustering	Achieved the estimation of cloud burst
Celik et.al [16]	Analyzed the relationships between stream flow and climatic variables	Association rules mining	Identified interesting relationships between climatic variables to the Kizilirmak River flow in turkey