Global Journal of Computer Science and Technology
Volume 11 Issue 22 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Classification of Arial Data Based on Data Mining Clustering Algorithm

By Prof.G.Ramaswamy, Dr. Vuda.Sreenivasarao, Dr.Popuri.Ramesh Babu, P.V.S.S.Gangadhar

Defence University college

Abstract - The Arial data contains date periodically observed with parameters of texture (min, max), flora, and density (min, max). The proposed Arial prediction system cluster and analyze, three input features that is average texture, flora, average density according to number of days to predict Arial for Surveillance applications. The proposed system realizes the k-means clustering algorithm for grouping similar features based on user intended period, further the system analyze using PCA (Principal Component Analysis) on same data.

Keywords : Data mining, Arial data, cluster algorithm, Principal Component Analysis.

GJCST Classification : H.2.8

A Classification of Arial Data Based on Data Mining Clustering Algorithm

Prof.G.Ramaswamy[α], Dr. Vuda.Sreenivasarao[Ω], Dr.Popuri.Ramesh Babu[β], P.V.S.S.Gangadhar[ψ]

Abstract -The Arial data contains date periodically observed with parameters of texture (min, max), flora, and density (min, max). The proposed Arial prediction system cluster and analyze, three input features that is average texture, flora, average density according to number of days to predict Arial for Surveillance applications. The proposed system realizes the k-means clustering algorithm for grouping similar features based on user intended period, further the system analyze using PCA (Principal Component Analysis) on same data.

Keywords : Data mining, Arial data, cluster algorithm, Principal Component Analysis.

I. Introduction

The Arial data contains date periodically observed with parameters of texture (min, max), flora, and density (min, max). The amount of data kept in computer files and databases is growing at a phenomenal rate. At the same time, the users of these data are expecting more sophisticated information from them. For example a marketing manager is no longer satisfied with a simple listing of marketing contacts, but wants detailed information about customers past purchases as well as predictions of future purchases. Simple structured/query language queries are not adequate to support these increased demands for information. Data mining steps in to solve these needs. Data mining is often defined as finding hidden information in a database. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association, sequence or path analysis, classification, clustering, and forecasting. This paper deals with implementation of an automated Arial prediction system for agriculture applications using data mining tools such as clustering (k-means algorithm) and Principle Component Analysis (PCA). For making accurate decision on large observation is an important factor, but with increasing information the clustering algorithm faces various limitations/problems. Among them current clustering techniques do not address all the requirements adequately (and concurrently). Dealing with large number of dimensions and large number of data items can be problematic because of time complexity. The effectiveness of the method depends on the definition of "distance" (for distance-based clustering). If an obvious distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces. The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways. The main objective of this paper is to develop an accurate and efficient Arial prediction system for agriculture applications using data mining tools such as clustering (k-means algorithm) and PCA.

II. Data Clustering

Clustering is a divided number of groups of similar data objects. Each group called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in numerical analysis, mathematics and statistics. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. In practical perspective clustering plays an outstanding performance in data mining applications such as, computational biology ,information retrieval and text mining, scientific data exploration ,marketing, medical diagnostics, spatial database applications, and Web analysis, etc.

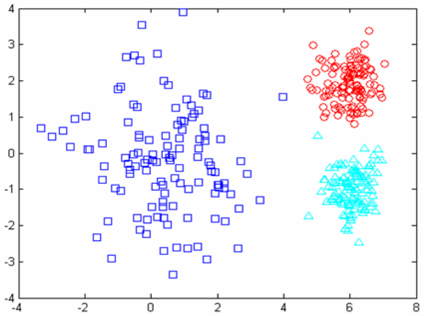

Figure1 : Clusters distribution of a data set

Author[α] : Principal, Priyadarshini Inst.Of Tech.&Sci, Guntur, India.
Author[Ω] : Professor in CIT, Defence University college, Debrezeit, Ethiopia.
Author[β] : Dean Faculty of Engineering, Malineni Perumallu Education, Society's Group of Institutions, India.
Author[ψ] : Scientist C, NIC, Govt of India, India.

