MOL2NET, International Conference Series on Multidisciplinary Sciences
USEDAT-08: USA-Europe Data Analysis Training Program Workshop, UPV/EHU, Bilbao-MDC, Miami, USA, 2020

Application of hard K -mean technique in conjunction with fuzzy C-mean algorithm in clustering the pre-monsoon thunderstorm and non-thunderstorm days of Kolkata, India

Sweta Chakraborty*[a,b,d],Sarbari Ghosh [a,c,d]

[a] Department of Atmospheric Sciences, University of Calcutta, Kolkata, (India)

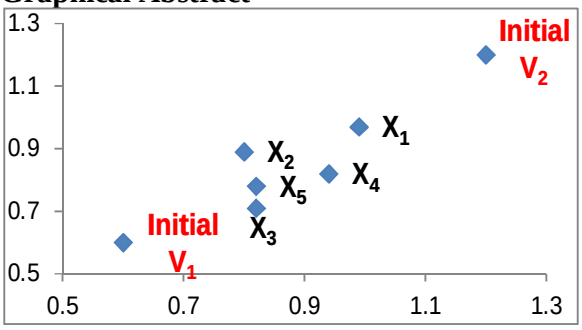
[b] Jagadis Bose National Science Talent Search, Kolkata,(India)

[c] Department of Mathematics, Vidyasagar Evening College, Kolkata,(India)

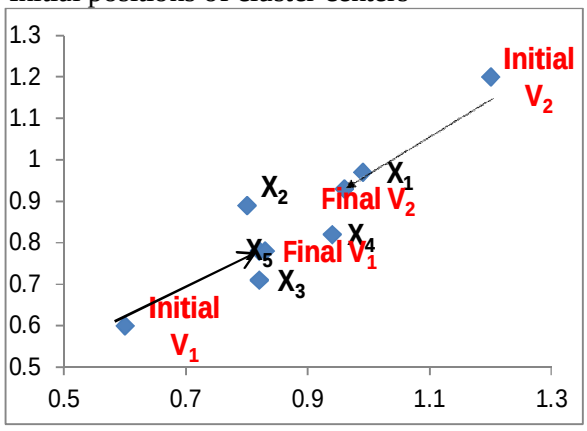
[d] Centre for Interdisciplinary Research and Education, Kolkata,(India)

.
.
.

Graphical Abstract



a. Initial positions of cluster centers



b. Final positions of cluster centers and their shifts

Fig.1 a) Initial positions of cluster centers, b) Final

Abstract.

The present study mainly aims at clustering of pre-monsoon thunderstorm (TS) and non thunderstorm (NTS) days over Kolkata ($22^0 32´$ N, $88^0 20´$ E) (India) using hard K-mean technique, backward selection procedure and fuzzy c- mean algorithm (FCM). The study involves the numerical values of the parameters observed at 0000 UTC and is performed in two stages. In the first stage , the hard C-mean technique is applied to cluster the days of a semi-supervised data set in the above mentioned two categories and the backward selection procedure is used to find the best possible combination of the theoretically influential atmospheric parameters that play the dominant role in the categorization on basis of performance score (PC). Though FCM technique is usually applied to supervised data set, this technique is applied to the semi-supervised data set of parameters to clarify the result obtained in the first stage.

positions of cluster centers and their shifts.	The final iteration in the first stage shows that the combination of maximum vertical velocity and P-PLCL at 1000 hpa level performs best in detecting the thunderstorm days so far the present dataset is concerned. It is interesting to note that this finding is also supported by FCM in the second stage of the study, where in the final iteration the center of the cluster consisting of thunderstorm days moves closer to the parameters , maximum vertical velocity and P-P_{LCL} at 1000 hpa level (the parameters, P and P_{LCL} represent respectively the pressure at the reference level and that at the corresponding lifting condensation level which is also considered as the cloud base) than that of the other cluster containing the non- thunderstorm days. Key words: Hard K-mean technique, Fuzzy C-mean algorithm, Backward selection procedure, Pre-monsoon thunderstorms, Lifting condensation level.

