Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 10, October 2014, pg.595 – 604

RESEARCH ARTICLE

Unsupervised Learning on Cosmic Ray Daily Harmonic Variations

Roopesh K. Dwivedi, P.K. Rai*
A.P. S. University Rewa (M.P.)-India
* pkrapsu@gmail.com

Abstract: Clustering is division of data into groups of similar objects. From a machine learning perspective cluster correspond to hidden patterns. In unsupervised learning we find cluster to represent a data concept. Since scientific organizations also generate large volumes of data, the challenges are to analyze the data using the recent data mining techniques, so as to arrive at meaningful conclusions. For real life applications, we have used the hourly cosmic ray intensity data from 1965 to 2006 to first derive for each day, the amplitude and phase of the harmonics of the daily variation (r_1, ϕ_1, and r_2, ϕ_2). We have applied the k-mean partitioning algorithm, the agglomerative hierarchical clustering algorithm BIRCH, and the density based partitioning algorithm DBSCAN on the above set of daily data containing r_1, ϕ_1, and r_2, ϕ_2 for each day. Many interesting clusters have been identified. The cluster analysis indicates that a very clear-cut 10-11 year periodicity is observed in the harmonics dataset even when all the four attributes are considered together. Moreover, similar characteristics are repeated after a gap of 10-11 years and many years occurring in pairs in the two sets (out of the 4 sets, each of about 10-11 years) are the outlier years. The years 1996 and 1997 are particularly emphasized as outliers. These results are similar to that reported in literature, though by statistical methods and by considering only r_1 and ϕ_1 and not all the four attributes taken together. As such the superiority of the mining technique is revealed in the real life situations.

Key Words: Clustering, Data mining, K-mean, BIRCH, DBSCAN, Cosmic ray harmonic

1. Introduction

The process of grouping a set of physical or abstract object is called clustering [JMF99]. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters [D93] [E93]. As a branch of statistics, cluster analysis has been studied extensively for many years. In cluster analysis main focus is on distance based cluster analysis [M96]. Many statistical analysis software packages or systems have built in feature for cluster analysis and they are being used as cluster analysis tools. These tools are based on k-means, k-medoids, and several other methods. Some such type software packages or systems are such as S-Plus, SPSS and SAS.

In machine learning, clustering is an example of unsupervised learning [S80] [JD88]. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labeled examples. Due to this reason, clustering is a form of learning by observation, rather than learning by examples. A group of objects forms a class, only if it is describable by a concept [KR90] in conceptual clustering. Conceptual clustering differs from conventional clustering. Conventional clustering measures similarity based on geometric distance.

Data mining efforts have focused on finding methods for efficient and effective cluster analysis in large database. Active themes of research focus on scalability of clustering methods [G02], the effectiveness of the methods for clustering complex shapes and type of data [K01] [HKT01], high-dimensional clustering techniques, and methods for clustering mixed numerical and categorical data in large databases [F99]. A very good introduction to contemporary data mining clustering techniques can be found in the textbook [HK01].

2. Clustering Technique

There are two main approaches to clustering- partitioning clustering and hierarchical clustering. The partition clustering techniques partition the database into a predefined number of clusters. This uses data partitioning algorithms, (which divide data into k subsets) such as K-Mean [H75] and K-Medioid algorithms or density based partitioning approach such as DBSCAN [EKSX96], GDBSCAN [SEKX98] etc. In hierarchical clustering techniques a sequence of partitions are made in which each partition is nested into the next partition in the sequence. It creates a hierarchy of cluster from small to big or big to small. The hierarchical techniques are classified as - agglomerative and divisive clustering techniques. Agglomerative clustering techniques (such as BIRCH [ZRL96], CURE [GRS98] etc.) starts with as many clusters as there are records, with each cluster having one record. Then pair of clusters are successively merged until the number of clusters reduce to k. At each stage, the pair of clusters that are merged are the ones nearest to each other. Some of the popular hierarchical clustering algorithms are BIRCH [ZRL96], CURE [GRS98] etc. Divisive clustering techniques take the opposite approach from agglomerative techniques. Divisive clustering starts with all the records in one cluster, and then tries to split that cluster into small pieces [HK01].

3. Daily Variations of Cosmic Rays

Solar modulation of galactic cosmic rays is an important aspect in the studies of soar-terrestrial-relationships. While these (cosmic rays) highly energetic positively charged particles traverse the vast interplanetary medium (impinging from the outside of the heliosphere), they are continuously modulated by the solar output and its variations. As such the study of the variability of the cosmic ray particles continuously recorded by the earth based monitors, provide an easy way to understand the fast changing solar output, which modulates these galactic particles both in space and time [R72] [A83].