Clustering is the subject of active research in several fields such as statistics, pattern recognition, data mining, grouping and decision making, pattern classification, bio informatics and machine learning. A very important characteristic of most of these application domains is that the size of the data involved is very large. So, clustering algorithms used in these application areas should be able to handle large data sets of sizes ranging from gigabytes to terabytes and even pica bytes. Typically, clustering algorithms paper on pattern matrices, where each row of the matrix corresponds to a distinct pattern and each column corresponds to a feature. Most of the early paper on clustering dealt with the problem of grouping small data sets, where the benchmark data sets used to demonstrate the performance of the clustering algorithms were having a few hundreds of patterns and a few tens of features. Fisher's Iris data is one of the most frequently used benchmark data sets. This data has three classes, where each class has 50 patterns and each pattern is represented using four features. However, several real-world problems of current interest are very large in terms of the pattern matrices involved. For example, in data mining and web mining the number of patterns is typically very large, whereas in clustering biological sequences, the number of features involved is very large. Cluster analysis is a way to examine similarities and dissimilarities of observations or objects. Data often fall naturally into groups, or clusters, of observations, where the characteristics of objects in the same cluster are similar and the characteristics of objects in different clusters are dissimilar. In one of the earliest books on data clustering, Underberg defines cluster analysis as a task, which aims to finding of natural groups from a data set, when little or nothing is known about the category structure. Bailey, who surveys the methodology from the sociological perspective, defines that cluster analysis seeks to divide a set of objects into a small number of relatively homogeneous groups on the basis of their similarity over N variables. N is the total number of variables in this case.

III. System Architecture

The Arial prediction system architecture is shown in figure 2.The overall system design consists of Input (Arial Data), modified input, Feature selection, Clustering Using k-means on Selected Feature And PCA on Selected Feature.

Output:

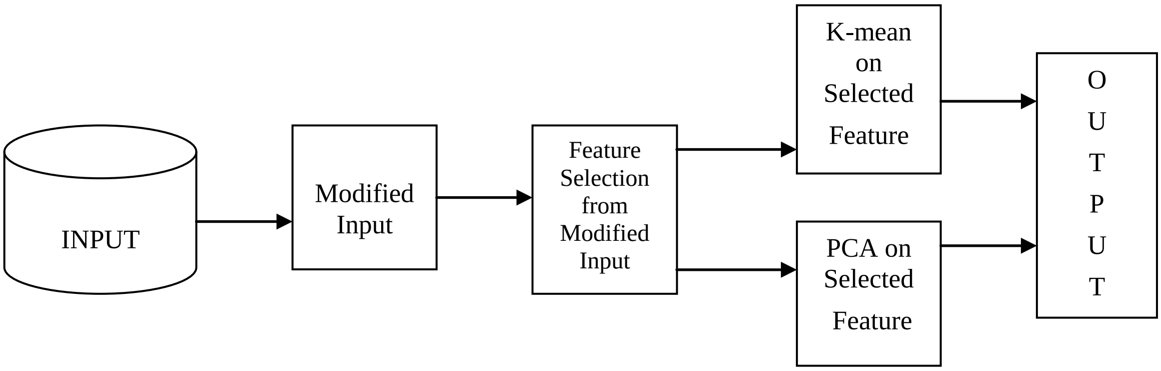
INPUT
Modified Input
Feature Selection from Modified Input
K-mean on Selected Feature
PCA on Selected Feature
O U T P U T

Figure 2 : Architecture of Arial prediction system

Apply k-means clustering algorithm on selected feature. Clustering is a division of data into groups of similar objects. Each group called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. K-means is a typical unsupervised learning clustering algorithm. It partitions a set of data into k clusters. However, it assumes that k is known in advance. Following is the summary of the algorithm:

1. Put K points into the representation of space by the objects that are being clustered. These points represent group Centroids.
2. Allocated each object to the group that has the closest centroids
3. When all objects have been allocated, again calculation of the positions of the K centroids.
4. Repeating of Steps 2 and 3 up to the centroids no longer move. This produces a separation of the objects into number of groups from which the metric to be minimized can be calculated.

