International Journal of Innovative Technologies
WWW.IJITECH.ORG
ISSN 2321-8665
Vol.04,Issue.06,
June-2016,
Pages:0973-0978

Identification of Dipoles in Climate Data

RAMIREDDY THUSHARA BHARATHI[1], J. VELMURUGAN[2]
[1]PG Scholar, Dept of CSE, Sri Venkateswara College of Engineering and Technology, Chittoor, AP, India,
E-mail: thusharacse@gmail.com.
[2]Associate Professor, Dept of CSE, Sri Venkateswara College of Engineering and Technology, Chittoor, AP, India,
E-mail: velmurugan85@gmail.com.

Abstract: The Dipoles are pair of regions where they possess strong negative correlation with the locations in other regions and strong positive correlation with the locations in the same regions. Majorly the pressure dipoles i.e. the tele-connections characterize the pressure anomalies for the two different locations at the same time by using correlations. Here the correlations, calculated through linear relationship between sets of data are used to represent the time series defined as Climate Indices. We have studied many methods such as clustering and community based approaches and we use the new KNN-NF Approach. This approach results in capturing the various climatic conditions and the anomalies that arise at various locations with respect to time.

Keywords: Dipoles, Correlations, Agglomerative Methods, Climate Indices, KNN-NF.

I. INTRODUCTION

The Datamining converts the raw data to the information of present. Datamining is a vast technology that encompasses different classification and clustering approaches. Now-a-days this extends its scope in every field such as in Medicine, Technology and Climatology. This offers usage in Neural Networks, Support Vector Machines and in other common techniques as Prediction, Outlier Detection and Regression. The climate gets varied in many regions of the world. These teleconnections are used to identify the changes that often takes place in the climatic conditions such as rainfall [15], hurricanes, droughts severities [24] etc. The Climate Indices here plays a key role in detecting the impact of various climatic parameters such as Pressure, Temperature and Precipitation. These together obtain the predicted values by considering the present values of the parameters needed. The extreme weather events generally depend on the seasonality, variability and the auto-correlation. The Dipoles are used to detect the climatic anomalies occurring at same or in an opposite polarities appearing at different or same locations at the same time. The dipoles are detected by the formation of correlation of different places. This correlation is further divided as positive and negative correlations. For the clustering approaches the agglomerative hierarchical approaches are then followed. Based on the characteristics the dipoles are classified as pressure dipoles that capture recurring and persistent large scale patterns of pressure and temperature dipoles that often defines the sudden fluctuations [14] occurred in the atmospheric temperature. There are many kinds of dipoles as per the study made by the scientists. Among all those the major kinds of dipoles are North Atlantic Oscillation (NAO) and Southern Oscillation Index (SOI). The paper is structured by sections. The sectionI includes basic introduction regarding the paper. The section II describes regarding the literature survey. The section III includes the dataset used in the implementation. The section IV includes the related work done in the existing system. The sectionV describes the implementation procedure of proposed system in detail. The sectionVI the results obtained. The conclusion and future work is then defined in the final sectionVII.

II. LITERATURE SURVEY

There have been done many studies regarding this clustering, detection of dipoles, teleconnections, forecasting and so on. Among those a paper by Jaya Kawale portrays the various climate indices and the data selection process used in datamining, it includes use of General Circulation Models [16] (GCMs) for the prediction of data. The paper proposed by Hetal Bharat Bhavsar provides description regarding the use of Graph Theory [21], this uses simple Shared Nearest Neighbor algorithm [11] where in doesn't cluster all the points considered. The paper given by Gince Keziban Orman enlightens different Community detection algorithms [12] and the comparison among them. In the paper presented by Michael Steinbach they have determined the climate indices for the Oceans using the Clustering process [8], here the Eigen Value analysis is used through the calculation of centroids of the clusters. The paper given by Aaron Clauset describes regarding the finding of the community structure in the large networks [4] formed while clustering, this uses the concept of modularity with the greedy approach. Other paper proposed by Jaya Kawale implements the detection of the dipoles in climate data [1] by the use of datamining techniques as Seasonality Removal and further use of the Community [9] based approaches along with the alternative approach named A1 Algorithm that depicts Nearest

Neighbor Approach. However few of these methods are used by overcoming the drawbacks found in those systems.