The spatial variations in the interplanetary space are seen by earth based monitors, in 24-hours, as daily variation (periodic in nature) of cosmic rays, superimposed on which are the aperiodic transient variations of varying magnitudes. For earth based detectors, the spatial variations are local time phenomena, whereas the transient variations are universal time phenomena which occur at the same instant of time all over the globe [L71]. As such, the variety of neutron detectors widely distributed in latitude and longitude of the earth record these transient variations almost simultaneously.

The existence of significant amplitudes of the first two harmonics of the spatial variations (i.e. the daily variations) of cosmic rays observed by neutron and meson monitors have been extensively studied for more than fifty years and their characteristics have been well reviewed in literature from time to time [PD71] [VB90] [SB06]. The pressure-corrected hourly cosmic ray intensity data from high counting rate super neutron monitors are continuously available for a number of neutron monitors since 1965, and so also for the same period, the near earth interplanetary medium parameters (in situ) such as solar wind and interplanetary magnetic field components. As such, it is instructive to study afresh the phenomena of the variability of the daily variations of cosmic rays, for the entire period from 1965 to 2006, using the amplitude and phase of the first two harmonics as a single entity, by applying the mining techniques available for the large databases.

4. Harmonic Variation

For obtaining the harmonics of the daily variations of cosmic rays, for each day, we have used the datasets containing hourly cosmic ray intensity from 1965 to 2006 of Kiel neutron monitor station available in the website www.ngdc.nova.gov. Before using these datasets to calculate the daily harmonic parameters r_1, ϕ_1, r_2 and ϕ_2 (where $r_{1,}$ ϕ_1 are the amplitude and phase of first harmonic and $r_{2,}$ ϕ_2 are the amplitude and phase of second harmonic), we have performed many preprocessing steps like cleaning, transformation and integration on the selected datasets for the entire period. We have also calculated the yearly vector averages of r_1, ϕ_1, r_2 and ϕ_2 from these daily harmonic variations for 'good days' only. For calculating the daily harmonic variations, as well as yearly vector averages, of r_1, ϕ_1, r_2 and ϕ_2 we have developed a comprehensive program in Visual Basic 6.0, which also takes care to ignore a few days in each year known to be contaminated by impulsive universal time transient variations of cosmic rays, thus retaining 'good days' only for analysis presented here, for the entire period from 1965 to 2006.

5. Experimental Results

For real life application, the calculated daily harmonic variations as well as yearly vector averages of r_1, ϕ_1, r_2 and ϕ_2 generated from above mentioned VB program have been used for the period of 1965 to 2006.

Experiment 1: We have applied the k-mean clustering algorithm to divide the daily harmonics data into k-clusters. We have chosen the different values of k = 4, 7, 10 in different observations. In our observation we have selected the standardized attributes of r_1, ϕ_1, r_2 and $\phi_{2,}$ for each day for the entire period from 1965 to 2006, to find the Euclidean distance, which is defined as

$$d(i,j) = \sqrt{} (|x_{i1}-x_{j1}|^2 + |x_{i2}-x_{j2}|^2 + \ldots\ldots\ldots + |x_{ip}-x_{jp}|^2)$$

where x_1, $x_2 \ldots x_p$ are the attributes of the objects and i and j are the objects.

In our experiments the object is represented by (Year, DayNo). Table 1 gives the summarized view of the results obtained for k = 7 clusters, for the entire period of 1956 to 2006, having good days. The year wise distribution of these days, for each cluster, is plotted in fig.1, where in most cases the occurrences of the maxima and / or minima are generally separated by 10-11 years.

ClusterID	No. of Days	Avg (r_1)	Avg(ϕ_1)	Avg(r_2)	Avg(ϕ_2)	Radius of Cluster
1	470	0.9005971	194.7072	0.4939104	23.48623	3.367247
2	3788	0.308137	210.2084	0.1113115	87.37534	1.397146
3	1348	0.2094155	291.139	0.1405728	-23.42823	1.756649
4	1519	0.2314179	82.3452	0.1464616	-5.124278	2.005774
5	2171	0.6803712	209.6058	0.1594635	17.86126	1.737811
6	3408	0.3110979	197.4382	0.1104767	-84.73829	1.422001
7	1832	0.3994687	204.4159	0.2986975	15.10687	1.860185

Table 1: Cluster wise averages of r_1, ϕ_1, r_2 and ϕ_2 along with the radius. Here radius represents the compactness of the cluster.

