



K-means Cluster Model for Climate Prediction of Vindhya Region

Shailendra Singh¹, Navita Shrivastava², R. K. Tiwari³

^{1,2}Awadhesh Pratap Singh University, Sirmour Road, Ananthpur, Rewa,
Madhya Pradesh 486003, INDIA

³Govt. New Science College, Rewa, Madhya Pradesh 486003, INDIA

Abstract: In the present work an attempt has made to develop a novel predictive model that uses the historical climate data (rain, wind, speed, temperature, dew point etc.) for prediction. The present study describes how to use a data mining technique and how to develop a system that uses numeric historical data to forecast the climate of vindhya region. Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as preprocessing steps for other data mining algorithms operating on the detected clusters. In this paper clusters are made by putting similar data into four groups. It is an unsupervised learning technique with the combination of various algorithms. Using k-means clustering algorithm we have developed clustering model for the training and test data set. The association between different climate attributes on the test data set i.e rainfall has been observed. The obtained cluster instances for the training data sets validate our model and clearly show the association of the climate attributes under taken on rainfall.

Keywords: k-means, unsupervised learning, data distribution, forecast.

I. Introduction:

Climate forecasting is a way by which scientists are using to predict climate change. To produce these forecasts an extensive suite of computational forecasting tools have been developed including a multi model ensemble approach that required thorough validation of each model's accuracy level in simulation inter-annual climate variability. The detection and analysis of predictive models is also part of statistics and machine learning^[1]. The choice of the predictor and response variables is relatively open in the data mining application. Through a lot of work has been done in this field of climate informatics in which various seasonal climate attributes have been used to develop the predictive models, but still there is a need to develop more robust, accurate, reliable and efficient model for climate prediction^[2-3]. An artificial neural network is a computing method inspired by structure of brains and nerves system. A typical neural network consists of a group of inter-connected processing units, which are also called neurons. Each neuron makes its independent computation based upon a weight sum of its inputs and passes the results to other neurons in an organized fashion^[4]. A neural network is trained from training data set. This makes neural network a desirable tool in dealing with complex systems.

The process of finding useful patterns and information from raw data is often known as knowledge discovery in database. 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. In hierarchical clustering, each data point is initially in its own cluster and then clusters are successively joined to create a clustering structure^[5-6]. This is known as the agglomerative method. In partition clustering, the number of clusters must be known apriori. The partitioning is done by minimizing a measure of dissimilarity within each cluster and maximizing the dissimilarity between different clusters.

II. K-Means Clustering

Clusters are natural grouping of data items based on similarity metrics or probability density models. K-means algorithm organizes objects into k-partitions where each partition represents a cluster. It is a partition method which finds mutual exclusive clusters of spherical shape. It generates a specific number of disjoint, non-hierarchical clusters.^[7] Clustering algorithms produces high quality clusters to yield lower inter cluster similarity a high intra cluster similarity. k-means algorithm achieve this on vast and highly structured data. But it has difficulties in handling unstructured natural data which often contain outlier data points. The efficiency of the algorithm has to be further tested on a comparatively larger data set.^[8] The outcome of the extracted data can be analyzed for the future planning and development perspectives.

III. Data and Method of analysis

The data used for this work was collected from the web (www.ipcc-data.org) and national climatic data center (NCDC) for vindhyar region. A monthly climate data for the period of 1901-2010 were collected and some data preprocessing has been applied in raw data sets. The flow chart for the proposed model is shown in the figure-1.