Introduction

Thunderstorm is a most spectacular mesoscale weather phenomenon resulting from the strong convective activity. This storm accompanied by a strong wind, lightning, heavy rain, and sometimes snow or hail. But, due to great potential to produce serious damage to human life and property, thunderstorms become a great concern for the atmospheric scientists.

During the transition period of pre-monsoon season (March-May), the differential air mass properties caused by moist warm southerly lower level wind flow from the Bay of Bengal and the cool dry westerly and North-Westerly upper-level wind that exist over this region favor exact climatological balance for the formation of the thunderstorms [1]. Pre-monsoon (March to May) or summer thunderstorms over Kolkata are generally known as “Nor’westers or ‘Kalbaishakhi’.

As thunderstorms formation is very complex in nature and thus no single parameter can be sufficient for its prediction [2]. Thunderstorm prediction is proved to be one of the most difficult tasks, due to their small spatial and temporal extension and the inherent

nonlinearity of their dynamics and physics [3]. Several studies have been conducted to explore the efficiency of various stability indices or meteorological parameters in representing the convective environment leading to the occurrence of a thunderstorm [4].

Fuzzy logic becomes more important in modern science. It is broadly used for forecasting of a complex system, data analysis and clustering. Clustering of data is the method of grouping data elements in such a way that objects in the same group are similar. Based on the data and the purpose of clustering, different similarity measures can be used to group items into classes.

Many previous studies have been performed by various researchers regarding the hard C-means as well as fuzzy clustering approach in the meteorological analysis. A short review of the literature is presented below:

1. Riordan et al. (2002)[5] proposed a hard K-means clustering algorithm framework for weather prediction in a Canadian airport. The study revealed that the K-means clustering could improve the accuracy of predictions of cloud ceiling and visibility at an airport by achieving direct, efficient, expert-like comparison of past and present weather cases.
2. Robson et al. (2005)[6] presented an efficient method for identifying temperature anomaly in the various regions of North America using fuzzy cluster analysis. The study showed that the single linkage cluster performed better than average linkage.
3. Lolis et al.(2008)[7]applied hard K-means clustering algorithm for classifying the temporal and spatial variability of winter cloud cover over Southern Europe and Mediterranean region and its relation with the general atmospheric circulation during the period 1950-2005. The study revealed that the atmospheric circulation can affect the cloudiness variability of the atmosphere.
4. Sonmez et al. (2011) [8] used the hard K- means clustering algorithm to reclassify rainfall regions of Turkey during the period of 1977-2006 using daily rainfall data and investigated their spatial and temporal variability in relation to the North Atlantic Oscillation.

5. Nath et al. (2015) [9]applied a fuzzy, C-means (FCM) clustering technique to investigate the track of tropical cyclones over the North Indian Ocean (NIO) for the period (1976-2014). The results indicated that each cluster has the unique features in terms of their genesis location, trajectory, seasonality, landfall, travel duration, accumulated cyclone energy and Intensity.

6. Saha et al. (2015) [10] studied a fuzzy clustering based method for predicting Indian monsoon using the El-nino data. The study revealed that the proposed ensemble approach surpassed the conventional approach.

7. Varsha et al(2018) [11] suggested a fuzzy Rule-based classification algorithm for annual rainfall prediction in Kerala. In this study five meteorological parameters,

Sea Level Pressure (SLP), Sea Surface Temperature (SST), humidity, zonal (u), and meridional (v) winds were used. The proposed models generated three hundred and eighteen fuzzy IF–THEN rules and fuzzy reasoning for prediction of rainfall clusters and the results were systematic in nature.

