An Efficient and Accurate Clustering Mechanism to Predict Weather Data

Himesh Parmar[1],Swarndeep Saket[2]
[1] Student LJIET, Ahmedabad, Gujarat, India
[2]Assistant Professor, Department of Computer Engineering, LJIET, Ahmedabad, Gujarat, India
[1]Computer Engineering,
[1]LJIET, Ahmedabad, India

Abstract:Data mining is the pattern of sorting through large dataset to identify pattern and establish relationship to solve problem through data analysis. Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. So using purposed flows we will work on clustering approach for batter weather data analysis. Reliable weather forecasting is one of the challenging tasks. One of most common difficulty is the accuracy and efficiency. In this paper we try to improve accuracy and efficiency using efficient clustering mechanism. Here accuracy of this approach is also measured.

Keywords: Clustering; Weather Forecasting; Convex-Hull; K-Means; Prediction.

I. INTRODUCTION

Meteorological data mining is a form of data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved hidden pattern can be transformed into usable knowledge. Useful knowledge can play important role in understanding the climate variability and climate prediction. 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 [6].

Weather condition can be described as the state of the atmosphere at a given time and place [7]. Weather forecasts are made by collecting quantitative data about the current state of the atmosphere. Weather forecasting entails predicting how the present state of the atmosphere will change. The main issue arise in this prediction are dimensional characters, data redundancy, missing data, skewed data, invalid data etc. To overcome this issues, it is necessary to analyze and simplify the data before proceeding with other analysis. Some data mining techniques are appropriate in this context.

To make an accurate prediction is one of the scientifically and technologically challenging problem facing by meteorologist all over the world in the last century. There are several approaches that have been used for weather prediction. This is due to mainly two factors: first, it is used for many human activities and secondly, due to the opportunism created by the various technological. In some cases, advance numerical analysis has used for weather prediction but in most of the situations clustering techniques are used for different types of predictions.

Clustering is a division of data into group of similar objects. Each group called a cluster consist of objects that are similar amongst themselves and dissimilar compare to the objects of another group. Representing data by few clusters leads to simplification of data. Clustering is the unsupervised classification of pattern into groups (clusters) [8]. In unsupervised classification, called clustering or exploratory data analysis, no labeled data are available. The goal of clustering is to separate a finite unlabeled data set into a finite and discrete set of "natural," hidden data structures, rather than provide an accurate characterization of unobserved samples generated from the same probability distribution.

1.1 Clustering procedure

It is important to understand the basic process of clustering. This has been simplified in the following flowchart. The chart shows how the process starts with given data samples and finally results into formation of clusters, their validation and finally interpretation of results [9].

1. **Feature Selection or Extraction**: In order to reduce the work load and simplify the design process feature selection or extraction is immensely important. In feature selection we have to select the most relevant attributes. Feature extraction generates new features using optimization. It utilizes some transformations to generate useful and novel features from the original ones.
2. **Design of Clustering Algorithm:** This second step generally starts with the appropriate selection of a 'corresponding proximity measure', and the construction of a criterion function. Patterns are grouped according to their resemblance with one another. All clustering algorithm are implicitly connected to define the proximity measure. Some clustering algorithm work directly on the proximity matrix. Clustering algorithms have been developed to solve different problems in specific fields. Therefore, it is important to design an appropriate clustering strategy.

