# CLUSTERING TECHNIQUES FOR UNSUPERVISED LEARNING

**P.K. Rai***

**Roopesh K. Dwivedi***

## Abstract

In data mining, division of data into groups of similar objects is known as clustering. From a machine learning perspective cluster correspond to hidden patterns, the search for cluster is unsupervised learning, and the resulting system represents a data concept. From a practical point of view clustering plays an outstanding role in data mining application such as scientific data exploration etc. For doing the scientific data exploration we have used solar interplanetary and geomagnetic data (SIGD) of a long period from the years 1965 to 2006. We have applied the agglomerative hierarchical clustering algorithm and k-mean partitioning algorithm on the data and many interesting clusters have been identified and discussed in this paper.

**Key Words**: Data Mining, Clustering, Agglomerative, K-mean, DBSCAN, SIGD.

* Head, Computer Centre, A.P. S. University Rewa (M.P.) India.

**1. Introduction**: The process of grouping a set of physical or abstract object is called clustering. A cluster is a collection of similar data objects that are similar to one another within the same cluster and are different to the objects in other clusters. Cluster analysis has been studied comprehensively for many years in the subject Statistics, focusing mainly on distance based cluster analysis. Cluster analysis tools based on k-means, k-medoids, and several other methods have been built into many statistical analysis software packages or systems such as S-Plus, SPSS and SAS.

Clustering is an example of unsupervised learning in the field of machine learning. The goal of clustering is basically to find data points that naturally group together, splitting the full data set into a set of categories. Therefore, clustering and unsupervised learning do not rely on predefined classes and class-labeled examples. For this reason clustering is form of learning by observation, rather than learning by examples. In conceptual clustering, a group of objects forms a class only if it is describable by a concept. This differs from conventional clustering, which 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, the effectiveness of the methods for clustering complex shapes and type of data, high-dimensional clustering techniques, and methods for clustering mixed numerical and categorical data in large databases.

General references regarding clustering include [H75] [S80] [JD88] [KR90] [D93] [E93] [M96] [JMF99] [F99] [K01] [HKT01] [G02]. The textbook [HK01] contains very good introduction to contemporary data mining clustering techniques.

There is a close association between clustering techniques and many other disciplines. Clustering has always been used in statistics and science. The classic introduction into pattern recognition framework is given in [DH73]. Typical applications include speech and character recognition. Machine learning clustering algorithms were applied to image segmentation and computer vision [JF96]. Clustering can be viewed as a density estimation problem. Clustering is

also widely used for data compression in image processing, which is also known as vector quantization [GG92].

This paper has an emphasis on clustering in data mining. Such clustering is characterized by large datasets with many attributes of different types. Clustering in data mining was brought to life by intense developments in information retrieval and text mining [CKPT92] [SKK00] [ DFG01], spatial database applications, for example, GIS or astronomical data, [XEKS98] [SEKX98] [EFKS00], sequence and heterogeneous data analysis [CSM01], Web applications [CMS99] [HC01] [FWZ01], DNA analysis in computational biology [BY99], and many others. A large amount of application-specific developments are resulted that are beyond the scope of this paper.

**2. Clustering Paradigm**: 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 use data partitioning algorithms, which divide data into k subsets. They attempt to determine k-partitions that optimize a certain criterion function. The partition clustering algorithms are of two types: (i) K-mean algorithms: where each cluster is represented by the center of gravity of the cluster and (ii) K-medoid Algorithms: where each cluster is represented by one of the objects of the cluster located near to center.

The hierarchical clustering techniques do a sequence of partitions, 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 of two types- agglomerative and divisive clustering techniques. Agglomerative clustering techniques start 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. If the merging is continued, it terminates in a hierarchy of cluster which is built with just a single cluster containing all the records, at the top of the hierarchy. The opposite approach is taken in divisive clustering techniques from agglomerative techniques. This starts with all the records in one cluster, and then tries to split that cluster into small pieces.

**3. Experimental Results:** For real life application, the solar interplanetary and geomagnetic data (SIGD) have been used for the years 1965 to 2006. This SIGD data is available in the net in different files and in different formats. The SIGD data consist of variety of attributes that include Year, Month, Day, rotation number (RNO), solar wind speed (V), B, Bz, Ap, Kp and the derived value of VB (=V*B).

By performing so many preprocessing steps (e.g. cleaning, transformation, standardization etc) we have transformed the data in a single database file SIGD to apply the proposed clustering techniques in this hope that there may be some patterns or periodicity in the cluster as well as for doing outlier analysis. For doing this so many comprehensive programs have been developed using Visual Basic 6.0.