Materials and Methods

In the present study, the thunderstorm (TS) as well as non thunderstorm days (NTS) of the pre-monsoon season (March, April, May) of the years 2012-2017 are collected from IMD, Alipore. The quantified meteorological parameters for different atmospheric levels, such as potential temperature (θ), equivalent potential temperature(θe), temperature (T), pressure (P), pressure at lifting condensation level(P_{LCL}), wind speed, convective available potential energy(CAPE), geopotential height (Z) of 0000UTC are taken from the Department of Atmospheric Sciences of University of Wyoming .The meteorological parameters utilized in this study for clustering are mainly vertical wind shear(dv/dz) at 1000-850 hPa level (X_1) ,maximum vertical velocity(X_2),P-P_{LCL} at 1000 hPa level (X_3) , conditional instability ($\partial\theta es/\partial z$) for 700-600 hPa layer($X_4$), wind speed(m/s) at 850 hPa level (X_5).

Study with semi-supervised data set is a real-life approach for data analysis. So this approach is used for clustering the pre-monsoon days of Kolkata, India.

The thermodynamic and dynamic parameters which may be considered important for thunderstorm formations are selected for the study. Among them some meteorological parameters are derived from the collected primary data using the following formulas:

Here maximum vertical velocity is calculated from the √ 2*CAPE, P and P_{LCL} represent the pressure at the reference level and that at the corresponding lifting condensation level; $(\partial\theta es/\partial z)_{700\text{-}600}$ represents the conditional instability of the following layers, where, θes is the saturated equivalent potential temperature calculated from the standard formulae introduced by Bolton (1980)[12],

$\theta es = \theta \exp(L_v r_S/0.24\ AT)$ where L_v is the latent heat of vaporization of water and equal to 539 cal/g][θ=potential Temperature in K.

$\theta = T\left(\frac{P_0}{P}\right)^{R/C_p}$ where P_0 is the standard pressure, usually taken as 1000hpa, P is the pressure at reference level , R is the gas constant of air, and C_P is the specific heat capacity at a constant pressure.

AT =(273.5+T) where AT is the absolute temperature in K,T being the dew point temperature in deg C.

r_S=0.622*(es/(P-es)) where rs is the saturation value of water vapor mixing ratio in g/kg based on the parcel temperature and pressure .

es= 6.12 exp17.67 T/(T+ 243.5) where es is the saturated vapour pressure of water in hpa .

Vertical wind shear for horizontal wind =$\partial v/\partial z$, where v and z are respectively the wind speed in m/s and geopotential height in meter.

θe is the equivalent potential temperature in K.

The study involves the following clustering methodologies:

1. Hard C-mean technique with Backward Selection Procedure.
2. Soft C-mean technique with semi-supervised dataset of parameters.

1. Hard C-mean Technique and Backward Selection Procedure:

The 1st part of the present study focused on a centroid based, semi-supervised learning approach known as the hard C- means or fuzzy K-means algorithm [13]. In a hard clustering method each data point belongs to exactly one cluster data point is grouped into crisp clusters. In this study, the simulation of basic k-means algorithm is implemented using Euclidean and Manhattan distance metric to the semi-supervised data set to cluster the days in two groups with centers C_1 (TS) and C_2 (NTS). This approach is followed by the backward selection procedure for every combination of parameters, so that one parameter is removed at each iteration on the basis of the value of the performance score (PC) and HK skill score.

- General algorithm for hard C-mean technique with Euclidean and Manhattan metrics :

Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points (here X=1,2,3,4,5)
and $V=\{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers (here V=1,2)

1. Select 'c' cluster centers randomly (here c=2).
2. Calculate the distance between each data point and cluster centers using the Euclidean and Manhattan distance metrics as follows:

$$(\text{Euclidean Distance})_{XY} = \sqrt{\sum_{k=1}^{m}(X_{ik} - X_{jk})^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

$$(\text{Manhattan Distance})_{XY} = \left| X_{ik} - X_{jk} \right| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2)$$

where $i = (X_{i1},X_{i2}, \ldots ,X_{in})$ and $j = (X_{j1},X_{j2}, \ldots ,X_{jn})$ are two n dimensional data objects(here i=1,2 and j=1,2).

3. A data point (which denotes a day here) is assigned to a cluster, when its distance from the center of that cluster is the minimum of all.
4. New cluster center is calculated using the following formula:

$$V_i = \left(\frac{1}{c_i}\right) \sum_{1}^{c_1} x_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

where, 'c_i' denotes the number of data points in i^{th} cluster.