III. DATASET

Teleconnections [22] is one of the important parameters considered for the atmospheric based issues as for the climate indices. For this analysis of this project the NCEP/NCAR Reanalysis data [23] is considered assimilated for the years required for the detection. This includes the particular latitude and longitude needed for the particular years of data considered. There may be the other parametric data such as Sea Level pressure, Temperature, Humidity, and so on. The datasets can be collected from one of these meteorological centers as National Center for Atmospheric Research (NCAR), National Center for Environmental Prediction (NCEP), National Oceanic and Atmospheric Administration (NOAA) [3],[17], Physical Sciences Division (PSD) [6]. Among all the Climate Indices such as North Atlantic Oscillation (NAO), Southern Oscillation Index (SOI), El Nino Southern Oscillation (ENSO) [13], [20] and Sea Surface Temperature (SST), Southern Oscillation Index (SOI) is the most commonly used Climate indices.

IV. RELATED WORK

The Existing System [1] includes the calculation of dipoles for the opposite polarities on the yearly basis. It obtains the different dipoles formed in the numerous years of the same place. This includes the Data Smoothing, Obtain anomaly values by Data Normalization as Seasonality Removal, Finding Correlation by Edge Weight Estimation. Then the process includes formation of clusters and finally finding the dipoles. The Data Smoothing includes removal of inconsistencies, Data Normalization includes finding the anomaly time series and Edge Weight Estimation includes computation of similarities in between the places forming networks [2][10]. The clusters can be formed by the use of clustering approaches [25] as k-means clustering and the other community based methods as Walk trap [7] Community detection and the Nearest Neighbor Algorithm.

V. PROPOSED SYSTEM

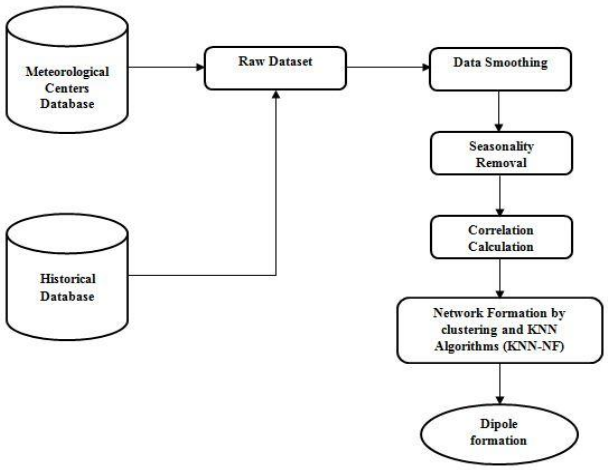

Fig.1. Architecture of Proposed System.

The Proposed System includes the identification of dipoles for the same polarities on the basis of various places. It obtains the dipoles formed in the numerous places in the same year and vice-versa as shown in Fig.1. The KNN-NF Approach defines well and effective way of identifying the dipoles in the data considered. It obtains the dipoles formed in the numerous places in the same year and vice-versa. The KNN-NF Approach defines well and effective way of identifying the dipoles in the data considered. To summarize the complete process of series of steps are followed as,

Step 1: Extract the data from either the historical data or from the meteorological centers. This results with the raw data.

Step 2: Then perform the data smoothing by removing the unnecessary data resulting the anomaly values through seasonality removal [18].

$$x_y(m) = x_y(m) - \mu_m \;//\; \text{where } m\epsilon\{1,2,\ldots\ldots,12\} \text{ and } y\epsilon\{1948,\ldots\ldots,2009\} \quad \textbf{(1)}$$

$$\mu_m = \frac{1}{end-start+1} \sum_{y=start}^{end} x_{y(m)} \quad \textbf{(2)}$$

Here the μ_m represents the mean of the month m for all the 12 months in a year by considering the moving average of three months. $x_y(m)$ represents the pressure values considered for the year y from 1948 to 2009 spanning about 63 years of data for each month m. The end and start represents the years 2009 and 1948. All the values of $x_y(m)$ are summed together for all the 63 years of data and are allowed to perform the mean of subtraction of all the years.

