Discovery and Analysis of Ocean Climate Indices Using DSNN Clustering Algorithm

Ravi D Patel[1], Bhavesh Tamawala[2] ,Kirti Sharma[3]

[1]*PG Scholer, CE Department, BVM Engineering College, aryanrdp@gmail.com*
[2]*Assistant professor, CE Department, BVM Engineering College, Bhavesh.tanawala@bvmengineering.ac.in*
[3]*Assistant professor,CE Department, B.V.M Engineering College, kirti.engineer@gmail.com*

Abstract—This Paper based on finding interesting spatio-tempral pattern from Earth Science data. The data consists measurements of various Earth Science variables (include Temperature and pressure) which are related with time series. Earth Science data has strong seasonal components that needs to be removed prior to pattern analysis, as the Earth Scientist are primarily interested in pattern that represent deviation from normal seasonal variations such as anomalous climate event (e.g. , E1 Nino) or tends (e.g., global warming). We used "monthly" Z Score to remove seasonality. After processing, we apply DSNN clustering algorithm to cluster the temperature time series associated with point on the ocean, yielding clusters that represent ocean regions with relatively homogeneous behavior. The centroids of these clusters are time series that summarize the behavior of these ocean areas and thus, represent potential OCIs (Ocean climate indices).To evaluate cluster centroid for their usefulness, we must determine which cluster centroids significantly influence the land area. For this task, we use variety approaches that analyze the correlation between potential OCIs and time series.

Keywords—Time series analysis, Clustering,Earth science data, scientific data mining.

I. INTRODUCTION

The Land, Ocean and Atmosphere processes are highly coupled i.e. climate phenomena occurring one location can affect the climate at another location. For understanding this affect, climate teleconnection is required to finding how the Earth's climate is changing and how global environment changes. To study teleconnection is by using climate indices, which are climate variability at a regional into a single time series i.e. Nino 1+2 indexes, which is defined as the average sea surface temperature anomaly region of the coast of Peru. Earth observation satellites or sensors are generating increasingly larger amounts of data which are combined with additional data from ecosystem models;create an opportunity forunderstanding andpredicting the behavior of the Earth's global ecosystem or Earth's Climate and how ecosystem responds to global environment change. However due to large amount of data, data mining techniques are used to facilitates the automatic extraction and analysis of interesting pattern from the earth science data.This data consists of sequence of global Earth snapshots of the Earth, typically available at monthly intervals, and include various land and ocean variable such as sea surface temperature (SST), pressure, Net Primary Production (NPP). NPP (Net Primary Production) is the net assimilation of atmospheric carbon dioxide (CO2) into organic matter by plants. Sea Level Pressure (SLP) and Sea Surface Temperature (SST) in Ocean region are the one on which most commonly used climate indices based. More Recently motivated by the massive amounts of new data being produced by satellite observation, Earth Scientist have been using eigenvaluesanalysis techniques such as principal components analysis (PCA) and singular value decomposition (SVD), to discover climate indices.[gcc] Because it's have some limitation, they present an alternative cluster-based methodology for the discovery of climate indices that overcomes limitations of eigenvalues analysis techniques. In this paper, we describe a high-dimensional nearest neighbor clustering (DSNN) algorithm and evaluate it on multi-dimensional spatio-temporal data set which overcomes the limitation of the Shared nearest neighbor (SNN).

The basic outline of this paper is as follows. Section 2 provides a description of the Earth science data that we use in our subsequent analyses; Section 3 discusses techniques to dealing with seasonality Data; and Section 4 shows the Our DSNN clustering approach. Section 5 Sections presents the results of applyingDSNN clustering algorithm to find climate indices that have astrong connection to land temperature. Section 6 provide conclusion and indicates future directions.