**Experiment1:** We have applied the k-mean clustering algorithm to divide the data into k-clusters. We have chosen the different value of k = 4, 7, 10 in different observations. In our observation we have selected the standardized attributes V,B, Ap and VB 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, Month and Day). Table1.1 give the summarize view of the results obtained

| ClusterID | No. of Days | Avg (V) | Avg(B) | Avg(Ap) | Avg(VB) | Radius of Cluster |
|---|---|---|---|---|---|---|
| C1 | 2809 | 390 | 7.4 | 10 | 2881 | 0.97 |
| C2 | 1109 | 633 | 6.8 | 27.9 | 4284 | 1.61 |
| C3 | 94 | 649 | 17.6 | 112 | 11393 | 5.72 |
| C4 | 2296 | 516 | 5.4 | 12.5 | 2797 | 1.1 |
| C5 | 3855 | 371 | 4.6 | 5.3 | 1703 | 0.85 |
| C6 | 467 | 524 | 12.9 | 49.3 | 6652 | 2.93 |
| C7 | 1408 | 452 | 10.4 | 20.2 | 4643 | 1.64 |

Table 1.1 Cluster wise average V, B, Ap and VB with radius. Here radius represent the compactness of the cluster.

**Experiment2:** Here again we use the same dataset discussed earlier but this time we are using the standardized yearly average of the V, B, Ap and VB and applied the k-mean clustering techniques on the dataset ( by taking optimized k=7). Table 1.2 give the summarize view of the results obtained.

| Cluster ID | No. of Years | Years | Avg (V) | Avg (B) | Avg (Ap) | Avg (VB) | Radius of Cluster |
|---|---|---|---|---|---|---|---|
| C1 | 8 | 1967,1969,1970,1972, 1977,1980,1987,1988 | 413 | 6.4 | 11.4 | 2682 | 0.69 |
| C2 | 5 | 1966,1971,1976,1986, 1995 | 444 | 5.8 | 12.7 | 2639 | 0.63 |
| C3 | 5 | 1982,1983,1984,1989, 1991 | 469 | 8.4 | 20.6 | 4018 | 1.12 |
| C4 | 10 | 1978,1979,1981,1988, 1990,1992,1999,2000, 2001,2002 | 431 | 7.4 | 14.7 | 3264 | 0.91 |
| C5 | 3 | 1974,1994,2003 | 528 | 6.8 | 19.6 | 3613 | 1.09 |
| C6 | 7 | 1968,1973,1975,1985, 1993, 2004,2005 | 469 | 6.2 | 14.2 | 2972 | 0.70 |
| C7 | 4 | 1965,1996,1997,2006 | 410 | 5.2 | 8.3 | 2145 | 0.73 |

Table 1.2 Cluster wise average V, B, Ap and VB with radius. Here radius represent the compactness of the cluster.

**Experiment 3:** Here again we use the same data set as experiment 2, and applied the agglomerative hierarchical clustering technique. Here again we have summarized the result in the following table 1.3.

| **Cluster ID** | **Years** | **Distance** |
|---|---|---|
| C1 | 1965, 2006 | 0.044 |
| C2 | 1969, 1977 | 0.21 |

| | | |
|---|---|---|
| C3 | 1968, 2005 | 0.22 |
| C4 | 1967, 1987 | 0.25 |
| C5 | 1983,1984 | 0.33 |
| C6 | C3, 1985 | 0.35 |
| C7 | 1993,2004 | 0.36 |
| C8 | C4, 1970 | 0.492 |
| C9 | 1990, 2000 | 0.494 |
| C10 | 1999, 2001 | 0.50 |
| C11 | 1976,1986 | 0.504 |
| C12 | C7, 1966 | 0.55 |
| C13 | C10, 1988 | 0.553 |
| C14 | 1971,1995 | 0.56 |
| C15 | 1982,1991 | 0.574 |
| C16 | C2, C8 | 0.578 |
| C17 | C1, 1996 | 0.635 |
| C18 | 1974,1994 | 0.642 |
| C19 | 1981,1992 | 0.690 |
| C20 | C6, 1975 | 0.695 |
| C21 | C13, 2002 | 0.71 |
| C22 | C16, 1972 | 0.72 |
| C23 | 1978,1979 | 0.77 |
| C24 | C22, 1998 | 0.78 |
| C25 | C11, C14 | 0.85 |
| C26 | C5,1989 | 0.86 |
| C27 | C24, 1980 | 0.89 |
| C28 | C9, C21 | 0.91 |
| C29 | C23, C28 | 0.85 |
| C30 | C12,C25 | 0.92 |
| C31 | C20, 230 | 1.06 |
| C32 | C19, C29 | 1.07 |
| C33 | C18, 1973 | 1.54 |
| C34 | C17, 1997 | 1.59 |
| C35 | C15, C26 | 1.74 |
| C36 | C27, C34 | 2.04 |
| C37 | C33, 2003 | 2.17 |
| C38 | C31, C36 | 2.53 |
| C39 | C32, C38 | 2.50 |
| C40 | C35, C37 | 2.95 |
| C41 | C39, C40 | 3.51 |

Table 1.3 Representation of the agglomerative hierarchical clustering process.

**Experiment 4:** For outlier analysis, again we have taken the same dataset as experiment 3, and applied the DBSCAN hierarchical clustering technique. In this technique we have to appropriately set the parameters € for finding €-neighborhood and MinPts to see whether the neighborhood is adequately dense or not. If its density does not exceed the threshold MinPts then it is marked as noise objects. We have done different observations with different € and MinPts Value. The optimized result obtained from €=1 and MinPts =1. For DBSCAN algorithm see reference [P01].