5. The distance between each data point and the new cluster centers is recalculated.
6. If no data point is reassigned then stop, otherwise repeat steps from 3 to 5.

This method is robust but the only limitation is that the accuracy of the result depends on the initial choice of the centers of the clusters.

The accuracy of the results is measured by proportion correct, where,

$$\text{Proportion Correct (PC)} = \frac{Number\ of\ right\ prediction\ of\ TS\ days}{Total\ number\ of\ days} \times 100 \quad \ldots(4)$$

Hanssen and Kuipers (HK) score [14-15] are calculated using the 2×2 contingency (Table 2) for each combination of parameters that are used to discriminate TS days from NTS days. Non-probabilities forecast of the categorical weather elements are verified by using the contingency table approach, which can serve as the beginning point for

examination of the strengths and weakness of the forecast. It gives information about the skill of the forecast as well as the type of errors that occur in the forecast.

The term categorical refers to the yes/no nature of the forecast verification at each grid point, then for each verification time it is scored as falling under one of the four categories of hits (A), false alarms (B), misses (C) or correct no rain forecasts (D). A number of hits (YY) (predicted and observed), B number of false alarms (NY) (predicted but not observed), C number of misses (YN) (observed but not predicted), D number of correct predictions of no rain (NN) (neither predicted nor observed)

$HK = (Acc)_{events} + (Acc)_{non-events} - 1 = AD - BC/(A + C)(B + D)$ ……(5)

$HK = \text{Correct forecast} - (\text{Correct forecast})_{random} /N - (\text{Correct forecast})_{random}$……(6)

Unbiased, Range : −1 to + 1; Perfect : 1.

The study has been started with all the five meteorological parameters as mentioned above. Finally, the backward selection procedure helps to select the combinations of the effective parameters to cluster the TS and NTS days. It is worth mentioning that the main focus here is to group the TS days as perfectly as possible.

2. Soft C-mean technique with semi-supervised dataset:

A semi-supervised fuzzy clustering algorithm with feature discrimination is imposed in the second part of the study. Though fuzzy clustering algorithm is normally applied on the unsupervised data set but in this study it is applied on the semi-supervised dataset to determine the veiled structure in the data set. The method produces a soft partition of a given dataset. Here, the objective function J_1 of fuzzy C-means has been extended in two ways:

1. The fuzzy membership degrees in clusters were incorporated into the formula;
2. An additional parameter m was introduced as a weight exponent in the fuzzy membership.

The extended objective function, denoted as J_m, is

$J_m(P,V)=\sum_{i=1}^{k}\sum_{xk\in X}(\mu_{ci}(x_k))^m \|x_k - v_i\|^2$ ……..……(6)

where P is a fuzzy partition of the dataset X formed by $C_1,C_2,\ldots..C_k$. The parameter m is a weight that determines the degree to which partial members of a cluster affect the clustering result.

Like hard C-mean technique, the fuzzy C-mean algorithm also tries to find a partition by searching for prototypes v_i that minimize the objective function J_m. Unlike C-means, however, the fuzzy C-means algorithm needs to search for membership functions μ_{ci} that minimize J_m.

A constrained fuzzy partition $\{C_1,C_2\ldots..C_k\}$ can be a local minimum of the objective function J_m only if the following conditions are satisfied:

$$\mu_{ci}(x)=\frac{1}{\sum_{j=1}^{k}\left(\frac{\|x-v_i\|}{\|x-v_j\|}\right)^{\frac{2}{m-1}}} \qquad 1\le i \le k;\ x\in X \quad \ldots\ldots\ldots\ldots\ldots\ldots(7)$$

$$v_i=\frac{\sum_{x\in X}(\mu_{ci}(x))^m x}{\sum_{x\in X}^{n}(\mu_{ci}(x))^m} \qquad 1\le i \le k; \quad \ldots\ldots\ldots\ldots\ldots\ldots(8)$$