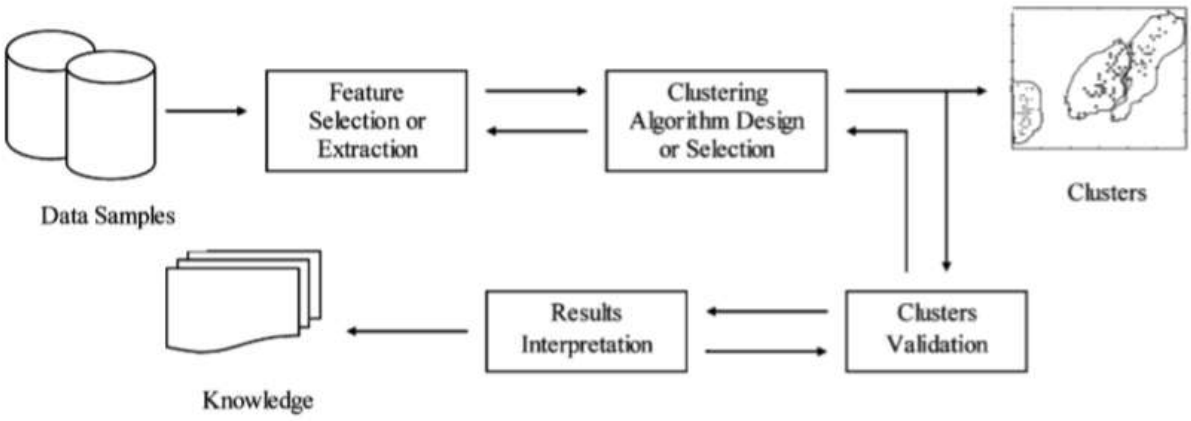

Figure 1. Clustering procedure [9]

3. **Validation of Cluster:** Different approaches lead to different clusters and are used for same algorithm or parameter identification. The correctness of clustering algorithm results is verified using predefined criteria and techniques. These assessments should be objective and have no preferences to any algorithm. It is used for finding pattern in noise.
4. **Result interpretation:** The main goal of clustering is to provide the users with meaningful insights into the original data. They can effectively solve the problems encountered. Further analyses, even experiments, may be required to guarantee the reliability of the extracted knowledge.

II. RELATED WORK

Rahul Day and Sanjay Chakraborty et al., [1] Density based clustering approach is incrementally used to predict the future weather conditions in this paper. One famous pre-processing approach, known as Convex-Hull is also used before fed the pollutant data into the clustering algorithm. This Convex-Hull method is strictly used to convert unstructured data into its corresponding structured form. These structured data is efficiently and effectively used by the DBSCAN clustering algorithm to form resultant clusters for weather derivatives. This forecasting database is totally based on the weather of Kolkata city in west Bengal and this forecasting methodology is developed to mitigating the impacts of air pollutions and launch focused modelling computations for prediction and forecasts of weather events. Here accuracy of this approach is also measured.

Shobha N andDr. Asha T et al., [2]describes a data mining study of agricultural meteorological patterns collected from meteorological centre of Bengaluru district. They use K Means and Hierarchical clustering techniques to extract patterns like minimum and maximum air temperature, relative humidity in the interim of morning hours and in the interim of noon hours, rainfall and pan evaporation which gives great significance to predict probable result. The obtained results play a crucial role in the decision making for sustainable agriculture. Along with this we also compared these algorithms by applying Connectivity, Silhouette width and Dunn index formula which measures internal validation of clustering techniques.

Dawei Wang and Wei Ding et al., [3]in this study they developed a framework for learning patterns from the spatiotemporal system and forecasting extreme weather events. In this framework, they learned patterns in a hierarchical manner: in each level, new features were learned from data and used as the input for the next level. Firstly, they summarized the temporal evolution process of individual variables by learning the location-based patterns. Secondly, they developed an optimization algorithm for summarizing the spatial regularities, SCOT, by growing spatial clusters from the location-based patterns. Finally, they developed an instance-based algorithm, SPC, to forecast the extreme events through classification. They applied this framework to forecasting extreme rain fall events in the eastern Central Andes area. Their experiments show that this method was able to find climatic process patterns similar to those found in domain studies, and our forecasting results outperformed the state-of-art model.

Zahra Karevan, Johan A.K. Suykens et al., [4] a data-driven modeling technique is proposed for temperature prediction. To investigate local learning, Soft Kernel Spectral Clustering (SKSC) is used to find similar samples to the test point to be used for training. Due to the high dimensionality. Finally, the predicted values by LS-SVMs are averaged based on the membership of the test point to each cluster. In the experimental results, the performance of the proposed method and "Weather underground" are compared and it is shown that the data-driven technique is competitive with the existing weather temperature prediction sites. For the case study, the prediction of the temperature in Brussels is considered.