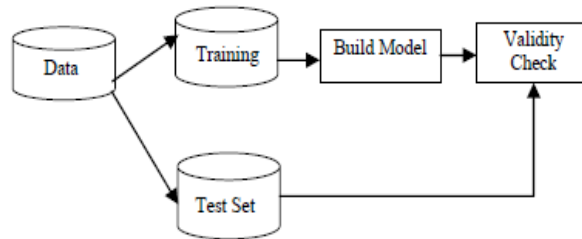


Fig. 1- Proposed model for prediction

Technique/ algorithms for proposed model along with appropriate data set have been identified for the development of the proposed model. Extraction, transformation and cleaning of the data set for desired parameter or attributes have been performed. We start out with initial set of means and classify cases based on their distances on their centers. We compute the cluster means using the cases that are assigned to the clusters. Then we reclassify all cases based on the new test set of means. This step repeats until cluster means do not change between successive steps. Finally we calculate the means of cluster once again and assign the cases to their permanent cluster.^[9]

The data set were inputted in as per the k-means algorithms in the WEKA. We have received clusters centroids for different attributes of the given data sets. The cluster instances for the cluster 0 and cluster 1 can be viewed from the run information of the model.

IV. Result and Discussion:

In the present analysis we used clustering method for the climate prediction of vindhya region for longitude 24.53° N and Latitude 81.30°E. We have developed the model based on k-means algorithm. The data sets of the climate parameters have been inputted after preprocessing. The output of the model for full training set of rainfall show that the cluster instances for the cluster 1 and cluster 0 have 75% and 25% among the 11 attributes selected for the training data set of rainfall for the period of 1901-2010. The cluster centroids for the for the full data set cluster 0 and cluster 1 are given in the table-1. Now we have extended the method for testing data set, which contain four climates attribute as shown in the run information. The impact of the first three attributes on the rainfall is reported in terms of cluster instances values for the cluster 0 and cluster 1. We found that 67% instances of the full data set directly affect the rain. Our finding validate model for the prediction of climate. The run information of the k-means clustering from WEKA tool is for the reference.

Clustering algorithms maps a new data item into one of several known clusters. K-means algorithm has biggest advantage of clustering large data sets and its performance increases as number of clusters increases. The k-means is a widely used partitioned clustering method in the industries. The k-means algorithm is the most commonly used partitioned clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time.