Experiment2: Here again we have used the same dataset discussed earlier but this time we are using the standardized yearly vector averages of the r_1, ϕ_1, r_2, $\phi_{2,}$ and applied the k-mean clustering technique (by taking optimized k=7). Table 2 gives the summarized view of the results obtained.

Cluster ID	No. of Years	Years	Avg (r_1)	Avg (ϕ_1)	Avg (r_2)	Avg (ϕ_2)	Radius of Cluster
1	5	1965, 1976, 1977, 1986,1987	0.204	207.16	0.032	26.71	1.528
2	9	1966,1967, 1972, 1978, 1979, 1980, 1988, 1992, 2001	0.276	206.70	0.031	12.64	1.249
3	3	1995, 1996, 1997	0.161	165.46	0.031	-43.20	1.376
4	4	1968, 1971, 1974, 2004	0.312	212.17	0.045	27.39	1.046
5	13	1969, 1970, 1981, 1982, 1983, 1984, 1985, 1989, 2000, 2002, 2003, 2005, 2006	0.307	217.47	0.028	35.28	1.068
6	3	1968, 1973, 1975, 1985, 1993, 2004, 2005	0.294	198.74	0.016	33.53	1.594
7	5	1965, 1996, 1997, 2006	0.279	189.84	0.048	-0.15	1.604

Table 2: Cluster wise averages of r_1, ϕ_1, r_2 and ϕ_2 along with the radius. Here radius represents the compactness of the cluster.

Experiment 3: Here again we have used the same data set as in experiment 2, and applied the agglomerative hierarchical clustering algorithm BIRCH. This technique we have used to identify the most similar as well as dissimilar years. Initially all the years are recognized as independent clusters. The years which participate in the formation of the clusters in earlier iterations will be the most similar years and the years which participate in the formation of the cluster at latter passes are recognized as the outliers. To implement this algorithm we have developed a comprehensive program in Visual Basic. The Table 3 shows the result obtained from of this algorithm.

Iteration	ClusterID	Years	Distance Between Clusters
1	C1	**1965,1986**	0.369356691837311
2	C2	**1984,1985**	0.44943380355835
3	C3	**1969,2006**	0.500886201858521
4	C4	**1970,2003**	0.610835909843445
5	C5	**1983,1989**	0.660654306411743
6	C6	1966,C4	0.707045435905457
7	C7	1980,C6	0.693011581897736
8	C8	1981,C3	0.713279068470001
9	C9	2000,2005	0.863325655460358
10	C10	1982,C8	0.911831319332123
11	C11	1987,C1	0.982805967330933
12	C12	C7, C10	0.984204351902008
13	C13	1978,1979	1.02457904815674
14	C14	1988,2001	1.03400552272797
15	C15	1973,1975	1.07646477222443
16	C16	C12, C14	1.11020398139954
17	C17	2002,C5	1.11371171474457
18	C18	1971,1974	1.13175892829895
19	C19	1990,C9	1.17439067363739
20	C20	1972,1977	1.19488656520844
21	C21	C13, C16	1.50459170341492
22	C22	2004,C18	1.55087602138519
23	C23	C17, C19	1.55425572395325
24	C24	1992,20	1.58612680435181
25	C25	1998,C24	1.62597835063934
26	C26	1968,C22	1.76180410385132
27	C27	1994,C15	1.78886198997498
28	C28	C21, C26	1.80191421508789
29	C29	C2, C23	1.8233550786972
30	C30	C28, C29	2.0625364780426
31	C31	**1993**,C27	2.09876656532288
32	C32	**1995,1997**	2.11792087554932
33	C33	**1996**,C32	2.27046942710876
34	C34	**1967**,C25	2.46455574035645
35	C35	**1976**,C31	2.78186106681824
36	C36	C30, C34	2.8255832195282
37	C37	C11, C36	2.69106459617615
38	C38	**1999**, C35	3.37300777435303
39	C39	C37, C38	3.44127893447876
40	C40	**1991**, C39	4.70050859451294
41	C41	C33, C40	6.32997274398804

Table 3: Results obtained from agglomerative hierarchical clustering algorithm “BIRCH” when applied in the dataset of yearly vector averages of r_1, ϕ_1, r_2 and $\phi_{2.}$

