Indian Journal of Science and Technology, Vol 8(14), DOI: 10.17485/ijst/2015/v8i14/73049, July 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

# Modified Projected Space Clustering Model on Weather Data to Predict Climate of Next Season

**S. Gokila[1*], K. Ananda Kumar[2] and A. Bharathi[3]**

[1]Bharathiar University, Coimbatore - 641 046, Tamil Nadu, India; goks_do@yahoo.co.in
[2]Department of MCA, Bannariamman Institute of Technology, Erode - 638401, Tamil Nadu, India; anandakumark@bitsathy.ac.in
[3]Department of IT, Bannariamman Institute of Technology, Erode - 638401, Tamil Nadu, India; bharathia@bitsathy.ac.in

## Abstract

**Objectives:** Four seasons of Indian weather are interdependent. Prediction of seasonal weather supports many fields to work successfully. The objective of proposed model to work on Weather Pattern identification in Initial phase of prediction which has to include unequal weight of attributes. Methods: The projected space clustering model is suitable to handle the non-sequence patterns of data set. The existing projected space clustering eliminates the least weighted attribute. The framework suggested in this paper incorporate modified projected space cluster which work on complete set of attributes to form pattern wise clusters which is dynamic in number for each season. Next part of framework is seasonal weather prediction using ANN, works on dynamic set of clusters. **Findings:** Dynamic nature of clusters formed in modified projected space clustering completely eliminates the error rate arise because of fixed number of cluster. The extreme events patterns formed as a separate clusters are not eliminated as outline. The result of these clusters gives the study report of each season, like the changes of climate pattern, the frequency of extreme event and weather prediction of next season. **Application/ Improvement:** The modified projected space clustering work well on unequal complete set of attributes to form a cluster of different pattern. For each duration numbers of clusters are dynamic based on the pattern variation in climate data.

**Keywords:** Climate, Data Mining, Dynamic Clustering, Forecasting, Projected Space, Weather, Weather Season

## 1. Introduction

Climate plays major role in deciding future of all the sectors because it is used for many of human activates. Predicating accurate climate of future is a challenging job for all the climate scientists. Weather is day-to-day variation in a particular region, whereas the climate is a long term fusion of the variation. The weather conditions are obtained from automated weather stations, ground observation, Doppler radar, aircraft, sensors and satellites. Weather data includes temperature, Wind Speed, Evaporation Radiation, Sunshine, Cloud Form, Humidity, Precipitation and Rain fall. Weather data are generally classified as synoptic data for climate data and used in weather forecast models (Mathematical calculations). Climate data are official data provided after some quality control on synoptic data. Weather varies for time to time and for each region. In data mining work weather data can be include in spatio-temporal data sector[4]. As the nature of region varies the quality control on weather considers nature of the region to create official climate data from weather data. The nature of region is predicted based on the latitude and longitude in which it is located[1].

Meteorological departments apply many mathematical models on weather data to predict future climate. The mathematical models are the equation to be solved predicts some value. The models are run with the help of efficient computers. The forecasting charts are the

* Author for correspondence

analysis result of mathematical model which are in visuals. Forecasting accuracy of model is good only for short term prediction. The accuracy falls off for long term because the long term calculation works on large weather data set which includes many attributes and more variation of readings. There the data mining techniques used to do either descriptive mining (describe general properties) or predictive mining (attempt to predict based on inference of data) on large volume of data to provide accurate forecasting even for long days and accurate prediction about climate for long term. The Figure 1 expresses this work flow of weather prediction and the roll played by data mining in that.