Here X= a semi supervised data set of parameters, (i= 1,2…n), k=2

c_i=the number of clusters to form (here c=2)

m= the parameter in the objective function (here m is taken to be 2)
ε= threshold for the convergence criteria (here $\varepsilon=10^{-2}$)
v_i= initialized prototype : $\{v_1,v_2\}$
This iteration continues until the center of new cluster, v_i and center of the previous cluster $v_{Previous}$ isles than or equal to the threshold value,
i.e., $\sum_{i=1}^{2}\|v_i^{Previous} - v_i\| \leq \varepsilon$ …… (9) is reached.

Results and Discussion.

In the first stage of the study using hard-C mean technique with Euclidean and Manhattan metrics, it is noted that Euclidean metric consistently performs better than Manhattan's in clustering the days (TS and NTS) of the semi-supervised data set, so far the present atmospheric parameters are concerned. Hence the iterations with Euclidean metric are given importance in the present analysis though in table 1 both the results are presented. The results of proportion correct (PC) and Hanssen and Kuipers discriminant (HK) skill score are also presented in the table 1.

The combination of all the five parameters at the start step produces 43.7 % correct prediction of TS days. The next steps are performed with the all possible combinations of the four, three, two parameters respectively and the result shows that PC value increases when the less number of parameters are involved. The combination of X_2, X_3, X_4, X_5 gives PC value 46.09,HK skill score 0.06, the combination of X_2, X_3, X_5 gives PC value 50, HK skill score 0.08 and the combination of X_2 and X_3 produces the highest PC value 54.5, HK skill score 0.05 for TS days so far the present parameters are concerned.

In the second stage, the iteration is to clarify the final combinations of parameters as obtained in the first stage as far as possible. So, the FCM algorithm is started with two unlabelled data set of five parameters. It is to be observed that the numerical values of some parameters are overlapping in two situations.

It is interesting to note that the final iteration of fuzzy C-mean technique clearly indicates that the center V_1 of C_1(TS) approaches the cluster containing X_2, X_3, X_5, whereas V_2 of C_2(NTS) gets closer to the cluster center containing X_4 and X_1, though X_4 is almost at the same distance from the two centers [Table 3,Fig 1.a,b]. Hence the result of the hard C-mean Technique may possibly be justified as follows:

The result is slightly biased towards NTS days while clustering the pre-monsoon days of the urban area, Kolkata (India) in two groups, TS and NTS on the basis of the observations at 0000 UTC with the help of the hard C-mean technique. Since X_4 lies almost on the overlapping portion of the boundaries of the two clusters, therefore X_4 may be partly responsible for this bias.

Conclusions

The study in the first stage reveals that the hard C-mean technique with Euclidean metric performs better than that with Manhattan metric. Based on "Proportion Correct" (PC) and "Hanssen and Kuipers discriminant" (HK) Skill Score and backward selection procedure the study brings out the following observations:

- It is observed that on the basis of the observations at 0000UTC the combination of two parameters among the five namely, the maximum vertical velocity and P-

P_{LCL} furnishes better results than the other combinations in clustering of a semi-supervised set of days.

- The last iteration of backward selection procedure is ignored as only one parameter is not sufficient to predict any atmospheric situation.
- Not only that, another interesting observation is that the bias becomes minimum and PC becomes maximum once the parameter X_4 is excluded.

In the second stage of the study, the Fuzzy C-means algorithm helps clarify the reasons behind the results obtained in the first stage, where the correct classifications are slightly biased towards the NTS days. The reason may be that the number of available data is more in NTS days than that of TS days and the parameter X_4 has overlapping numerical values in two situations.

In this stage, among the five parameters initially considered for clustering into two groups, TS (C_1 with the center V_1) and NTS (C_2 with the center V_2), the center V_2 of NTS days finally gets closer to the cluster formed by the parameters vertical wind shear at 1000-850 hPa level (X_1) and conditional instability for the layer (700-600) hPa (X_4) and V_1 gets closer to X_4 ,maximum vertical velocity (X_2), P-P_{LCL} at1000 hpa level(X_3),wind speed(m/s) at 850hpa level (X_5). That may be one of the underlying reasons for bias in the results towards the NTS days in the first stage and comparatively low percentage ($\leq$57) of correct classification of TS days.