Step 3: In the next stage construct the networks [5] for the obtained data by calculating the correlation values. This obtains the various values of anomaly data [27] for different time periods. Then to form the networks threshold the graph resulting in both the positive and negative correlation values.

Step 4: The correlation can be calculated by using the edge eight estimation using different correlation measures. The correlation measures considered are such as pearson correlation and other similarity measures. By these correlation values the networks are constructed by similar values obtained between the two series of data. The correlation values are calculated by the use of linear measure namely pearson correlation.

$$r_{xy} = \frac{\sum_{n=1}^{n}(xi-x')(yi-y')}{(n-1)sxsy} \quad \textbf{(3)}$$

Then among the correlation values obtained after the filtering the edges of a graph, these are further divided as positive and negative correlations. The positive correlation values depict the occurrence in same polarity and the negative correlation values depict the occurrence in opposite polarity as shown in Figs.2 and 3. Here these positive values considered are the values taken after finding the threshold value for the data obtained. And the data within the threshold area is termed to be the positive correlation and the other

data values lying outside the formed clusters [19] called outliers are pruned by filtering the edges of graph. Similarly through the negative values considered and by threshold value the negative correlation values are shortlisted by filtering the outlier edges lying in the graph. Thus in this way both the positive and negative correlation values are obtained separately. The graphs observed in both fig.4 and 5 depicts the positive and negative correlation between the series of data.

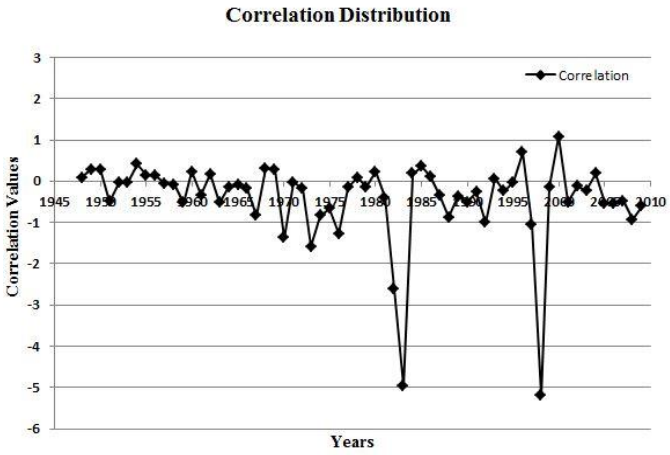

Fig. 2. Distribution of correlation.

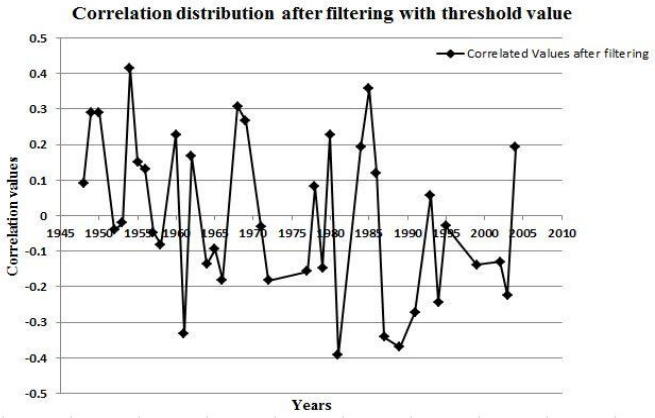

Fig. 3. Distribution of Correlation after thresholding the graph.

