Patel et al., International Journal of Advanced Research in Computer Science and Software Engineering 4(1), January - 2014, pp. 742-745

Volume 4, Issue 1, January 2014 **ISSN: 2277 128X**

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Discovery and Analysis of Ocean Climate Indices Using DSNN Clustering Algorithm

Ravi D. Patel
PG scholar Computer Engineering Department, B.V.M Engineering College, India

Bhavesh A. Tanawala
Assistant professor Computer Engineering Department, B.V.M Engineering College, India

Kirti J. Sharma
Assistant professor Computer Engineering Department, B.V.M Engineering College, India

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

NASA's Earth observation satellites are generating increasingly larger amounts of data. This remotely sensed data, combined with additional data from ecosystem models, offers an unprecedented opportunity for predicting and understanding the behavior of the Earth's ecosystem. However, due to the large amount of data that is available, data mining techniques are needed to facilitate the automatic extraction and analysis of interesting patterns from the Earth Science data. Teleconnection are the simultaneous variation in climate and related process over widely separated point on the Earth. For example, E1 Nino, the anomalous warming of the eaten tropical region of the Pacific, has been linked to climate phenomena such as drought in Australia and heavy rainfall along the Eastern coast of South America [Tay98]. For this paper; we will be concerned with teleconnections, such as E1 Nino, that involve the relationship of the ocean to land climate. However due to the large amount of data is available, data mining techniques are used to facilitate 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, and ecologists who work at the regional and global scale have identified NPP as a key variable for understanding the global carbon cycle and the ecological dynamics of the Earth.[1][3]

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 SNN clustering approach. Section 5 Sections presents the results of applying SNN clustering algorithm to find climate indices that have a strong connection to land temperature. Section 6 provide conclusion and indicates future directions.

II. EARTH SCIENCE DATA AND CLIMATE INDICES

The Earth science data for our analysis consists of global snapshots of measurement values for a number of variables (e.g., temperature, pressure and precipitation) collected for all land and sea surfaces (see Figure 1). These variable values are either observations from different sensors, e.g., precipitation, Sea Level Pressure (SLP), sea surface temperature (SST), and are typically available at monthly intervals that span a range of 10 to 50 years. [1] For the analysis presented here, we focus on attributes measured at points (grid cells) 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, and SLP, which is available for a grid. Most of the well-known climate indices based upon SST and SLP are shown in Table 1. The spatial and temporal nature of Earth Science poses a number of challenges. For instance, Earth Science time series data is noisy, has cycles of varying lengths and regularity, and can contain long term1trends. In addition, such data displays spatial and temporal autocorrelation, i.e., measured values that are close in time and space tend to be highly correlated, or similar. [1][3]

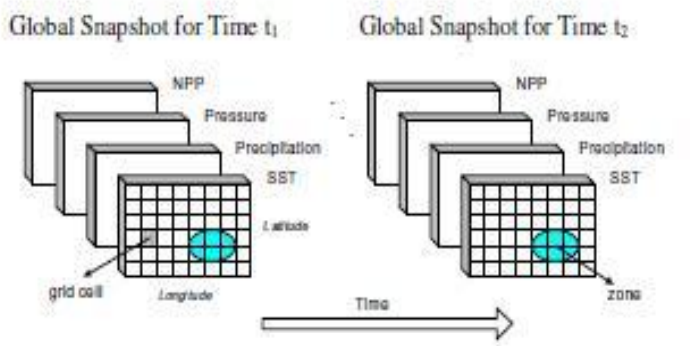

Figure 1: A simplified view of the problem domain. [1]

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^oW-90^oW and 0^o-10^o S
NINO 4	Sea surface temperature anomalies in the region bounded by 1500^oW-1600^oW and 5^oS -5^o N
NP	Area-weighted sea level pressure over the region 30N-65N, 160E-140W

III. DEALING WITH THE SEASONALITY OF DATA

Pattern derived from Earth Science data are often dominated by the presence of seasonal variation in data. Although yearly patterns such as spring, summer and winter or rainy season are important. Earth scientists are primarily interested in patterns that represent derivation from normal seasonal cycles, such as drought, floods, heats waves etc. Such events become apparent only if the seasonal components of the climate time series are removed. [1][11]