Jan Skapa, Marek Dvorsky, Libor Michalek, Roman Sebesta, Petr Blaha et al., [5] deals with using a K–means clustering which is used for decision what parameter related to weather affects propagation of radio waves in mobile telecommunication network. There were analyzed parameters from a meteorological service as well as the parameters related to Global System of Mobile Communication network. For this purpose, we studied and used theory of data mining. The second part of the paper is focused on the significant weather parameters as results of K–means analyses. Consequently, there have been found some dependencies between weather conditions and receive level using a mathematical tools of correlation analysis via MATLAB.

III. PROPOSED WORK

Accurate weather forecasting is one of the challenges in climate informatics. It involves reliable predictions for weather elements like temperature, humidity, and precipitation. Weather Forecasting is an approach which is used to forecast the weather based on the previous or current weather conditions, for a particular region and particular time period, using some science, algorithm and technology. The main objective of an efficient clustering mechanism is to improve accuracy and efficiency of weather forecast prediction.

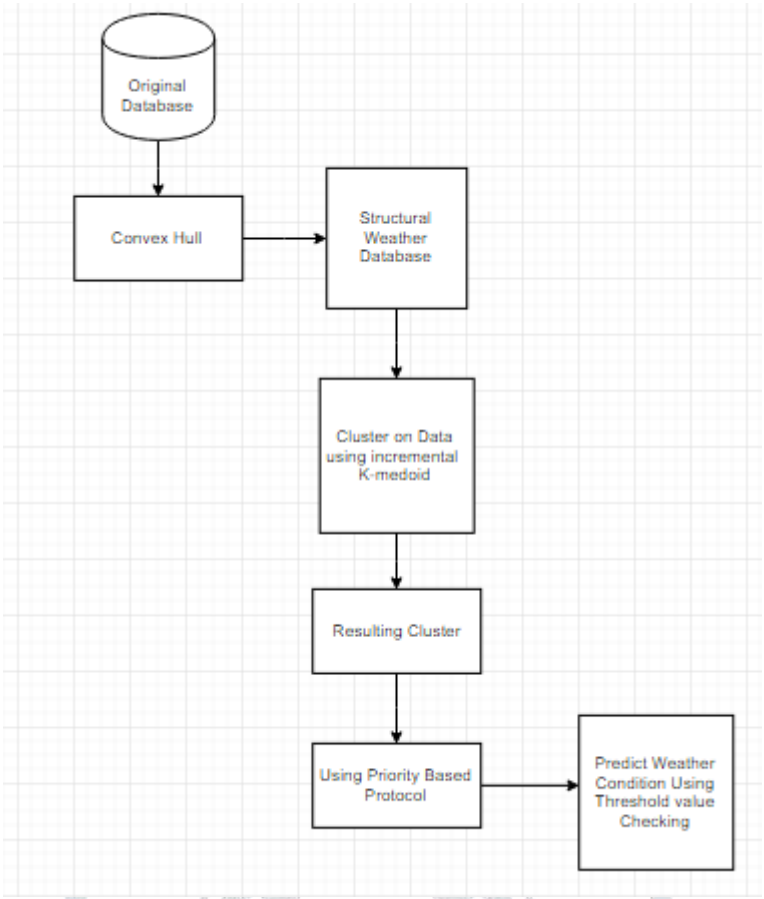

Figure 2: Proposed Method

In order to enhance accuracy and performance of data using k-medoid clustering technique has been proposed and following are the steps.

1. Collect weather data of recent year and store them into the Database.
2. Use Convex-Hull to get structural data. It includes all extreme data and then store them into 'structural Weather Database'.
3. Structural weather database splits into parts on the basis of day and night and rain flow ratio.
4. Apply K-medoid Clustering to create clusters using structural data.
5. Finally, find the resulting clusters
6. Then the priority based protocol is used on those resulting clusters to give the weather prediction on the basis of last three years.
7. From the final result, the probable weather conditions and also a Rain flow and storm can be predicted

3.1 K-MEDOID CLUSTERING

K-Medoids clustering is one such algorithm. Rather than using conventional mean/centroid, it uses medoids to represent the clusters. The medoid is a statistic which represents that data member of a data set whose average dissimilarity to all the other members of the set is minimal. Therefore a medoid unlike mean is always a member of the data set. It represents the most centrally located data item of the data set.