From the above observations it may be successfully used to cluster inferred that the conjunction of hard C-mean and fuzzy C-mean techniques may be successfully used to cluster the data of a semi-supervised data set and explain the results as far as possible.

It is also expected that this methodology may be applied in the other fields too for clustering different situations that depend on the parameters having overlapping quantified values.

The present work has definite scope for improvement provided some more potential parameters for TS/NTS generation can be incorporated therein.

Acknowledgements:

Thanks to the Regional Meteorological Centre, India Meteorological Department, Alipore, for supplying the necessary rainfall data.

Our sincere gratitude to Late Prof. Utpal Kumar De, Ex-emeritus Prof. of School of Environmental Studies, Jadavpur University, without whose active participation and encouragement this study could not have been possible.

References

1 Desai, B. N.; Rao, Y. P. On the cold pools and their role in the development of Nor'westers over West Bengal and East Pakistan. Ind. J. Meteor. Geophys. 1954, 5, 243-248.

2 Lorenz, E. N. Deterministic non periodic flow. J. Atmos. Sci. 1963, 20,130-141.

3 Orlanski,I. A rational subdivision of scales for atmospheric processes. Bull. Am. Meteorol. Soc.1975, 56,527-530.

4 Ghosh,S.; Sen.P.K.; De,U.K. Identification of significant parameters for the prediction of pre-monsoon thunderstorms at Calcutta.Int.J.Climatol.1999,19,673-681.

5 Riordan,D.; Hansen,B.K. A fuzzy case-based system for weather prediction. Eng Int Syst. 2002, 3, 139–146.

6 Robeson,S.M.; Doty,J.A. Identifying Rogue Air Temperature Stations Using Cluster Analysis of Percentile Trends. Bull. Am. Meteorol. Soc.2005,18,1275-1287.

7 Lolis,C.J. Winter cloudiness variability in the Mediterranean region and its connection to atmospheric circulation features. Theor. Appl. Climatol. 2009, 96,357–373.

8 Sönmez,I.; Kömüşcü,A.U. Reclassification of rainfall regions of Turkey by K-means methodology and their temporal variability in relation to North Atlantic Oscillation (NAO). Theor. Appl.Climatol.2011, 106,499–510.

9 Nath,S.; Kotal,S.,D.; Kundu,P.K. Application of Fuzzy Clustering Technique for analysis of North Indian Ocean Tropical Cyclone Tracks. Tropical Cyclone Research and Review.2015,4,110-123.

10 Saha,M.;Mitra.P.;Chakraborty.A. Fuzzy Clustering-Based Ensemble Approach to Predicting Indian Monsoon. Adv. Meteorol.2015, 2015, 1-12.

11 Varsha,K.S.; Maya,L.Pai. Rainfall Prediction Using Fuzzy C-mean Clustering and Fuzzy Rule-Based Classification. Int. J. Pure. Appl. Math.2018, 119, 597-605.

12 Bolton, D. The computation of equivalent potential temperature. Mon. Weather Rev.1980, 108, 1046-1053.

13 Zeng,S.; Vaughan,M.; Liu,Z.; Trepte.C.; Kar,J.; Omar,A.; Winker,D.; Lucker,P.; Hu,Y.; Getzewich,B.; Avery,M. Application of high-dimensional fuzzy k-means cluster analysis to CALIOP/CALIPSO version 4.1 cloud–aerosol discrimination. Atmos. Meas. Tech.2019, 12, 2261–2285.

14 Hanssen, A. W.; Kuippers, W. J. A. On the relationship between the frequency of rain and various meteorological parameters. Verhand. K. Nederlands. Meteorl. Inst.1965, 81, 2-15.