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. 2

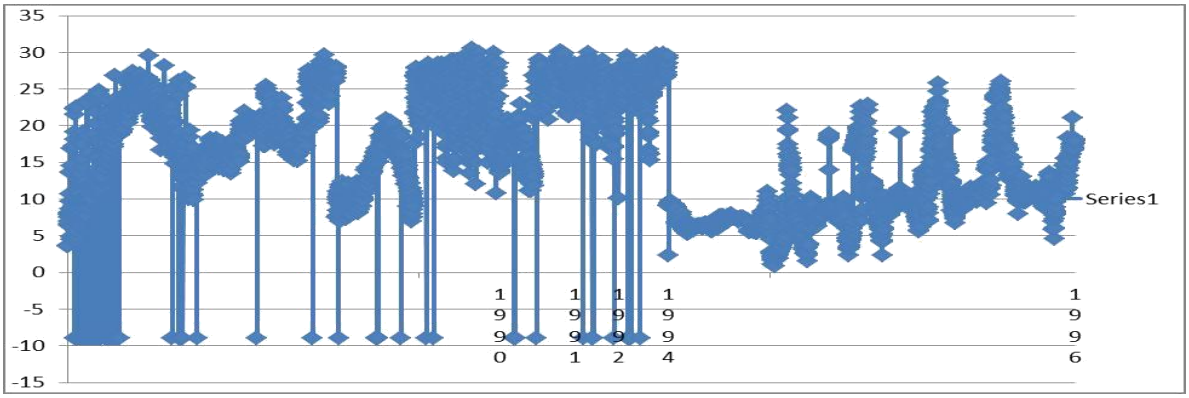

Figure 2.1: Before applying Monthly Z-score

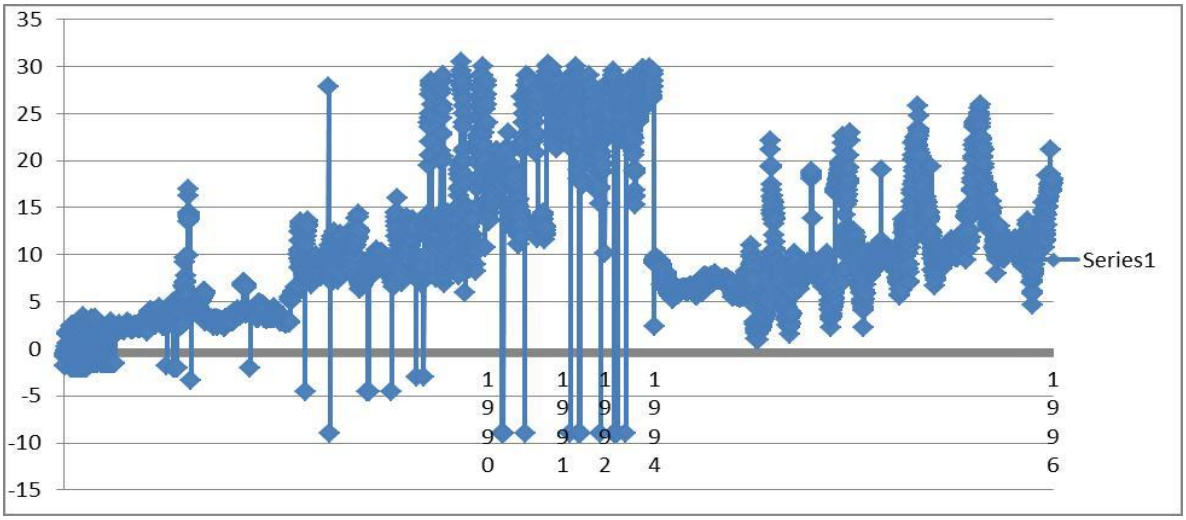

Figure 2.2 :After Applying Monthly Z-score

IV. AN SNN BASED CLUSTERING APPROACH

If we apply a clustering algorithm to cluster the temperature time series associated with points on the ocean, we obtain 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. Consequently, clustering is an initial and key step in using data mining for the discovery of OCIs. [1] For processing our Earth science data, we used K-means clustering algorithm, which is efficient and simple. 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][1]

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 a inferior efficiency, since users are required to have a deep understand on spatiotemporal data set 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]

V. SNN CLUSTERING OF OCEAN DATA

We used SNN clustering on the one set of data that we have for the ocean, sea surface temperature (SST). For each of these data sets we clustered over time periods, from 1990 through 1996. In this technique, it first 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 similarity and 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 SNN 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. These features allow the algorithm to find clusters that other approach over look, i.e., clusters of low or medium density which represent relatively uniform regions surrounded' by non-uniform or higher density areas. With respect to Earth Science data, SNN clustering produces high quality clusters, which are almost always geographically contiguous, and automatically selects the number of clusters. [1][3]

VI. CONCLUSION AND FUTURE WORK

In this paper we demonstrated that clustering can provide an alternative approach to eigenvalue analyses (based on PCA or SVD) for finding ocean climate indices. Specially, by using the SNN 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; and still other indices may represent potentially new Earth science phenomena.[3][1] With the Literature review, we said that DSNN can reduce 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.[2]

From this review we showed that cluster based indices generally outperforms SVD derived indices, both in terms of direct correlation with the known indices. It should be noted that SVD results were obtained by using the data for the entire ocean (SST), for the entire globe. Clustering on the other hand, automatically identifying regions that may be of interest. We also conclude from review that DSNN clustering algorithm is perform better then SNN clustering algorithm in High Dimensional Dataset. [1]

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", .In Proceedings 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 .