V. Output

Training Set	Testing Set																																																								
=== Run information === Scheme: weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10 Relation: rewa Instances: 12 Attributes: 11 Test mode: evaluate on training data === Clustering model (full training set) === k-Means ===== Number of iterations: 4 Within cluster sum of squared errors: 2.3427640507177223 Missing values globally replaced with mean/mode Cluster centroids:	=== Run information === Scheme: weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10 Relation: TCVtoR Instances: 12 Attributes: 4 Temperature Cloude Cover Vpressure Rainfall Test mode: evaluate on training data === Clustering model (full training set) === kMeans ===== Number of iterations: 4 Within cluster sum of squared errors: 1.4843695117982518 Missing values globally replaced with mean/mode Cluster centroids:																																																								
<table><tr><th>Attribute</th><th>Full Data (12)</th><th>Cluster 0 (3)</th><th>Cluster 1 (9)</th></tr><tr><td>1901-1910</td><td>81.835</td><td>271.1433</td><td>18.7322</td></tr><tr><td>1911-1920</td><td>88.1692</td><td>269.15</td><td>27.8422</td></tr><tr><td>1921-1930</td><td>92.3475</td><td>313.13</td><td>18.7533</td></tr><tr><td>1931-1940</td><td>91.4292</td><td>290.0067</td><td>25.2367</td></tr><tr><td>1941-1950</td><td>92.4367</td><td>304.6033</td><td>21.7144</td></tr><tr><td>1951-1960</td><td>93.06</td><td>287.3267</td><td>28.3044</td></tr><tr><td>1961-1970</td><td>88.8092</td><td>291.2833</td><td>21.3178</td></tr><tr><td>1971-1980</td><td>90.31</td><td>278.3167</td><td>27.6411</td></tr></table>	Attribute	Full Data (12)	Cluster 0 (3)	Cluster 1 (9)	1901-1910	81.835	271.1433	18.7322	1911-1920	88.1692	269.15	27.8422	1921-1930	92.3475	313.13	18.7533	1931-1940	91.4292	290.0067	25.2367	1941-1950	92.4367	304.6033	21.7144	1951-1960	93.06	287.3267	28.3044	1961-1970	88.8092	291.2833	21.3178	1971-1980	90.31	278.3167	27.6411	<table><tr><th>Attribute</th><th>Full Data (12)</th><th>Cluster 0 (4)</th><th>Cluster 1 (8)</th></tr><tr><td>Temperature</td><td>74.2117</td><td>193.935</td><td>14.35</td></tr><tr><td>Cloude Cover</td><td>36.5117</td><td>62.5775</td><td>23.4787</td></tr><tr><td>Vpressure</td><td>18.6817</td><td>29.9025</td><td>13.0712</td></tr><tr><td>Rainfall</td><td>25.5483</td><td>29.4925</td><td>23.5762</td></tr></table>	Attribute	Full Data (12)	Cluster 0 (4)	Cluster 1 (8)	Temperature	74.2117	193.935	14.35	Cloude Cover	36.5117	62.5775	23.4787	Vpressure	18.6817	29.9025	13.0712	Rainfall	25.5483	29.4925	23.5762
Attribute	Full Data (12)	Cluster 0 (3)	Cluster 1 (9)																																																						
1901-1910	81.835	271.1433	18.7322																																																						
1911-1920	88.1692	269.15	27.8422																																																						
1921-1930	92.3475	313.13	18.7533																																																						
1931-1940	91.4292	290.0067	25.2367																																																						
1941-1950	92.4367	304.6033	21.7144																																																						
1951-1960	93.06	287.3267	28.3044																																																						
1961-1970	88.8092	291.2833	21.3178																																																						
1971-1980	90.31	278.3167	27.6411																																																						
Attribute	Full Data (12)	Cluster 0 (4)	Cluster 1 (8)																																																						
Temperature	74.2117	193.935	14.35																																																						
Cloude Cover	36.5117	62.5775	23.4787																																																						
Vpressure	18.6817	29.9025	13.0712																																																						
Rainfall	25.5483	29.4925	23.5762																																																						
	Time taken to build model (full training data) : 22 seconds																																																								

1981-1990	83.6083	255.3967	26.3456	=== Model and evaluation on training set === Clustered Instances 0 4 (33%) 1 8 (67%)
1991-2000	74.2117	224.54	24.1022	
2001-2010	87.6217	278.4897	23.999	
Time taken to build model (full training data) : 20 seconds === Model and evaluation on training set === Clustered Instances 0 3 (25%) 1 9 (75%)				

References:

- [1]. Fahrmeir L. and Tutz T. Multivariate statistical modelling based on generalised linear models. New York: Springer Verlag, 2000.
- [2]. 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.
- [3]. Cavazos T. Using self-organization maps to investigate extreme climate event. Journal of Climate 2000; 13: 1718-1732.
- [4]. 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
- [5]. Badhiye S. S., Wakode B. V., Chatur P. N. "Analysis of Temperature and Humidity Data for Future value prediction", IJCSIT Vol. 3 (1), 2012, 3012 – 3014
- [6]. Bartok J., Habala O., Bednar P., Gazak M., and Hluch L., "Data mining and integration for predicting significant meteorological phenomena," Procedia Computer Science, p.37 – 46. 2010.
- [7]. Singhal S and Jena M. "A study on WEKA Tool for Data Pre-processing, Classification and Clustering", IJITEE, Vol (2)-6.pp-250, 2013.
- [8]. Kalyankar M.A and Alaspurkar S.J." Data mining technique to analyse the metrological data", IJARCSSE, Vol (3)-2 pp 114-118, 2013.
- [9]. Joshi A. and Kaur R., "Comparative study of various clustering techniques in data mining", IJARCSSE, Vol.(3)-3 pp 55-58, 2013.