15 Woodcock,F.The Evaluation of Yes/No Forecasts for Scientific and Administrative Purpose. Mon. Weather Rev.1976, 104, 10, 1209-1214.

Tables:

Sl no	Combination of Parameters(00Z)	Steps performed	Euclidean Distance (Hanssen and Kuipers discriminant) HK Skill Score	Total Proportion Correct(PC)		Manhattan Distance (Hanssen and Kuipers discriminant) HK Skill Score	Total Proportion Correct(PC)		Selection of parameters/combination of parameters based on Proportion Correct and HK Skill Score
				PC of TS	PC of NTS		PC of TS	PC of NTS	
1	X_1,X_2,X_3,X_4,X_5	**Starting**	0.02	53.27		0.08	51.40		Starting Combination
				43.7	58.8		45.1	63.88	
2	X_1,X_2,X_3,X_4	**Step-1**	0.03	53.27		-0.02	50.47		Not Selected
				43.9	59.0		40.9	57.1	
3	X_1,X_2,X_3,X_5		-0.06	49.53		-0.08	48.59		Not Selected
				37.8	55.7		36.1	54.9	
4	X_1,X_2,X_4,X_5		-0.01	52.33		-0.07	49.53		Not Selected
				41.2	57.5		37.1	55.5	
5	X_1,X_3,X_4,X_5		-0.02	51.4		-0.07	49.53		Not Selected
				40.5	57.1		36.3	55.4	
6	$\mathbf{*X_2,X_3,X_4,X_5}$		**0.06**	**56.07**		**-0.06**	**51.4**		**Selected**
				46.9	60		36.4	55.2	
7	$\mathbf{*X_2,X_3,X_5}$	**Step-2**	**0.08**	**57.94**		**0.08**	**57.01**		**Selected**
				50	61.9		48.3	60.5	
8	X_2,X_3,X_4		0.09	57.01		-0.08	50.47		Selected
				48.6	61.1		33.3	55.4	
9	X_2,X_4,X_5		0.05	56.07		0.1	56.07		Not Selected
				46.7	59.7		47.8	62.2	
10	X_3,X_4,X_5		0.05	55.14		-0.02	54.2		Not Selected
				45.7	59.7		40	57.4	
11	$\mathbf{*X_2,X_3}$	**Step-3**	**0.05**	**58.88**		**0.03**	**56.07**		**Selected**
				54.5	59.3		45.4	58.8	
12	X_3,X_5		0.06	57		0.08	54.21		Not Selected
				48	59.		40	57.5	

					7				
13	X_2,X_5		0.13	57		0.08	54.21		Not Selected
				48.9	63.7		46.1	61.8	
14	X_2,X_4		0.05	56.07		0.03	53.27		Not Selected
				46.7	59.7		45	58.2	
15	X_3,X_4		0.07	56.07		0.02	56.07		Not Selected
				47.1	60.3		45	58.6	
16	X_4,X_5		0.07	56.07		-0.03	51.4		Not Selected
				47.2	60.5		39.4	56.8	

Table 1. Selection of the parameters based on Proportion correct and HK skill score in hard C- means algorithm with Euclidean and Manhattan Distance metrics (* Denotes the best combination in each step.

Forecast	Observed		Total
	Yes	No	
Yes	A	B	A+B
No	C	D	C+D
Total	A+C	B+D	A+B+C+D=n

Table 2. 2×2 contingency table for yes/no forecast verification

Parameters	**TS**	**NTS**
V_1(Initial)(Cluster 1)	0.60	0.60
V_2 (Initial)(Cluster 2)	1.20	1.20
Wind Shear(1000-850) X_1	0.99	0.97
Max Vertical Velocity X_2	0.80	0.89
P-Plcl X_3	0.82	0.71
Conditional Instability X_4	0.94	0.82
Wind Speed At 850 Mb X_5	0.82	0.78
V_1(Final)(Cluster 1)	0.83	0.78
V_2(Final) (Cluster 2)	0.96	0.93

Table 3. Shift of the centers of initial V_1, V_2 and final V_1, V_2