Overview of Clustering Algorithm for Weather Data

Himesh Parmar[1], Swarndeep Saket [2].

[1]Student (*Master of Engineering*), *Computer Engineering,* L.J. Institute of Engineering and Technology *, Gujarat, India.*

[2] Assistant Professor, Computer Department; Technology Department, L.J. Institute of Engineering & Technology
, Gujarat, 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.

Keyword :-Clustering; Weather Forecasting; Convex-Hull; DBSCAN; K-Means.

1. INTRODUCTION

Forecasting is very important for prediction of the future events. Science and computer technology together has made significant advances over the past several years and using those advanced technologies and few past patterns, it grows the ability to predict the future [6]. 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 [1]. Accurate weather forecasting is one of the challenges in climate informatics. It involves reliable predictions for weather elements like temperature, humidity, and precipitation [4]. Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc.

2. WORKING TECHNOLOGY

2.1 BASIC CONCEPT

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) [7]. 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.

2.2 WORKING MECHANISM

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.

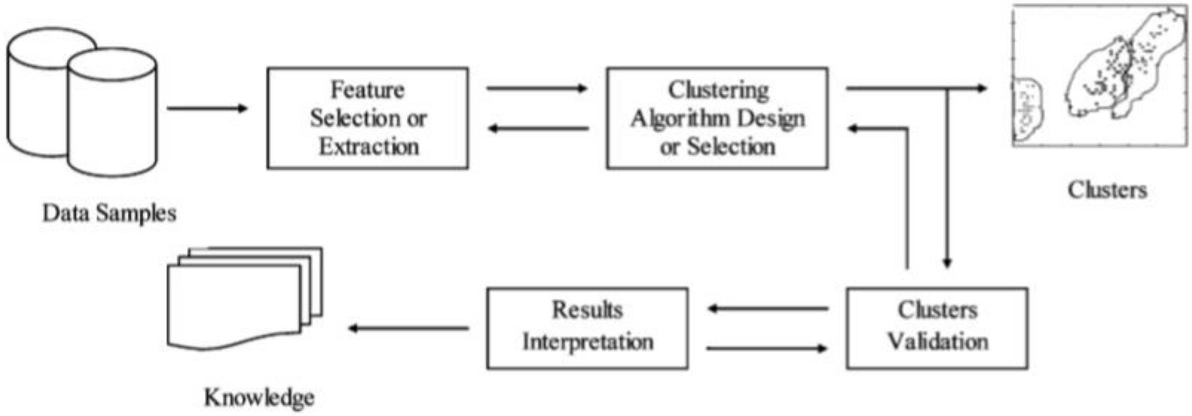
Data Samples
Feature Selection or Extraction
Clustering Algorithm Design or Selection
Clusters
Clusters Validation
Results Interpretation
Knowledge

Fig.1. Clustering procedure [8].

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.

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.

2.3 APPLICATIONS OF CLUSTERING:-

- **Marketing**: Finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- **Biology:** Classification of plants and animals given their features;
- **Libraries:** Book ordering;
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- **City-planning:** Identifying groups of houses according to their house type, value and geographical location;
- **Earthquake studies:** Clustering observed earthquake epicenters to identify dangerous zones;

- **WWW:** Document classification; clustering weblog data to discover groups of similar access patterns.

2.4 CLASSIFICATION OF CLUSTERING ALGORITHMS:-

Clustering is a division of data into groups of similar objects. Data objects are represented by few well defined clusters. Major clustering techniques can be classified into the following categories.

1) **Partitioning Methods:** Perhaps the most popular class of clustering algorithms is the combinatorial optimization algorithms and iterative relocation algorithms. These algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained. In a basic iterative algorithm, such as K-means- or K-Medoids, convergence is local and the globally optimal solution cannot be guaranteed. Because the number of data points in any data set is always finite and, thereby, also the number of distinct partitions is finite, the problem of local minima could be avoided by using exhaustive search methods [7].
2) **Hierarchical Clustering:** The hierarchical method group data instances into a tree of clusters. There are two major methods under this category. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are closed to one another, until all of the groups are merged into one (the top most level of hierarchy), or until a termination condition holds. The divisive approach, also called the top down approach, starts with all the objects in the same cluster. In each successive iteration a cluster is split up into smaller clusters, until each object forms a cluster, or a termination condition holds [7].
3) **Density based Clustering:** Density based clustering algorithm try to find clusters based on density of data points in s region. The key idea of density based clustering is that for each instance of cluster the neighborhood of a given radius (Eps) has to contain at least minimum number of instances (MinPts). One of the most well-known density-based clustering algorithm is DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Other density based clustering methods are OPTICS (Ordering Points to Identify the Clustering Structure) and DENCLUE (Density-based clustering).
4) **Grid-Based Methods:** The grid-based clustering approach uses a multiresolution grid data structure. It quantizes the object space into finite number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects, yet dependent on only the number of cells in each dimension in the quantized space [7].

3. MODULATION TECHNIQUES:

A. 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 centered around 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.

4. RELATED WORKS:

4.1 LITERATURE REVIEW

1) Convex-Hull & DBSCAN Clustering to Predict Future Weather[1]:

Density based clustering approach is incrementally used to predict the future weather conditions in this paper. One famous preprocessing 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 modeling computations for prediction and forecasts of weather events. Here accuracy of this approach is also measured.