II. Earth Sciences Data and Climate Indices

The Earth science data for our analysis consists of globalsnapshots of measurement values for a number of variables(e.g., temperature, pressure and precipitation) collected for all sea surface and land (see Figure 1). These variable values are either observations from different sensors and are typically available at monthly intervalsthat span a range of 10 to 50 years e.g., precipitation, Sea Level Pressure (SLP), sea surface temperature(SST). [1]

For the analysis presentedhere, we focus on attributes measured at points (gridcells) on latitude-longitude spherical grids of different resolutions,e.g., land temperature, which is available at a resolution of 0.5^{o} x $0{:}5^{o}$ and SST, which is available for a grid.Most of the well-known climate indices based uponSST and SLP are shown in Table 1.The spatial and temporal nature of Earth Science creates a number of challenges. Earth Science timeseries data is noisy, has cycles of varying lengths and regularity,and can contain long term trends. In addition,such data displays spatial and temporal autocorrelation, i.e.,measured values that are close in time and space tend to behighly correlated, or similar. [1][3]

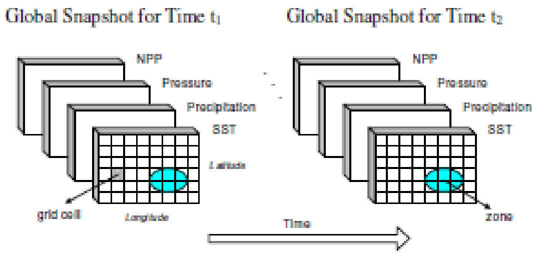


Figure 1: A simplified view of the problem domain.[1]

Table 1: Description of well-known climate indices. [8]

Index	Description
SOI	(southern Oscillation Index) Measure the SLP anomalies between Darwin and Tahiti
NAO	Normalized SLP difference anomalies between Ponta, Delgada, Azores and Stykkisholur, Iceland.
NINO 1+2	Sea surface temperature anomalies in the region bounded by 80^{o}W-90^{o}W and 0^{o}-10^{o} S

NINO 4	Sea surface temperature anomalies in the region bounded by 1500^{o}W-1600^{o}W and 5^{o}S -5^o N
NP	Area-weighted sea level pressure over the region 30N-65N, 160E-140W

III. DEALING WITH THE SEASONALITY OF DATA

Earth science data are spatio-temporal in nature. Mining Pattern derived from Earth Science data are often difficult due to presence of seasonal variation in data. Although Earth scientists are primarily interested in patterns that represent derivation from normal seasonal cycles, such as drought, floods, heats waves; instead of yearly patterns such as spring, summer and winter or rainy season are important. Such events become apparent only if the seasonal components of the climate time series are removed. [1].Monthly Z-Score: This transformation takes the set of values for a given month e.g. all January, calculate the mean and standard deviation for the set of monthly values and then standardizes each value by calculating its Z–Score i.e. by subtracting the mean and divided by the standard deviation. It is quite different than others since it uses the monthly mean and standard deviation of instead of the overall mean and standard deviation. The month-by-month used in this transformation causes seasonal fluctuations to disappear.[3][1].The graph represents in figure 2.1 shows how temperature varies yearly, here this data contain the seasonality data. For finding Earth scientist interested pattern from this data we need to apply Monthly Z-Score. Figure 2.2 shows the result after applying Monthly Z-score in which the seasonality removed from the data.