The papering flow of "clustering using k-means on selected data" is explained by flowchart shown in figure 3. Start the procedure, next accept starting period, ending period data from user, and accept k value from user. Accept clustering data option,

4) Average Texture
5) Flora
6) Average Density

Cluster Arial data using k-means algorithm on selected feature (between starting and ending period days). If select 4 as our option than average Texture data is cluster, if select 5 than rain, if select 6 than average density data is cluster using k-means clustering algorithm. Display final results and Stop the procedure.

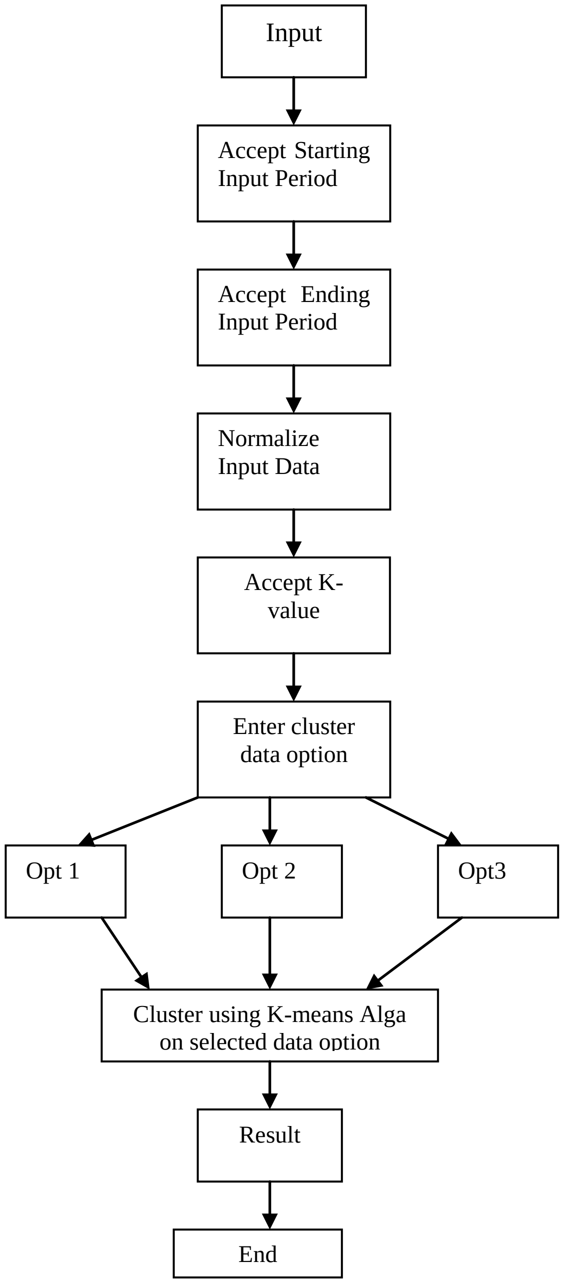
Input
Accept Starting Input Period
Accept Ending Input Period
Normalize Input Data
Accept K-value
Enter cluster data option
Opt 1
Opt 2
Opt3
Cluster using K-means Alga on selected data option
Result
End

Figure 3 : Clustering Arial Data using K-means Algorithm

Apply PCA algorithm on selected feature. The papering flow of "PCA on selected data" is explained by flowchart shown in figure 4.