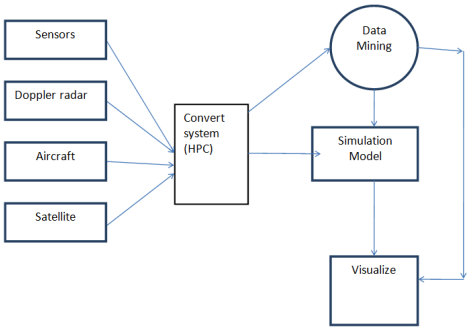


**Figure 1.** Data Mining in Weather Predication.

## 1.1 Clustering

Cluster analysis is exploring the structure of data. Core Cluster analysis is a clustering. Clustering analysis in a data is a unknown label class (unsupervised)[3,11]. So it is learned by observation not learned by example[2]. Clustering divide the data set into classes using the principle of "Maximum intra class similarity and Minimum inter class similarity". It doesn't have any assumption about the category of data. The basic clustering techniques are Hierarchical, Partitional, Density based, Grid based and Model based clustering. Some sort of measure that can determine whether two objects are similar or dissimilar is required to add them into particular class. The distance measuring type varies for different attribute type. Clustering can also used to detect outline in data which may occur due to human error or some abnormal events occurred while creating data set[2,11]. Cluster work well on scalable, heterogeneous and high dimensional data set. In all the clustering algorithms user defined parameters are given as input to find either similarity, dissimilarity among clusters and for root attribute of cluster and for maximum or minimum number of clusters.

## 1.2 Projected Space Clustering

In some data set the clusters compared may vary in attributes. The projected space clustering filter the less important dimension[8,9]. Treat that as outline and eliminate that dimension. The clustering done with the remaining data set. In that the clusters varies in attributes. The qualities of clusters are identified and that are optimized. Cluster formation done using any of the basic clustering method discussed above.

The variation in the basic projected space clustering is important in data set like weather data. Because the dimension removed as outline may influence a predication of climate. For example during the summer season the rain attribute may have less reading but that is also one of the value decides the climate of next season. So the attributes must be kept as such even it is less important in that cluster. Modified projected space clustering suggested paper is necessary to ensemble the data with all level of attributes.

In[4] reported Spatio-Temporal pattern in climate data using clustering. Cluster the climate data reduces the computational complexity. Climate data pre-processed to eliminate anomaly and the outline. Spatial similarity algorithm creates a cluster of similar weather stations on 46/77 latitude/longitude with similar climate behavior. In[5] used data mining techniques to find the increase and decrease in global temperature. One of the aims of climate analysis is to find the increase or decrease in global temperature. K-means clustering algorithm used to group data set with minimum temperature and also applies J48 Classification algorithm to fine the suitable attribute to split the data set. The size of K=5. The data set taken for study is 112 years long. In[6] applied K-means algorithm with the size of k=4 applied on nine years of data to form a cluster. Seasonal study is of great help for agriculturalists; the clusters used for summer, rainy, spring and autumn. The study of the clustering analysis was done for variation in rain, temperature, humidity and wind speed compare to same season of each nine years. Such study helps to predict the climate of all the four seasons in near future. All these clustering required user input to decide the number of cluster if it is KNN or K Means clustering method. When the clustering is decision tree base the initial attribute selection plays a major roll. These problems are talked in automated projected space clustering.

In[7] developed SUBCAD projected space clustering algorithm, in which the data from data set are selected in

sampling technique to form a initial set of clusters. The remaining data are added into any of the initially formed cluster in which the new one have less impact on cluster quality. Later the clusters are pruned to optimize. The problem in this method is to identify the initial clusters with all attributes. In[8] found graph partition algorithm called CLICK. The weight of each attribute is calculated. The graph based partition of data by applying weighted attributes. The attributes are vertex of graph the edge between the vertex. The problem with this CLICK is handling high dimension data. This also eliminates the less weight attributes. Needs the user input to find threshold values in calculating attribute weight. In[9] developed AT-DC algorithm based on decision tree clustering. AT-DC forms a single top level cluster with entire data set. Then the sub clusters are formed. The sub clusters are accepted when thequality higher than the original clusters formed in previous step. AT-DC handles the high dimensional data automatically that is without user input to form a projected space cluster, but it doesn't handle outliner in data set.

In[10] developed PROCAD algorithm to ensemble data in projected space. It handles high dimension data and also the outline without user input. In the initial state the attribute weight are calculated based on which the clusters are formed. The quality of cluster is analyzed as good when it contains desired number of dimension and data point. Even in this PROCAD algorithm attribute with less important are not considered to form cluster. And also all these algorithms work only for categorical data not for numerical data. The modified model of projected space cluster in this paper suggest the procedure form a clusters of one season data pattern with required amount of dimension even few attribute are less in requirement which decides the weather of next season. Here the data is completely numerical.