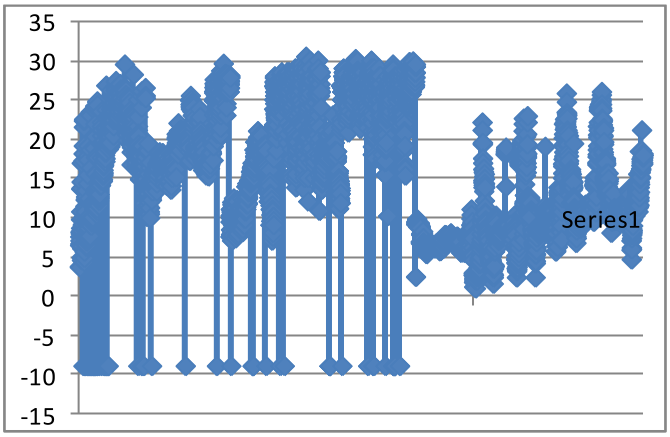


Figure 2.1: Before applying Monthly Z-score

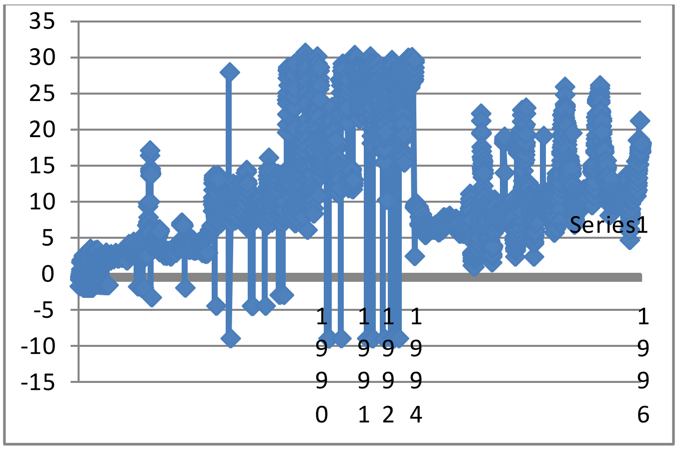


Figure 2.2 :After Applying Monthly Z-score

VI. AN DSNN BASED CLUSTERING APPROACH

There are various methods or technique used for clustering Earth Science Data. We obtain clusters that represent ocean regions with relatively homogeneous behavior by applying clustering algorithm in the temperature time series associated with points on the ocean. The centroids of these clusters are time series that summarize the behavior of these ocean areas, and thus, represent potential OCIs. Consequently, clustering is an initial and key step in using data mining for the discovery of OCIs. [1]

In first paper, they introduce Shared Nearest Neighbor (SNN) for processing our Earth science data to overcome the limitation of K-means clustering algorithm, which is density based algorithm. K-means has disadvantaged that when it is tries to cluster all the data, and because of this, cluster quality suffers greatly, particularly if the data is noisy, as with Earth science data. Also, the number of clusters has to be specified in advance for K-mean clustering. Furthermore, K-means produces clusters that sometimes consist of “chunks” which are geographically widely separated. It can be interesting and useful, for our work in detecting OCIs, we wanted clusters that are geographically contiguous, or nearly so. The clusters produces by SNN clustering algorithm are high quality clusters, which are automatically discovers the “correct” number of clusters, and almost always geographically contiguous. [5]

Here we also find out that there are many disadvantages of SNN clustering algorithm in high-Dimensional Data set. In SNN there is no enough process for outlier, which result in redundant pointless computation and also definition of thresholds for core points, outliers are not clearly provided.Those points with a higher link strength than the threshold is defined as core points. This method with threshold often have an inferior efficiency, since users are required to have a deep understand on spatiotemporal data set and also the procedure defining core points is not good enough that is it’s not exact define core point directly by threshold.[2]

Then we brought the high dimensional nearest neighbor clustering algorithm (DSNN)to overcome SNN’s limitation. This refined algorithm can reduce the spatio-temporal complexity effectively, and refined many performances, such as outliers, core points, clustering results and so on. [2]