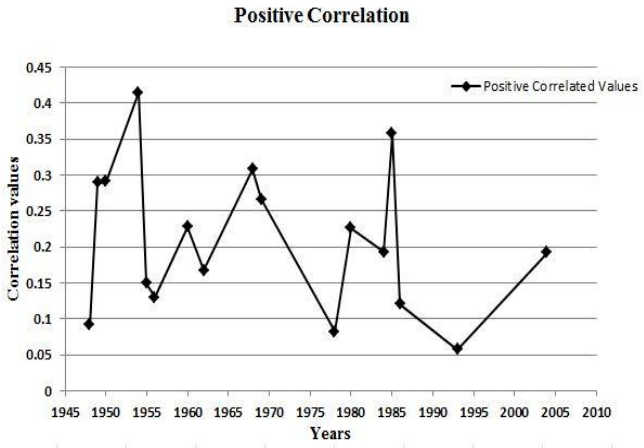

Fig.4. Positive Correlation values within threshold obtained

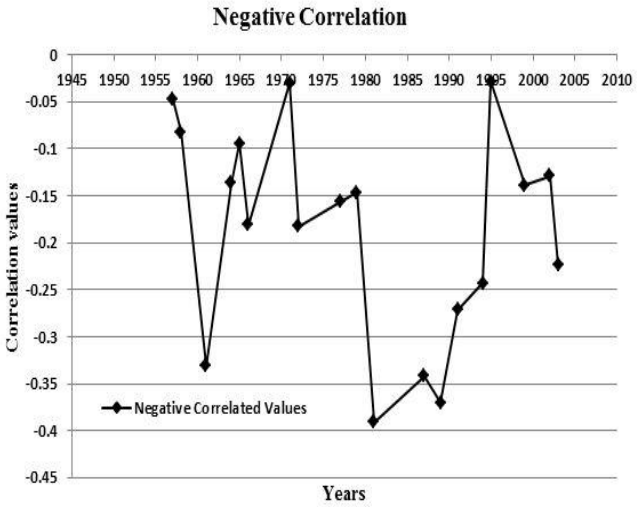

Fig.5. Negative Correlation values within threshold obtained.

In the next stage construct the networks for the obtained data by calculating the correlation values. This obtains the various values of anomaly data for different time periods. Then apply the algorithm KNN-NF Approach for finding the dipoles which in parallel for the networks, perform pruning and also finds the similarity distance measures to the data. By applying the above equations it results in the anomaly values for the series of time with respect to the location.

A.Network Formation by Clustering and KNN Approach (KNN-NF)

From the obtained values of correlation the networks are formed for both the negative and positive correlation values. Thus in this project a new approach is used named KNN-NF approach which internally includes the functionalities of both the clustering and the classification techniques. These provide the complete process of network formation and the identification of dipoles. This algorithm KNN-NF specified works initially by separately grouping the moving average of five each as a community. For the grouped communities the distance measure is calculated and based on this the networks are formed. For the formed networks the KNN approach is applied such that it calculates the similarity values in between the different nodes in a network [5]. This process continues by considering only the filtered nodes. The nodes of a network are again here filtered by making an allowance for only the densely connected nodes. And the rest of the nodes that are not in densely connected regions are pruned. This algorithm possesses the basic characteristics in finding the dipoles in the regions. This works for building both the positively and negatively edged graphs of the regions. The positively obtained ones specifies the identification in same polarities and the vice versa for the negatively obtained data.Initially the dataset can be collected either form the meteorological centers or from the historical database [26] collected form the web portals. This obtains the raw data which is further given for seasonality removal and in finding the correlation values. Then find the both N

and P values of negative and positive correlation values depicting by the span of 5 years period. Then perform the hierarchical clustering for the obtained data by forming the clusters. This in parallel applies the classification approach namely KNN approach to the data which further calculates the similarities existing in between the different nodes of years considered. For the data obtained by applying this algorithm pruning is done in various phases. After each phase the same hierarchical approach is applied for the densely obtained regions.

Algorithm:

Algorithm: New KNN-NF Approach for finding Dipoles
Require: The Data obtained through Edge Weight Estimation

τ = Correlation Values Threshold
D = Values of Dataset
n = total number of years
P = Positive Threshold Values
N = Negative Threshold Values
For i = 1 to n do
τ = Mean (Values of location dataset with respect to time period considered)
D = $-\tau \leq x \leq +\tau$;
{For every node in dataset D, the values lying in between the threshold are defined by x}
if ($D \leq +\tau$ && $D > 0$: true : false) then
P = D // defining the positive correlation values through the true obtained
end if
if ($D \geq -\tau$ && $D < 0$: true : false) then
N = D // defining the negative correlation values through the true obtained
end if
end for
for i = 1 to n do