## 2. Modified Projected Space Clustering Model

The model separates the entire process into three phases. The phase one starts with cluster formation of each season base on the patterns in weather data, second phase is to find the extreme event in each season from the clusters formed in first phase of model, the third phase is to identify the cluster which relate to the weather of next season weather data.

The model suggested here is to work specifically on Indian weather data set. Naturally India weather segregated into four seasons. The clustering has to be done for all the four seasons (Winter December to February, summer or Pre Monsoon – March to May, Rainy or Monsoon – June to August, autumn or Post Monsoon – September to November) each of three months long[13]. End of clustering four set of clusters of each seasons is derived. The cluster formation is expressed in Figure 2. Clustering doesn't take any input from the user. The patterns of cluster including all the attributes are in projected space. The first day of pattern is formed as initial cluster. To find the cluster of next day pattern, distance between the existing clusters pattern and the new day pattern is calculated. The new day pattern is included in the cluster which gives distance less than the threshold value. If no distance is within threshold then the new cluster with new data is formed. This process is continued for all the day's data point in one season and for all the four seasons separately.

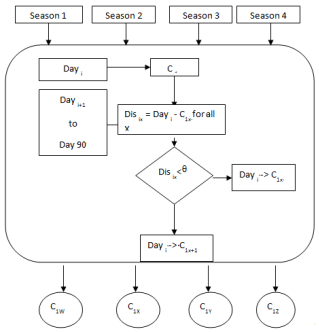


**Figure 2.** Seasonal Cluster.

End of clustering $C_{1W}$ (Clusters of Season 1), $C_{2X}$ (Clusters of Season 2), $C_{3Y}$ (Cluster of Season 3) and $C_{4Z}$ (Clusters of Season 4). Where W, X, Y, Z are number of clusters in each season. So the number of cluster is dynamic, no data point is omitted from clustering, many different patterns in particular season can be found, number if occurrences of extreme events can be found.

The extreme high and low events in each season can be identify from the cluster with less data point in it. This cluster also include for the season vice comparison even as extreme. Because these may have some interdependency either with previous season cluster or with next season clusters. And it is an alert to the application area when the same is repeated in almost all the years of same season. This identification can be takes places for $C_{1W,}$ $C_{1X,}$ $C_{1Y,}$ and $C_{1Z}$ clusters separately.

The main aim of meteorological is accurate weather prediction for many applications. One sub part of this prediction is analyzing of weather seasonal dependency. The clusters of four seasons are cross compared to find the particular pattern of cluster which has more influence in climate of next season. The Prediction using Neural Network gives accurate result when compared with traditional algorithms[12,14].

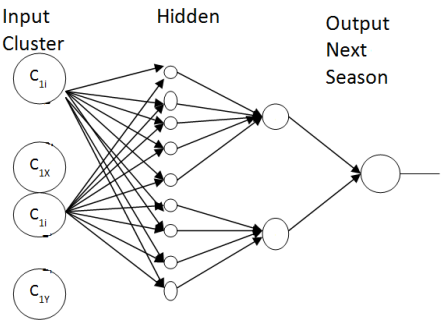


**Figure 3.** ANN of Season Comparison.

The ANN base model can be trained to predict the weather of next season by identifies impact of the cluster in precious season. Here the number of cluster in each season differs. Each Cluster of one season is compared with all the clusters of other season to identify the pattern which influences the climate relevance of immediate next season. The Figure 3 expresses above explanation of third phase of projected space model. The input later is the clusters set of two seasons, for example $C_{1X}$ and $C_{1Y}$ are the clusters of two seasons. These set are unequal in number of clusters. Each of which is compared in hidden layer to predict exact pattern of $C_{1X}$ which have major influence in climate of next season.

## 3. Discussion of Model

Few domains like weather data consists of seasonal vice result dependency. Especially in India the climate of one season mostly depends on the previous season. This is like a chain activity. The attributes recorded in this domain is uniform in all the season but the importance of attribute in one will not have same level of importance in another season and vice versa. And the season pattern of weather will also vary. Less and more variation in pattern may also influence the climate, so no pattern or dimension to be eliminated in forming season wise patterns. In this case the clusters of different seasons to be compared have unequal weight of complete dimensional attribute.