The working of K-Medoids clustering algorithm is similar to K-Means clustering. It also begins with randomly selecting k data items as initial medoids to represent the k clusters. All the other remaining items are included in a cluster which has its medoid closest to them. Thereafter a new medoid is determined which can represent the cluster better. All the remaining data items are yet again assigned to the clusters having closest medoid. In each iteration, the medoids alter their location. The method minimizes the sum of the dissimilarities between each data item and its corresponding medoid. This cycle is repeated till no medoid changes its placement. This marks the end of the process and we have the resultant final clusters with their medoids defined. K clusters are formed which are centeredaround the medoids and all the data members are placed in the appropriate cluster based on nearest medoid.

Input:

- k: number of clusters
- D: the data set containing n items

Output:

- A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoids

Method:

1. The algorithm begins with arbitrary selection of the K objects as medoid points out of n data points (n>K).
2. After selection of the K-medoid points, associate each data object in the given data set to most similar medoid.
3. Randomly select non-medoid object O.
4. Compute total cost, S of swapping initial medoid object O.
5. 5. If S>0, swap initial medoid with the new one. 6. Repeat steps until there is no change in the medoid.

IV. RESULT ANALYSIS

Experiment of proposed method is executed on computer having Intel (R) Core (TM) i5-5200U CPU@2.20GHz with 4GB RAM having Windows 10(64 bit) operating system.The parameter selected for result analysis of proposed system is accuracy and efficiency.In this section accuracy and efficiency is calculated and compared with other researchers developed method

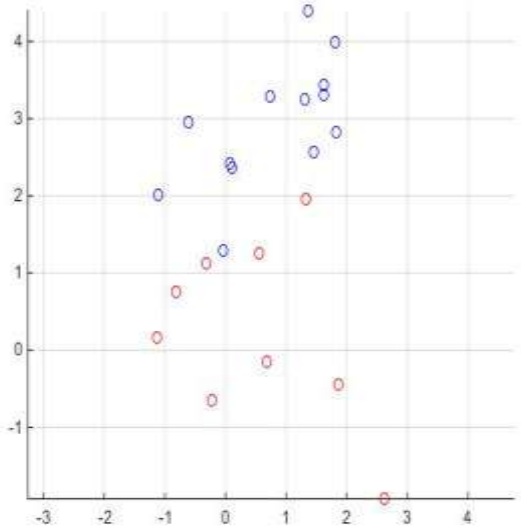

Figure 3: Temperature Cluster Data for Month Sep. 2106

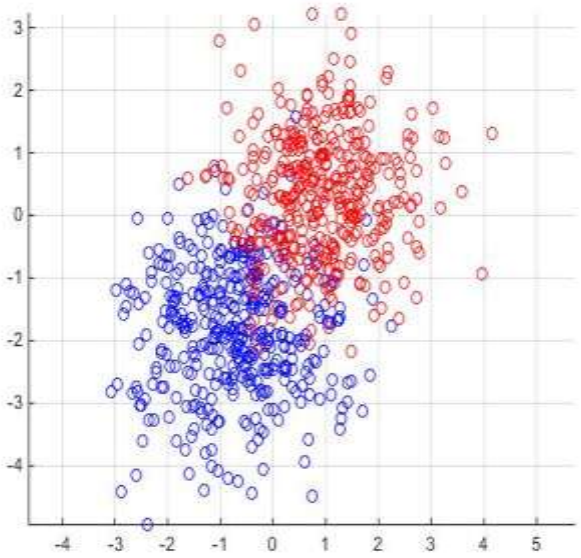

Figure 4: Rain fall Cluster Data for Month Sep. 2106

Date	Temperature	Humidity
04/09/16	22.1530	0.5917
05/09/16	16.8387	0.8729
06/09/16	17.2896	0.8133
07/09/16	21.4484	0.7304