{Apply KNN Approach to the data & in parallel form networks by hierarchical clustering}
P = By moving 5 years of span form networks by sorting as $P_1, P_2, P_3, \ldots\ldots, P_n$
N = By moving 5 years of span form networks by sorting as $N_1, N_2, N_3, \ldots\ldots, N_n$
for each node of $P_1 \ldots P_n$ and $N_1 \ldots N_n$ do
val = (pow (P[i]-P[n],2));
{Prune the values resulting only the values of densely formed regions in a network & repeat till the minimum distances are obtained through these dense regions}
if ($val(P) \geq -\tau$ && $val(N) \leq +\tau$) then
return (P[i],N[i]) // Dipoles are detected
else return "no dipoles formed" //Also can do same process for negative values too
end if
end for

The same process continues till it obtains the reductant data with minimum distance measure between the nodes that includes the similar occurrence of pressure anomalies termed as dipoles are obtained.

VI. EXPERIMENTAL RESULTS

Finally through the above process, i.e., by the use of KNN-NF approach specified in this obtains the correlation values of both the places deliberated.

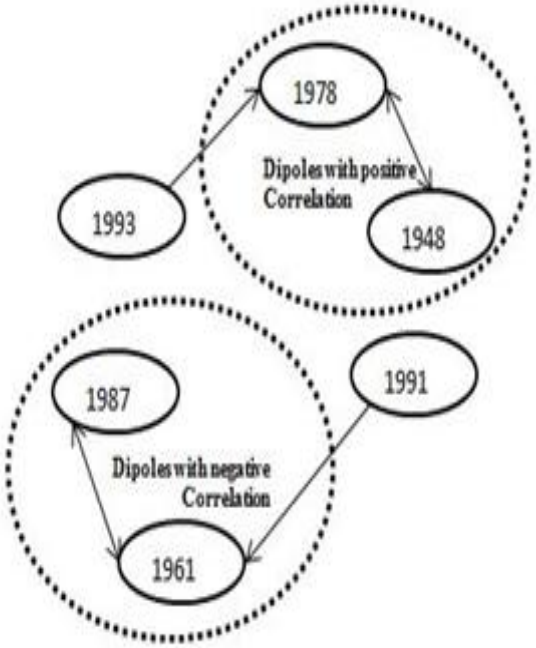

Fig.6. Network depicting the dipoles formed with the positive and the negative correlations.

Fig.6 shows the dipoles formed in the network along with their neighboring nodes for both positive and negative correlation values. For the further analysis the facts are shown in the following scheduled table displaying the dipoles formed in both the positive and the negative correlation. This is tabulated along with their neighboring node values connected to these main nodes in the network.

TABLE I: Dipoles of Positive and Negative Correlation Values in a Network

Year	Positive Correlation	Negative Correlation
1948	0.09123561	-
1961	-	-0.331515062
1978	0.081869598	-
1987	-	-0.341824157
1991	-	-0.271387857
1993	0.057809642	-

* Here '-' defines the pruned values.

TABLE II: Dipoles Formed with Positive Correlation

Year	Positive Correlation
1993	0.057809642
1978	0.081869598
1948	0.09123561

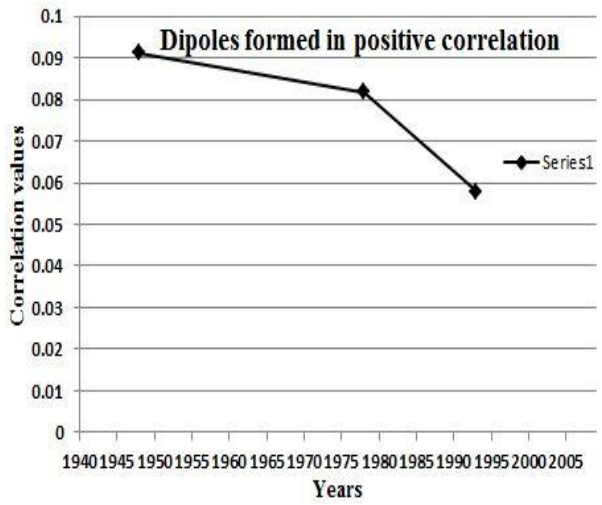

Fig.7. Dipoles in positive correlation.