Experiment 4: For outlier analysis, again we have taken the same dataset as in experiment 2 and 3, and applied the DBSCAN clustering technique. In this technique we have to appropriately set the parameters ε for finding ε-neighborhood and MinPts to see that the neighborhood is adequately dense or not. If its density does not exceed the threshold MinPts then it is marked as noise objects. Since there are only 42 objects (one for each year) so we have taken MinPts =2. We have done different observation with different ε Value. The years obtained as the noise years in different observations (for different values of ε) are summarized in Table 4

Noise years in different observations (for different values of ε)					
ε = 0.75	ε = 1	ε = 1.25	ε = 1.5	ε =1.75	ε= 2
1968, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2004	1975, 1977, 1991, 1993, 1996, 1997, 1998, 1999, 2004	1991, 1993, 1996, 1997, 1998, 1999, 2004	1991, 1993, 1996, 1997, 1998	1993, 1996, 1997	1996, 1997

Table 4: Results obtained from DBSCAN algorithm when applied in the dataset of yearly vector averages of r_1, ϕ_1, r_2 and $\phi_{2.}$

6. Results and Discussion:

By analyzing the results presented in the above experiments from 1 to 4, many interesting patterns and periodicities have been detected. They are as follows:

(i) A critical examination of table 1, reveals that the average values of the amplitude r_1 and r_2 as well as radius (radius denotes the degree of compactness of the cluster) are very high in cluster C1 in comparison to other clusters. As such, the days which belong to this cluster are the outlier days. This means that on these days of cluster C1, r_1, $r_{2,}$ ϕ_1, ϕ_2 are sparsely distributed showing un-associativeness. By observing Figure 1 (which plots the year-wise number of days distributed in different clusters), the 10-11 year periodicity have been detected in almost all the clusters, including cluster C1 which belongs to outlier days i.e. where r_1, ϕ_1, r_2, ϕ_2 differ widely.

(ii) By thoroughly observing column "Year" of table 2, it is apparent that most of the years of 1970's and 1990's are quite distributed in different clusters whereas the most of the years of 1980's and 2000's occur together in cluster C5 and the radius of the cluster C5 is very low. It means that these years are quite similar in their characteristics. This result indicates the 10-11 year periodicity as detected in (i). Also by observing clusters C1, C6 and C7, where the radius of the clusters are relatively high, the years belonging to these clusters are either consecutive or having 10-11 year periodicity, as was seen in (i) above.

(iii) A critical examination of the table 3, it is noticed here again that the most similar years are either consecutive or having periodicity of the order of 10-11 years. Also, most of the years of 1990's have been identified as special years (i.e. outliers).

(iv) By observing table 4, for $\varepsilon = 0.75$, it is seen that the most of the years of 1970's and 1990's are identified as noise years or outliers. Here again we observed the 10-11 year periodicity as detected in (i), (ii) and (iii). Also for $\varepsilon = 2.0$ the years 1996, 1997 are detected as noise years or outliers i.e. these years are significantly different than all other years when all the 4 components r_1, ϕ_1, r_2, ϕ_2 are considered together as one entity for each year, in the entire period of study 1965-2006.

Since these results are obtained just by applying the mining methodology, it would be instructive to compare the findings with those obtained by simple statistical techniques and reported in literature [SB06], considering the four attributes r_1, ϕ_1, r_2 and $\phi_{2,}$ one-by-one separately. It is important to underline that it is only the mining techniques, which considers all the four attributes together as one entity for each year. The results of 10-11 year periodicity, and a gap of 10-11 years in the similarity characteristics and the outliers (particularly the years 1996 and 1997), have been identified earlier by doing statistical analysis and by using only r_1, ϕ_1 values but not in combination with r_2, $\phi_{2,}$ which comes out here only by the mining techniques so easily. As such, the experimental results 1 to 4 have proved beyond doubt the significance of the mining technique in the real life situation also.

7. Conclusion

Since clustering is an unsupervised learning technique, so the results obtained form (i) to (iv) must be further explored and analyzed by some of the more directed data mining techniques for better understanding of the behavior of the daily harmonic variation parameters r_1, ϕ_1, r_2 and $\phi_{2,}$ both on a day-to-day as well as on annual average basis. In future, interplanetary medium parameters can also be clubbed for better understanding of the link between them and cosmic ray variations by using clustering techniques.