VII. CLUSTERING OF OCEAN DATA USING DSNN CLUSTERING ALGORITHM.

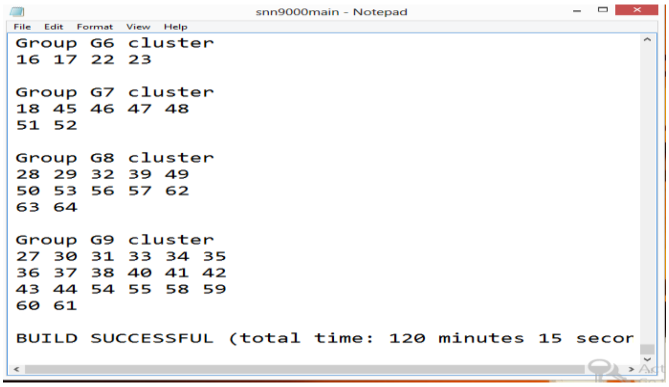
We used DSNN clustering on the one set of data that we have for the ocean, sea surface temperature (SST). Foreach of these data sets we clustered over time periods, from 1990 through 1996. In this technique, in first step all the data given as input to Distance Based Outlier algorithm which gives set of core points and also outlier. In second step, set of core points and set of data given as input to SNN algorithm and perform clustering algorithm on the data. Finds the nearest neighbors of each data point and then redefines the similarity between pairs of points in terms of how many nearest neighbors the two points share. Using this definition of similarity, our algorithm identifies core points and then builds clusters around the core points. The problem with varying densities and high dimensionality are solved by use of a shared nearest neighbor definition of similarityand the use of core points handles problems with shape and size.Furthermore, the number of clusters is automatically determined by the location and distribution of core points. Another novel aspect of the DSNN clustering algorithm is that the resulting clusters do not contain all the points, but rather, contain only points that come from regions of relatively uniform density and also it has separate process for outlier, which reduce the computation effectively. With respect to Earth Science data, DSNN clustering produces high quality clusters, which are almost always geographically contiguous, and automatically selects the number of clusters. [1]

DSNN Clustering Algorithm

1. Transfer an algorithm to find distance-based outliers, in order to achieve better precision with refined sample set.[7]
2. Based on this sample set, Identify the k nearest neighbors for each point in sample set (the k points most similar to a given point, using a distance function to calculate the similarity).
3. Calculate the SNN similarity between pairs of points as the number of nearest neighbors that the two points share. The SNN similarity is zero if the second point in not in its list of k nearest neighbors, and vice-versa.
4. Calculate the SNN density of each point: number of nearest neighbours that share Eps or greater neighbors.
5. Detect the core points. If the SNN density of the point is equal or greater than MinPts then classify the point as core.
6. Form the cluster from the core points. Classify core points into the same cluster if they share Eps or greater neighbors.
7. Identify the noise points. All non-core points that are not within a radius of Eps of a core point are classified as noise.
8. Assign the remainder points to the cluster that contains the most similar core point.[5]

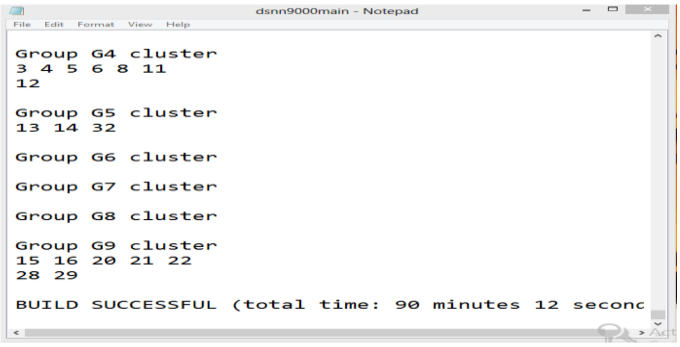
VIII. DSNN AND SNN'S RESULTS

After applying Dsnn and Snn clustering algorithm on high dimensional data following results are taken.



```
Group G6 cluster
16 17 22 23

Group G7 cluster
18 45 46 47 48
51 52

Group G8 cluster
28 29 32 39 49
50 53 56 57 62
63 64

Group G9 cluster
27 30 31 33 34 35
36 37 38 40 41 42
43 44 54 55 58 59
60 61

BUILD SUCCESSFUL (total time: 120 minutes 15 secor
```