The above table 2 and Fug.7 depicts the dipoles formed in the positive correlation. This shows that the dipoles are identical in the magnitude of time and location.

TABLE III: Dipoles Formed with Negative Correlation

Year	Negative Correlation
1987	-0.341824157
1961	-0.331515062
1991	-0.271387857

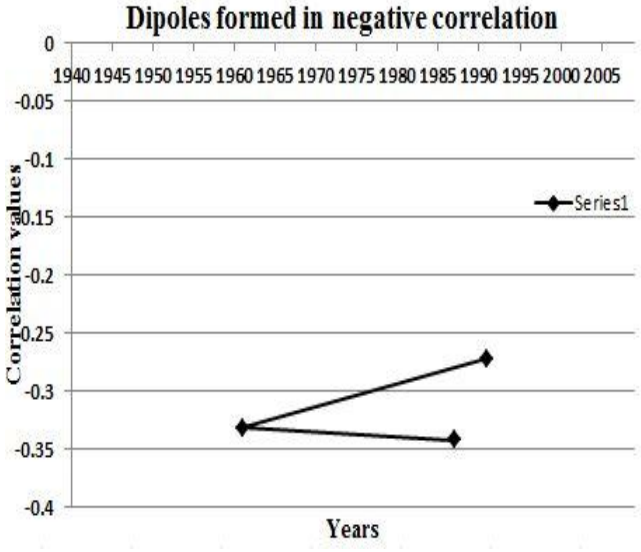

Fig.8. Dipoles in negative correlation.

The above table3 and Fig.8 depicts the dipoles formed in the negative correlation. This shows that the dipoles are not identical in the magnitude of time in different locations.

VII. CONCLUSION

The problem of finding the dipoles tends to be a very prominent and interesting concept for the climate scientists. The project identification of dipoles in climate data presents a newfangled approach named KNN-NF algorithm by merging the two different approaches as network formation and KNN approach. This methodology applied here seems to afford the better results in finding the dipoles. Through the algorithm and process used in this project the dipoles in both the positive and negative are obtained. The positive dipoles are represented in the same time and same location whereas the negative dipoles are represented in vice-versa i.e., in the same time but different locations. In future for further investigation and improvements this can be extended with the forecasting of the occurrence of dipoles in the different regions using the various better approaches and also the dipoles detection can be done with respect to different locations at same time. For the accomplishment of forecasting it should consider the obtained data as testing dataset for calculating and identifying the dipoles for the training dataset considered.

VIII. ACKNOWLEDGMENT

I whole-heartedly express my gratitude and esteemed regards to Mr. J.Velmurugan, Dr.M.Venkatesan, my parents, friends and well-wishers who helped me in carrying out this work successfully.

IX. REFERENCES

[1]Jaya Kawale, Michael Steinbach, Vipin Kumar, Discovering Dynamic Dipoles in Climate Data, with proceedings of 2011 in SIAM International Conference on Datamining of PRDT11, DOI 10.1137/1.9781611972818.10.

[2]Steinhaeuser, K., Chawla, N. V., Ganguly, A. R. An exploration of climate data using complex networks. KDD Workshop on Knowledge Discovery from Sensor Data, pp. 23{31, 2009.

[3]Storch, H. V. and Zwiers, F. W. Statistical analysis in Climate Research. Cambridge University Press, 1999. Taylor, G. H. Impacts of the El Ni~no/southern oscillation on the paci_c northwest. Technical report, Oregon State University, USA, 1998.

[4]Aaron Clauset, M. E. J. Newman and Cristopher Moore, Finding community structure in very large networks, arXiv:cond-mat/0408187v2, 2004.

[5]Tsonis, A. A., Swanson, K. L., Roebber, P. J. What Do Networks Have to Do with Climate. In Bulletin of the American Meteorological Society, vol. 87, no.5, pg. 585, 595, 2006.

[6]GEOPHYSICAL RESEARCH LETTERS, with the volume VOL. 32, letter L15710, doi:10.1029/2005GL022709, 2005.

[7]Pons, P., Latapy, M. Computing Communities in Large Networks Using Random Walks. Journal Graph Algorithms Applications. 10(2): 191-218, 2006.