Acknowledgement: The authors would like to thank Prof. S.P. Agrawal for his valuable suggestions and insights on the database as well as the daily harmonic variation parameters amplitude and phase, which have been used here as a real life application for the unsupervised learning. We would also like to thank the organizations and personnel's, those who are constantly involved in updating the cosmic ray databases and providing it in public domain through World Wide Web.

References

[A83] Agrawal, S.P., "Solar Cycle Variations of Cosmic Ray Intensity and Large-Scale Structure of the Heliosphere", Space Science Reviews, Springer Netherlands, Vol.34, No.2, pp. 127-135, 1983.

[D93] Dubes, R.C., "Cluster Analysis and Related Issues". In Chen, C.H., Pau, L.F., and Wang, P.S. (Eds.) Handbook of Pattern Recognition and Computer Vision, World Scientific Publishing Co., River Edge, NJ.47, pp. 3-32, 1993.

[E93] Everitt, B., "Cluster Analysis (3rd ed.)". Edward Arnold, London, UK, 1993.

[EKSX96] Easter, M., Kriegel, H.-P., Sander, J. and Xu, X., "A Density-Based Algorithms for Discovering Clusters in Large Spatial Databases with Noise", In Proceeding 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, pp. 226-231, 1996.

[F99] Fasulo, D., "An analysis of recent work on clustering algorithms". Technical Report, UW-CSE01 -03-02, University of Washington, 1999.

[G02] Ghosh, J., "Scalable Clustering Methods for Data Mining". In Nong Ye (Ed.), Handbook of Data Mining, Lawrence Erlbaum, 2002.

[GRS98] Guha, S., Rastogi, R. and Shim, K., "CURE: An Efficient Clustering Algorithm for Large Databases". In Proc. of the ACM SIGMOD Conference, Seattle, W.A., pp. 73-84,1998.

[H75] Hartigan, J., "Clustering Algorithms". John Wiley & Sons, New York, NY, 1975.

[HKT01] Han, J., Kamber, M., and Tung, A. K. H., "Spatial clustering methods in data mining: A survey". In Miller, H. and Han, J. (Eds.) Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2001.

[JD88] Jain, A., and Dubes, R., "Algorithms for Clustering Data". Prentice-Hall, Englewood Cliffs, NJ., 1988.

[JMF99] Jain, A.K, Murty, M.N., and Flynn, P.J., "Data Clustering: A Review". ACM Computing Surveys, 31, 3, pp. 264-323, 1999.

[K01] Kolatch E., “Clustering Algorithms for Spatial Databases: A Survey”. PDF is available on the Web, 2001.

[KR90] Kaufman, L. and Rousseeuw, P., “Finding Groups in Data: An Introduction to Cluster Analysis”. John Wiley and Sons, New York, NY, 1990.

[L71] Lockwood, J.A., “Forbush Decreases in Cosmic Radiator”, Space Science Reviews, Springer Neherlands, Vol.12, No.5, pp. 658-715, 1971.

[M96] Mirkin, B., “Mathematic Classification and Clustering”. Kluwer Academic Publishers, 1996.

[PD71] Pomerantz, M.A. and Duggal, S.P., “The Cosmic Ray Solar Diurnal Anisotropy”, Space Science Reviews, Springer Netherlands, Vol.12, No.1, pp. 75-130, 1971.

[R72] Rao, U.R., “Solar Modulation of Galactic Cosmic Radiation”, Space Science Reviews, Springer Netherlands, Vol.12, No.6, pp. 719-809, 1972.

[S80] Spath, “Cluster Analysis Algorithms”. Ellis Horwood, Chichester, England, 1980.

[SB06] Singh, M. and Badruddin, “Study of the Cosmic Ray Diurnal Anisotropy During Different Solar and Magnetic Conditions”, Solar Physics, pp. 291-317, 2006.

[SEKX98] Sander, J., Ester, M., Kriegel, H.-P., and Xu, X., ” Density-Based Clustering in Special Database: the algorithm GDBSCAN and its application ”. In Data Mining and Knowledge Discovery, 2, 2, pp. 169-194, 1998.

[VB90] Venkatesan, D. and Badruddin, “Cosmic-Ray Intensity Variations in the 3-Dimensional Heliosphere”, Space Science Reviews, Kluwer Academic Publisher, Vol.52, pp. 121-194, 1990.

[ZRL96] Zhang, T., Ramakrishnan, R., and Livny, M., “BIRCH: An efficient data clustering method for very large databases”, In Proceeding of ACM SIGMOD International Conference on Management of Data, pp. 103-114, 1996.