Projected space clustering using in similar kind of domain is modified in this model to handle the clusters of projected space without removing any of the lease weighted attribute as outline. The model clusters the similar patterns in different clusters. Number of clusters dynamic to cover all possible patterns of weather. So the number of clusters handled for same season differs in each year. The analysis of dynamic number clusters identifies the variation in weather of each year and also the extreme events. The reason of weather variation and the influence of extreme event in this variation can be compared. In this comparison the dimension of attributes in clusters will be equal. Inter seasonal clusters with different attribute dimensional are back propagated using ANN trained to predict the weather of next dependent season. The suggested model will perform well in handling clusters of unequal weight of attributes.

## 4. Conclusion

Climate is not fixed; the fluctuation in the climate can be seen from year to year. Data mining application can help meteorological to create faster forecast and decisions and provide more performance and reliability than any other methods. Clustering techniques applied on climate data helps to produce similar pattern of climate with the consideration of spatial nature. Projected Space Clustering is good on non-contiguous numerical time series data. But the same needed some modification to handle all the attributes of data point without the omitting single unequal attribute. The new modified model of projected space cluster will solve the problem of handling unequal attribute and also the uneven number of clusters to predict the climate of particular season and also to identify the extreme event in any of four seasons in Indian weather data. The model can be enhanced further by changing the threshold value dynamically based on the previous year actual weather report of meteorological department in fixing threshold values. This is required because of changing nature of climate.

## 5. References

1. Badhiye SS, Wakode BV,Chatur PN. Analysis of temperature and humidity data for future value prediction. International Journal of Computer Science and Information Technologies (IJCSIT) 2012; 3(1):3012–14.

2. Han J, Kamber M. Data mining: concepts and techniques. San Francisco, CA: Morgan Kaufmann; 2006. p. 5. ISBN: 978-1-55860-901-3.
3. Soman KP, Diwakar S, Ajay V. Insight into Data Mining Theory and Practice. Delhi: PHI Learning; p. 7. ISBN: 978-81-203-2897-6.
4. Levy D. Spatio Temporal Pattern Detection in Climate Data. ITiCSE. 2013; 1(4):67–71.
5. Rajinikanth TV, Balaram VVSSS, Rajasekhar N. Analysis of Indian weather data sets using data mining techniques. In: Nagamalai D et al. (Editor). ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS. 2014; 1(1)89–94.
6. Kohail SN, El-Halees AM. Implementation of data mining techniques for meteorological data analysis (a case study for gaza strip). International Journal of Information and Communication Technology Research. 2011 Jul; 1(3):96–100.
7. Gan G, Wu J. Subspace clustering for high dimensional categorical data. ACM SIGKDD Explor. 2004; 6(2):87–94.
8. ZakiMohammed J, Peters M, Assent I, Seidl T. CLICKS: an effective algorithm for mining subspace clustersin categorical datasets. Data & Knowledge Engineering. 2007; 60(1):51–70.
9. Cesario E, Manco G, Ortale R. Top-down parameter-free clustering of high-dimensional categorical Data. Knowledge and Data Engineering, IEEE Transactions. 2007; 19(12):1607–24.
10. Bouguessa M. Clustering categorical data in projected spaces. Data Mining and Knowledge Discovery. 2015; 29(1):3–38.
11. Zaki MJ, Wagner Meira JR. Data mining and analysis fundamental concepts and algorithms. First Edition. New York: Cambridge University Press; 2014. ISBN: 978-0-521-76633-3.
12. Kang MY, Shin J-D, Kim B. Automatic subject classification of korean journals based on kscd. Indian Journal of Science and Technology. 2015 Jan; 8(1):452–56.
13. Attri SD, Tyagi A. Climate profile of India. New Delhi: India Meteorological Department, Environment Monitoring and Research Centre; 2010.
14. Gharehchopogh FS, Khaze SR, Maleki I. A new approach in bloggers classification with hybrid of k-nearest neighbor and artificial neural network algorithms. Indian Journal of Science and Technology. 2015 Feb; 8(3):237–46.