[8]Steinbach, M., Tan, P., Kumar, V., Potter, C and Klooster, S. Data mining for the discovery of ocean climate indices. In Mining Scientific Datasets Work- shop, 2nd Annual SIAM International Conference on Data Mining, 2002.

[9]Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre, Fast unfolding of communities in large networks, arXiv.0803.0476v2, 2008.

[10]Donges, J. F., Zou, Y., Marwan, N. Complex networks in climate dynamics. In European Physical Journal Special Topics, 174 (1), pp. 157{179, 2006.
[11]Ertoz, L., Steinbach, M., Kumar. V. A new shared nearest neighbor clustering algorithm and its applications. In Workshop on Clustering High Dimensional Data and its Applications, SIAM Data Mining, 2002.
[12]Gince Keziban Orman, Vincent Labatut and Hocine Cherifi, University of Burgandy, Qualitative Comparison of community detection algorithms, 2008.
[13]Gozolchiani, A., Yamasaki, K., Gazit, O., Havlin S. Pattern of climate network blinking links follows ElNi~no events. In Europhysics Letters, vol 83, issue 2, 2008.
[14]Fogarty, E. A., Elsner, J. B., Jagger, T. H., Tsonis, A. A. Network Analysis of U.S. Hurricanes, Hurri- canes and Climate Change, 1-15, 2009.
[15]Gadgil, S. and Vinayachandran, PN and Francis, PA and Gadgil, S. Extremes of the Indian summer monsoon rainfall, ENSO and equatorial Indian Ocean oscillation. In Geophysical Research Letters, 174 (1), pp. L12213{1, 2004.
[16]Jaya Kawale. Snigdhansu Chatterjee, Arjun Kumar, Stefan Liess, Michael Steinbach and Vipin Kumar, Anomaly Construction in Climate data : Issues and Challenges, 2010.
[17]Steinbach, M., Tan, P., Kumar, V., Potter, C and Klooster, S. Data mining for the discovery of ocean climate indices. In Mining Scientific Datasets Work- shop, 2nd Annual SIAM International Conference on Data Mining, 2002.
[18]Portis, D. H., Walsh, J. E., El Hamly, Mostafa and Lamb, Peter J., Seasonality of the North Atlantic Oscillation, Journal of Climate, vol. 14, pg. 2069-2078, 2001.
[19]Steinbach, M., Tan, P., Kumar, V., Potter, C., Klooster, S. Clustering earth science data: Goals, issues and results. In Proceedings of the 4th KDD Workshop on Mining Scientific Datasets, 2001.
[20]Taylor, G. H. Impacts of the El Nino/southern oscillation on the Pacific Northwest. Technical report, Oregon State University, USA, 1998.
[21]Hetal Bharat Bhavsar, Anjali Ganesh Jivani, An approach towards the Shared Nearest Neighbor Clustering Algorithm, 2010.
[22]Tsonis, A. A., Swanson, K. L., Wang, G. On the role of atmospheric teleconnections in climate. In Bulletin of the American Meteorological Society, vol.21, issue 12, 2008.
[23]E. Kalnay, et al, 1996. The NCEP/NCAR 40- Year Reanalysis Project Bulletin of the American Meteorological Society, Vol. 77, No. 3. (1 March 1996), pp. 437-470.
[24]https://climatedataguide.ucar.edu/climate-data/palmer-drought-severity-index-pdsi
[25]http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm
[26]http://www.ipcc.ch
[27]http://www.climatechange.cs.umn.edu

Author's Profile:

Mr. J. Velmurugan working as an Associate Professor, Department of Computer Science and Engineering in Sri Venkateswara College of Engineering and Technology, Chittoor. His Area of interest is Data Mining. He is pursuing his Ph.D in VIT University, Vellore and had his M.Tech in Dr.M.G.R. University, Chennai. He obtained Professional membership in ISTE Life Member and has teaching work experience of 8 years.

Ms.Ramireddy Thushara Bharathi received her B.Tech in 2014 in Computer Science and Engineering affiliated with JNTUA, Anantapuramu. She is now pursuing her M.Tech in Department of Computer Science and Engineering, Sri Venkateswara College of Engineering and Technology, Chittoor. Her area of interest is Data Mining.