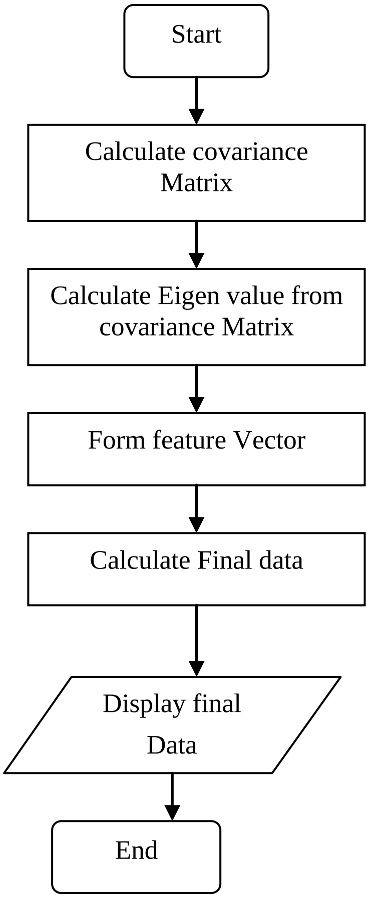
Start
Calculate covariance Matrix
Calculate Eigen value from covariance Matrix
Form feature Vector
Calculate Final data
Display final Data
End

Figure 4 : PCA method on Arial data

Start the procedure next accept starting period, ending period data from user, and accept k value from user. Accept PCA data option, Perform PCA Method on selected data option. Subtract mean from normalize option data store in adjusted data variable. Calculate adjusted data covariance matrix. Calculate Eigen vector, value from covariance matrix, next select eigenvector with highest Eigen value is feature vector. Next calculate final data, display final data and stop the procedure.