Figure:5.1 Output after applying SNN Clustering Algorithm



```
Group G4 cluster
3 4 5 6 8 11
12

Group G5 cluster
13 14 32

Group G6 cluster

Group G7 cluster

Group G8 cluster

Group G9 cluster
15 16 20 21 22
28 29

BUILD SUCCESSFUL (total time: 90 minutes 12 seconc
```

Figure:5.2 Output after applying DSNN Clustering Algorithm

From Figure: 5.1 and 5.2 we can say that using Snn clustering Algorithm for clustering Temperature Data around 10000 (high dimensional data) it will create 66 clusters, while Dsnn clustering algorithm create 33 clusters.

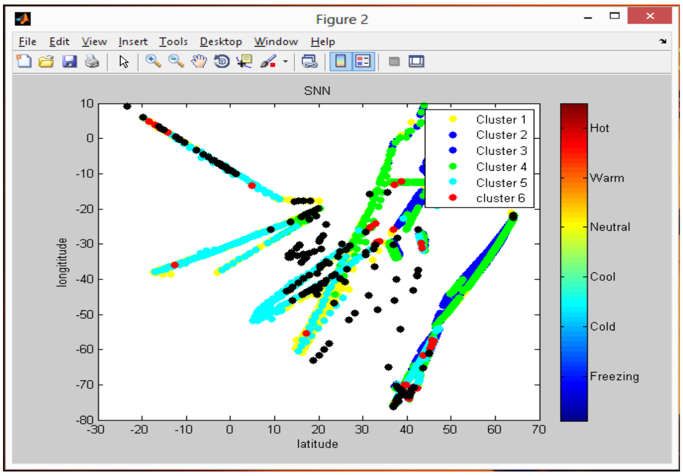


Figure :5.3 SNN Graph

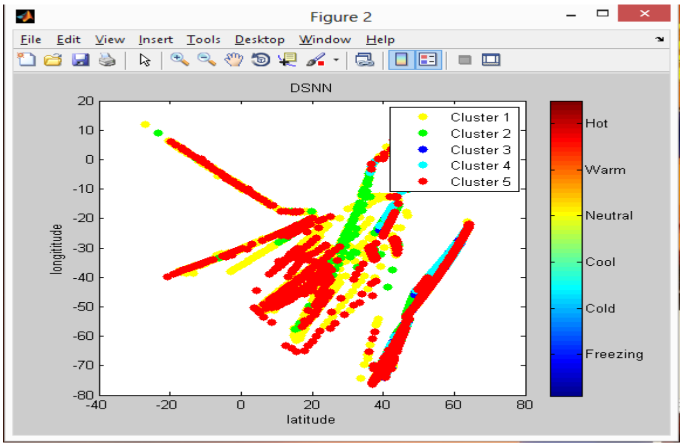


Figure : 5.4 DSNN Graph

From figure 5.3 and 5.4 we can say that DSNN Clustering Algorithm provide same result as SNN Clustering Algorithm provide but DSNN Clustering Algorithm work more on Outlier, Noncore, Core Points while SNN Clustering Algorithm does not.