2) Monitoring Weather based Meteorological Data: Clustering approach for Analysis[2]:

This paper describes a data mining study of agricultural meteorological patterns collected from meteorological center of Bengaluru district. They use K Means and Hierarchical clustering techniques to extract patterns like minimum (15 to 17°C) and maximum (28 to 29°C) air temperature, relative humidity (79 to 96%) in the interim of morning hours and (42 to 50%) in the interim of noon hours, rainfall (0 mm) and pan evaporation (5.22 to 7.2 mm) 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. The results indicate that Hierarchical technique performs better than K Means in terms of Connectivity (10.1671), Silhouette width (0.4084) and Dunn index (0.4619).

3) A Hierarchical Pattern Learning Framework for Forecasting Extreme Weather Events[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.

4) **Clustering-based feature selection for black-box weather temperature prediction[4]:**
In this paper, 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.

5) **K-mean Clustering and Correlation Analysis in Recognition of Weather Impact on Radio Signal[5]**:
This paper 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.

4.2 COMPARATIVE TABLE

Sr. No.	Paper Title	Method Used	Advantages	Disadvantages
1	Convex-Hull & DBSCAN Clustering to Predict Future Weather	Convex-Hull method and DBSCAN clustering	Predict some certain factors of weather, like temperature, humidity	Not predict rain fall, storm etc.
2	Monitoring Weather based Meteorological Data: Clustering approach for Analysis	K Means and Hierarchical clustering	Useful for predict weather parameters	More time consuming to predict.
3	A Hierarchical Pattern Learning Framework for Forecasting Extreme Weather Events	Location based patterns and spatial clusters	Useful for forecasting extreme weather events	Not predict other weather feature.
4	Clustering-based feature selection for black-box weather temperature prediction	Soft Kernel Spectral Clustering and Least Squares Support Vector Machines	Useful for temperature prediction	Not more efficient
5	K-mean Clustering and Correlation Analysis in Recognition of Weather Impact on Radio Signal	K-means clustering	Weather impact on Radio Signal	Not useful for overlap frames

4.3 CURRENT ISSUES:-

Clustering is one of the major topics of research related to data mining. Though lots of research in this area has been done, the topic is yet to be explored fully. Clustering is still in primary stage of research. This is basically because clustering is unsupervised learning and so the work starts without any assumptions and therefore, the scope for research increases. The main theme of the present research is to improve accuracy and efficiency for weather data and generate better and stable clusters.

5. CONCLUSIONS:

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.

6. ACKNOWLEDGEMENT:

The authors are very much grateful to Department of Computer Engineering, L.J Institute of Engineering & Technology, Ahmedabad, from Gujarat Technological University; for giving opportunity to do research. Two authors Himesh Parmar, and Prof. Swarndeep Saket (Jain) are also grateful to Dr. A. C. Suthar, Principal of L.J Institute of Engineering & Technology, Ahmedabad, from Gujarat Technological University; for giving constant encouragement to do research work in the Department.

7. REFERENCES:

[1] Ratul Dey and Sanjay Chakraborty,"Convex-Hull & DBSCAN Clustering to Predict Future Weather", 978-1-4799-6908-1/15 31.00 ©2015 IEEE.

[2] Shobha N and Dr. Asha T, "Monitoring Weather based Meteorological Data: Clustering approach for Analysis", p.no.-75-81, 978-1-5090-5960-7/17 31.00 ©2017 IEEE.

[3] Dawei Wang and Wei Ding, "A Hierarchical Pattern Learning Framework for Forecasting Extreme Weather Events", p.no.-1021-1026, 1550-4786/15 31.00 © 2015 IEEE.

[4] Zahra Karevan and Johan A.K. Suykens, "Clustering-based feature selection for black-box weather temperature prediction", p.no.-2722-2729, 978-1-5090-0620-5/16 31.00 2016 IEEE.

[5] Jan Skapa, Marek Dvorsky, Libor Michalek, Roman Sebesta, and Petr Blaha, "K-mean Clustering and Correlation Analysis in Recognition of Weather Impact on Radio Signal" ,p.no.-316-319,IEEE-2012.

[6] Sanjay Chakraborty, Prof. N.K.Nagwani and Lopamudra Dey, "Weather Forecasting using Incremental K-means Clustering", Published 2012 in ArXiv.

[7] Megha Mandloi, "A Survey on Clustering Algorithms and K-Means",p.no.-1-5, IJRETM-2014-02-04-514.

[8] Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms",p.no.-645-677, 1045-9227/$20.00 © 2005 IEEE.

[9] Archna Kumari, Pramod S. Nair and Sheetal Kumrawat, "An Enhanced K-Medoid Clustering Algorithm",p.no.-27-31, IJRITCC | June 2016

BIOGRAPHIES

	Himesh Parmar received the B.E. degree in Computer Sciences and Engineering from Gujarat Technological University in 2016 and student of M.E. in Computer Engineering, L.J Institute of Engineering & Technology, Ahmedabad, from Gujarat Technological University; Currently he is doing research work in Data Mining.
	Swarndeep Saket is *Assistant Professor* at L.J Institute of Engineering & Technology, Ahmedabad, from Gujarat Technological University; more than 2 years' experience in teaching.