IV. Result Analysis

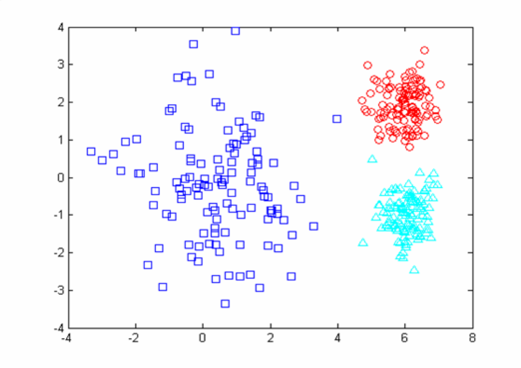

Figure 5 : distributed data set of the read information

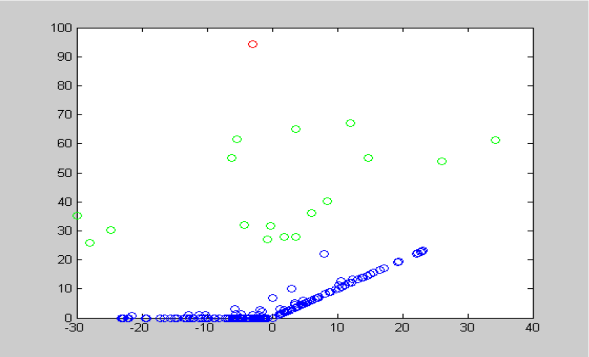

Figure 6 : the classified Component analysis data set for the reference model

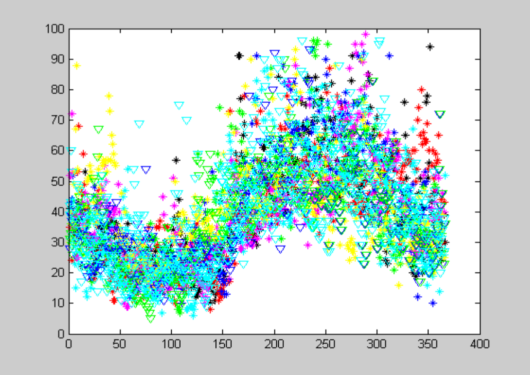

Figure 7 : obtained distribution reference for the given input data

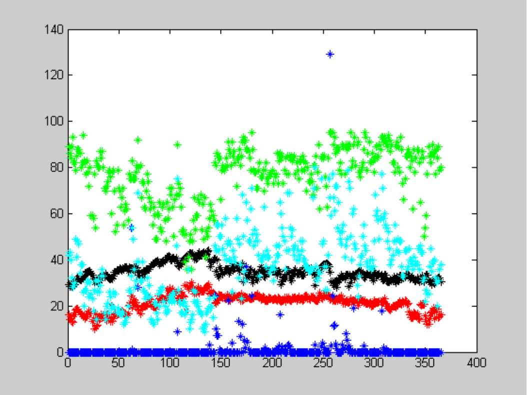

Figure 8 : Observations for obtained classified data set for the input information

V. Conclusion

The Arial data contains date periodically observed with parameters of texture (min, max), flora, and density (min, max).Farmer needs timely and accurate Arial data. In order to achieve this, data should be continuously recorded from stations that are properly identified, manned by trained staff or automated with regular maintenance, in good papering order and secure from tampering. The stations should also have a long history and not be prone to relocation. The collection and archiving of Arial data is important because it provides an economic benefit but the local/national economic needs are not as dependent on high data quality as is the Arial risk market. In this study, it was found that the data mining tools could enable experts to predict Arial with satisfying accuracy using as input the Arial parameters of the previous years. The K-means clustering and PCA algorithms are suggested and tested for period of 11 years with multiple features to early prediction of Arial for agriculture applications.

References References Referencias

1. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Prentice-Hall, New Jersey, 1988.
2. R. A. Fisher. The Use of Multiple Measurements on Taxonomic Problems, Annals of Eugenics, vol. 7, 179-188, 1936.
3. David J. Hand, Heikki Mannila and Padhraic Smyth. Principles of Data Mining, MIT Press, 2001.
4. R. W. Cooley. Web usage mining: discovery and application of interesting patterns from web data. PhD thesis, University of Minnesota, USA, 2000.
5. M. R. Anderberg, Cluster analysis for applications, Academic Press, Inc., London, 973.
6. K. D. Bailey, Cluster analysis, Sociological Methodology, (1975), 59–128.
7. R. J. Little and D. B. Rubin, Statistical analysis with missing data,John Wiley & Sons, 1987.
8. Guha. S, Rastogi. R, and Shim, 1998. Cure: An efficient clustering algorithm for large databases. In Proceedings of the ACM Sigmod Conference, 73-84, Seattle, WA.
9. Berry. M. W and Browne. M, 1999, Understanding Search Engines: Mathematical Modeling and Text Retrieval. SIAM.
10. P. Bradley, C. Reina, and U. Fayyad, Clustering very large databases using EM mixture models, in Proceedings of 15th International Conference on Pattern Recognition (ICPR'00), vol. 2, 2000, 76–80.
11. J. Macqueen, Some methods for classification and analysis of multivariate observations, in Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, 281-297.Lindsay I Smith, "A tutorial on Principal Components Analysis", February 26, 2002.
12. M.C. Todd, C.Kidd, D. Kniveton and T. J. Bellerby, A combined satellite infrared and passive microwave

technique for estimation of small-scale flora, J.Atmos, Oceanic technol., vol. 18, 724-75, 2001.

13. R.F. Adler and A.J. Negri, A satellite technique to estimate tropical convective and stratiform flora, J. Appl. Meteorol., vol. 27, 30-51, 1988.
14. D.I.F Grimes, E. Coppola, M.Verdecchia, and G. Visconti, A neural netpaper approach to real-time flora estimation for Africa using satellite data, J. Hydrometeorol., vol. 4, 1119-1133, 2003.
15. Degaetano, A spatial grouping of United States climate stations using a hybrid clustering approach. International Journal of Climatology, Chichester, Vol.21, 791-807, 2001.