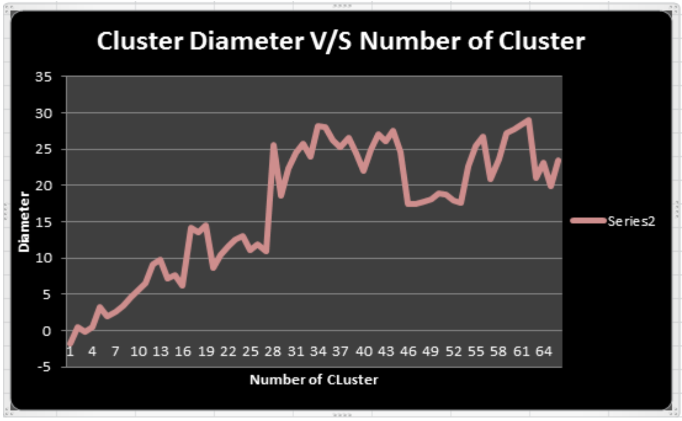


Figure: 5.5 Clusters Diameters Graph for SNN

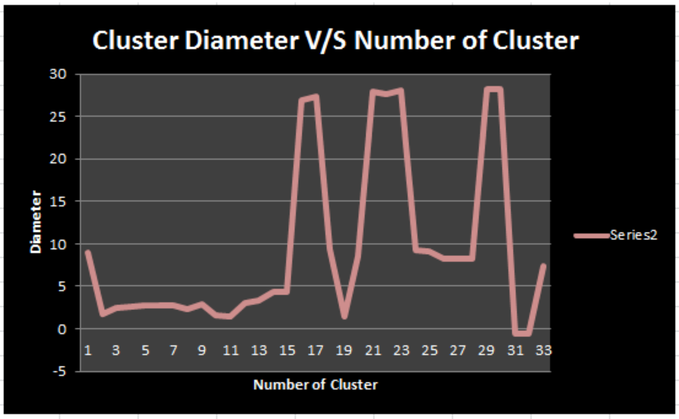
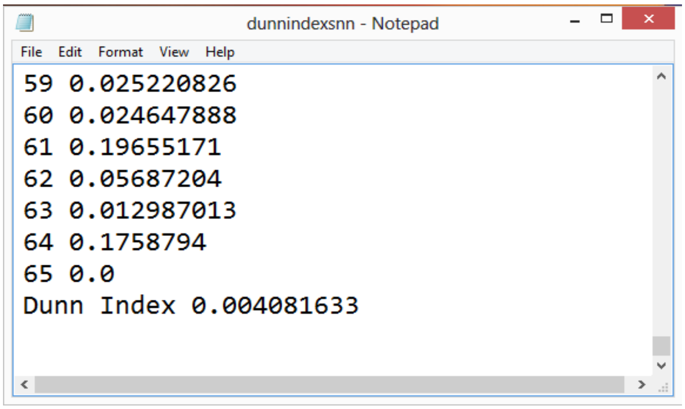


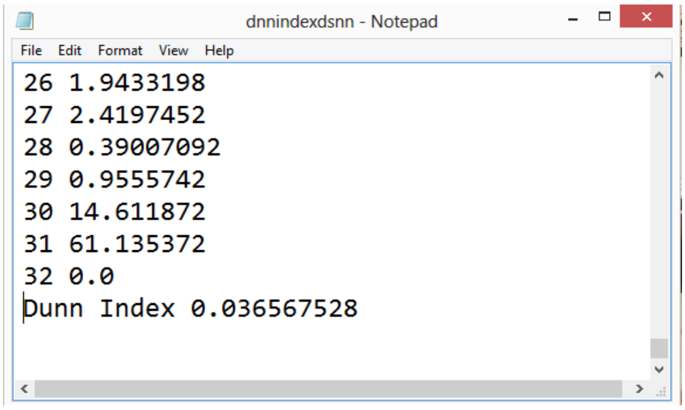
Figure: 5.6 Clusters Diameters Graph for DSNN

Figure 5.5 and 5.6 Shown Graph represents the number of cluster versus Diameter of that clusters. Diameter metric represents that if diameter is small then it's better cluster. From the graph (figure 5.6) we can say that DSNN Clustering Algorithm provide better cluster in comparison SNN Clustering Algorithm.



```
59 0.025220826
60 0.024647888
61 0.19655171
62 0.05687204
63 0.012987013
64 0.1758794
65 0.0
Dunn Index 0.004081633
```

Figure: 5.7 Output Dunn Index For SNN



```
26 1.9433198
27 2.4197452
28 0.39007092
29 0.9555742
30 14.611872
31 61.135372
32 0.0
Dunn Index 0.036567528
```

Figure: 5.8 Output Dunn Index For DSNN

Figure 5.7 and 5.8 represents Dunn Index For SNN and DSNN. Dunn Index is another metric which we used to compare two clustering algorithm. Large value of Dunn Index represents compact and well separated clusters. From Figure DSNN Clustering Algorithm gives higher value of Dunn Index then SNN Clustering Algorithm.