Output: Year 1973, 1997 and 2003 recognized as Noise and can be taken as outliers.

**4. Result and Discussion:**

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

**(i)** By observing the table 1.1, the average value for Ap in the clusters C2, C3, C6 and C7 are quite high and the radius of the clusters are also high in the order of Average Ap values. The high radius of the clusters indicate the most un-predictable behavior of V, B and Ap. Though by thoroughly observing the table it is also seen that if V is high and B is moderate low or its vice versa (besides both are high), then also the geomagnetic disturbance index are high(>20).

**(ii)** By thoroughly observing column "Year" of table 1.2, a 10-11 year periodicity is also identified. Here again analyzing cluster C3 and C5 also indicate the same result as (i).

**(iii)** By Thoroughly observing the table 1.3 , The 10-11 year periodicity have been clearly identified. Also year 1973, 1997 and 2003 have been identified special years or as outliers.

**(iv)** The output of experiment-4 confirms the result obtained in (iii).

**Conclusion:** Since clustering is an unsupervised learning technique, so the result obtained form (i) to (iv) must be further explored and analyzed by some of the more directed data mining techniques for better understanding the behavior of the solar interplanetary data values like solar wind speed (V), total interplanetary magnetic field IMF (B), and the product VB as well as geomagnetic disturbance index Ap.

**Acknowledgement:** The authors would like to thank Prof. S.P. Agrawal, ex vice-chancellor, APS University, Rewa(MP) India for his valuable suggestions and insights on the database, which have been used here as a real life application for efficiently mining the quantitative association rules. We would also like to thank the organizations and personnel's, those who are constantly involved in updating the solar interplanetary and geomagnetic databases and providing it in public domain through word wide web.

**References**

[BY99] Ben-Dor, A. and Yakhini, Z., "Clustering gene expression patterns". In Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB 99), 11-14, Lyon, France, 1999.

[CKPT92] Cutting, D., Karger, D., Pedersen, J., and Tukey, J., "Scatter/gather: a clusterbased approach to browsing large document collection". In Proceedings of the 15th ACM SIGIR Conference, 318-329, Copenhagen, Denmark, 1992.

[CMS99] Cooley, R., Mobasher, B., and Srivastava, J., "Data preparation for mining world wide web browsing". Journal of Knowledge Information Systems, 1, 1, 5-32, 1999.

[CSM01] Cadez, I., Smyth, P., and Mannila, H., "Probabilistic modeling of transactional data with applications to profiling, Visualization, and Prediction". In Proceedings of the 7th ACM SIGKDD, 37-46, San Francisco, CA., 2001.

[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, 3-32, World Scientific Publishing Co., River Edge, NJ.47, 1993.

[DFG01] Dhillon, I., Fan, J., and Guan, Y., "Efficient clustering of very large document collections". In Grossman, R.L., Kamath, C., Kegelmeyer, P., Kumar, V., and Namburu, R.R. (Eds.) Data Mining for Scientific and Engineering Applications, Kluwer Academic Publishers, 2001.

[DH73] Duda, R. and Hart, P., "Pattern Classification and Scene Analysis". John Wiley & Sons, New York, NY, 1973.

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

[EFKS00] Ester, M., Frommelt, A., Kriegel, H.-P., and Sander J., " Spatial data mining: database primitives, algorithms and efficient DBMS support". Data Mining and Knowledge Discovery, Kluwer Academic Publishers, 4, 2-3, 193-216, 2000.

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

[FWZ01] Foss, A., Wang, W., and Zaane, O., "A non-parametric approach to Web log analysis". 1st SIAM ICDM, Workshop on Web Mining, 41-50, Chicago, IL.48, 2001.

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

[GG92] Gersho, A. and Gray, R. M., "Vector Quantization and Signal Compression. Communications and Information Theory", Kluwer Academic Publishers, Norwell, MA, 1992.

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

[HC01] Heer, J. and Chi, E., "Identification of Web user traffic composition using multimodal clustering and information scent". 1st SIAM ICDM, Workshop on Web Mining, 51-58, Chicago, IL, 2001.

[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.

[JF66] Jain, A.K. and Flynn, P.J., "Image segmentation using clustering". In Advances in Image Understanding: A Festschrift for Azriel Rosenfeld, IEEE Press, 65-83, 1966.

[JMF99] Jain, A.K, Murty, M.N., and Flynn, P.J., "Data clustering: a review". ACM Computing Surveys, 31, 3, 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.

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

[P01] Pujari, A.K., "Data Mining Techniques", Universities Press, 2001.

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

[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, 169-194, 1998.

[SKK00] Steinbach, M., Karypis, G., and Kumar, V., "A comparison of document clustering techniques". 6th ACM SIGKDD, World Text Mining Conference, Boston, MA, 2000.

[XEKS98] Xu, X., Ester, M., Kriegel, H.-P., and Sander, J., " A distribution based clustering algorithm for mining in large special databases". In proceeding of the 14th ICDE, 324-331, Orlando,FL.