Table 1: Resultant Table of Temperature and Humidity

Year	Rainfall
2016	335.9
2015	566.5
2014	771.3
2013	513.1

Table 2: Resultant Table of Rainfall

Date	Probable temp range °C	Actual temp °C
04/09/16	15-23	22
05/09/16	15-23	18
06/09/16	15-23	17
07/09/16	15-23	21

Table 3: Predict Weather condition of Temperature

$$\text{Equation for calculating Accuracy} = \frac{\text{Number of matched records}}{\text{Total number of records}} \text{ X } 100$$

Compression Algorithm	Execution time	Accuracy
DBSCAN	2.199835	74.5%
K-Medoid	0.883495	78%

Table 4: Comparison Tableof both Algorithm

V. CONCLUSION

The overall goal of data mining process is to extract information from a large data set and transfer it into an understandable form for future use. Clustering is important in data analysis and data mining applications. Clustering is a division of data into group of similar objects. Clustering can be done by the different algorithms such as hierarchical- based, partitioning-based, grid-based and density-based

algorithms. In this paper, a new technique is introduced to predict the weather of upcoming days with the help of incremental K-Medoid clustering algorithm.

VI. REFERENCES

[1] RatulDey, Sanjay Chakraborty" Convex-Hull & DBSCAN Clustering to Predict Future Weather", 978-1-4799-6908-1/15 31.00 ©2015 IEEE, Year: 2015, Pages: 1-8.

[2] Shobha N, Dr. Asha T" Monitoring Weather based Meteorological Data: Clustering approach for Analysis",978-1-5090-5960-7/17 31.00 ©2017 IEEE, Year: 2017, Pages: 75-81.

[3] Dawei Wang, Wei Ding," A Hierarchical Pattern Learning Framework for Forecasting Extreme Weather Events",1550-4786/15 31.00 © 2015 IEEE, Year: 2015, Pages: 1021-1026.

[4] Zahra Karevan, Johan A.K. Suykens" Clustering-based feature selection for black-box weather temperature prediction", 978-1-5090-0620-5/16 31.00 2016 IEEE, Year: 2016 Pages: 2722-2729.

[5] Jan Skapa, Marek Dvorsky, Libor Michalek, Roman Sebesta, Petr Blaha" K-means Clustering and Correlation Analysis in Recognition of Weather Impact on Radio Signal", IEEE International Conference on Information Technology, Year: 2012, Pages: 316-319.

[6] Meghali A. Kalyankar, Prof. S. J. Alaspurkar, "Data Mining Technique to Analyse the Metrological Data", International Journal of Advanced Research in Computer Science and Sofrware Engineering, Year: February 2013, Volume 3, Issue 2, Pages: 119-122.

[7] K. Mumtaz, Dr. K. Duraiswamy, "A Novel Density based improved k-means Clustering Algorithm Dbkmeans", International Journal on Computer Science and Engineering, 2010, 213-218, ISSN: 0975-3397 213 Vol. 02, No. 02, Pages: 23-27.

[8] MeghaMandloi, "A Survey on Clustering Algorithms and K-Means", IJRETM-2014-02-04-514, Pages: 1-5.

[9] Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms", 1045-9227/$20.00 © 2005 IEEE, Year: 2005, Pages: 645-677.

[10] Ravi Shankar Sangam, Hari Om, "The k-modes algorithm with entropy based similarity coefficient", International Symposium on Big Data and Cloud Computing, 2015 Science Direct, Year: 2015, Pages: 93-98.

[11] Er. Arpit Gupta, Er.Ankit Gupta, Er. Amit Mishra," RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS", International Journal of Advance Technology & Engineering Research, Year: November 2011, Vol. 1, Issue 1, Pages: 39-47.

[12] Vaibhavi Mistry, Vibha Patel "Weather Condition Prediction Using Semi-Supervised Data Mining Technique",International Journal of Engineering Trends and Technology (IJETT), Year: Feb 2015, Pages: 179-183.

[13] https://en.wikipedia.org/wiki/K-medoids.html,12/10/2017,4.30 PM.

[14] http://www.zentut.com/Data Mining,12/10/2017,4.30 PM.