From these practical results we conclude that in high dimensional data DSNN clustering algorithm is better clustering algorithm then SNN clustering algorithm.

X. Conclusion and Future work

In this paper we demonstrated that DSNN clustering algorithm can provide an alternative approach to SNN clusteringalgorithm for finding ocean climate indices. Specially, by using the DSNN clustering algorithm, we found centroids of many clusters of SST data which are correspond to known climate indices and provide a validation of our methodology; other centroids are variants of known indices that may provide better predictive power for some land areas.

From the practical results , we said that DSNN can reduce number of clusters and computation effectively, at the same time, it can accurately judge core points and outliers, and gain better clustering performance than SNN algorithm with better clustering methods.

From the practical results, clusteringcanautomatically identifying regions that may be of interest. We also conclude from practical results that DSNN clustering algorithm is perform better then SNN clustering algorithm in High Dimensional Dataset.

In future work we implement another clustering algorithm on Earth Science Dataset and improve the results.

REFERENCES

[1] Steven Klooster, Christopher Potter, Vipin Kumar, Pang-Ning Tan, Michael Steinbach, "Discovery of Climate Indices using Clustering".In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Year-2003.

[2] Jian Yin,Xianli Fan, Yiqun Chen and Jiangtao Ren, "High-Dimensional Shared Nearest Neighbor Clustering Algorithm", Wang and Y. Jin (Eds.): FSKD 2005, LNAI 3614, pp. 494–502, 2005. Springer-Verlag Berlin Heidelberg 2005.

[3] Vipin Kumar, Pang-Ning Tan, Michael Steinbach,Steven Klooster,Christopher Potter,Alicie Torregrosa, "Mining Scientific Data: Discovery of Patterns in the Global Climate System", .InProceedings of the Joint Statistical Meetings (Athens, GA, Aug. 5–9). American Statistical Association, version-3, Year-2001.

[4] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Steven Klooster,Christopher Potter,Alicie Torregrosa, "Clustering Earth Science Data: Goals, Issues and Results", In Proceedings of the Fourth KDD Workshop on Mining Scientific Datasets,version-2 ,Year-2001.

[5] Michael Steinbach , Vipin Kumar, Steven Klooster, Christopher Potter, Pang-Ning Tan, "Data Mining for the Discovery of Ocean Climate Indices ". In Mining Scientific Datasets Workshop, 2nd Annual SIAM International Conference on Data Mining, Year-2002.

[6] Adriano Moreira, Maribel Y. Santos and Sofia Carneiro, "Density-based clustering algorithm – DBSCAN and SNN".

[7] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Steven Klooster, Christopher Potter,Alicie Torregrosa, "A New Shared Nearest Neighbor Clustering Algorithm and its Application". Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining, Year-2002.

[8] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Steven Klooster, Christopher Potter,Alicie Torregrosa, "Temporal Data Mining for the Discovery and Analysis of Ocean Climate Indices". In Proceedings of the KDD Temporal Data Mining Workshop, Year-2002.

[9] Anil Kumar Patidar, Jitendra Agarwal, Nishcol Mishra, "Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach".International Journal of Computer Applications © 2012 by IJCA Journal Volume 40 - Number 16 Year of Publication: 2012.

[10] Jozef Zurada and Medo Kantardzic, "New Generation of Data Mining Application", A Wiley Interscience Publication.

[11] PoojaMehtaa, Brinda Parekh, KiritModi, and PareshSolanki, "Web Personalization Using Web Mining: Concept and Research Issue". International Journal of Information and Education Technology, Vol. 2, No. 5, October 2012.