iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VIII **Month of publication:** August 2017

DOI: http://doi.org/10.22214/ijraset.2017.8151

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

A Review on Hybrid Approach Based on K-Mean and Ward's Algorithm

Sophiya[1], Saurabh Sharma[2]
[1,2]*Department of Computer Science Engineering, Sri Sai University*

Abstract*: *Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. In this paper, a study based on K-means and Ward's Algorithm with Honey Bee optimization is done for spatial data mining and finally an algorithm is created for data clustering also. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining. So, by this algorithm clustering can be done in a most appropriate way and can be used for further study.

Keywords: Data clustering, Data Mining, Algorithm, Honey Bee Optimization

I. INTRODUCTION

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. It is basically holds the complexity of spatial data types, spatial relationship and spatial autocorrelation. Spatial data are the data related to objects that occupy space. A spatial database stores spatial objects represented by spatial data types and spatial relationship among such objects [1] [2]. Moreover Geographical information systems are becoming rich deposits of spatial data in many application areas (i.e., geology, meteorology, traffic planning, emergency aids). The GISs provide the user with the possibility of querying a territory for extracting areas that exhibit certain properties, i.e., given combinations of values of the attributes. This explosively growing spatial data creates the necessity of knowledge/information discovery from spatial data, which leads to a promising emerging field, called spatial data mining or knowledge discovery in spatial databases [3].

Regionalization has been an important and challenging problem for a large spectrum of research and application domains, for example, climatic zoning [5], eco region analysis [6], hazards and disasters management [7], map generalization [8], location optimization [9], census reengineering [10], and health-related analysis [11]. Regionalization is essentially a special form of classification where spatial units are grouped together, based on a set of defined criteria and a set of contiguity or adjacency constraints [12]. Spatial clustering is an important component of spatial data mining. It aims group similar spatial objects into group or clusters so that objects within a cluster have high similarity in comparison to one another but are dissimilar to objects in other clusters [13]. Spatial clustering can be applicable for solving many problems. An important application area for the spatial clustering algorithm is social and economic geography. In the scope a classical methodical problem of social geography, "regionalization" can be considered [13].In this paper, we propose a new hybrid approach for data clustering making use of FCM and gravitational search optimization

II. PROPOSED WORK

RCDA [7] was proposed to enhance the quality of clustering by combining multiple clustering schemes to produce a more robust scheme delivering similar homogeneous basins. Cluster analysis is been used to identify, analyses and describe hydrological similar catchments. Cluster ensemble techniques aims to improve the clustering scheme by aptly combining multiple schemes. Computational validation of results is performed by analyzing the SSE of the clustering scheme obtained by RCDA, with another cluster ensemble method available in R software and the scheme with the lowest SSE is taken as optimum scheme. Spatial Clustering algorithm proposed in [2] for efficient processing of objects with neighborhood relations. Therefore, spatial clustering is determined by its spatial attributes as well as the attributes of objects in its neighborhood. Cluster with shortest distance based geomorphological discrepancy laws are combined. The drawback of this method is that regional homogeneity is not guaranteed.

Fuzzy based Cluster analysis [11] was used for partial distribution of the sites to a region. A catchment is classified as belonging

to a group on the basis of its dissimilarity with other catchments in the region in a multi-dimensional space of attributes affecting their flood response. It is not legitimize to assign catchment to one group or another, when it resembles more than one catchment. The fuzzy clustering algorithm (FCA) allows a catchment to have partial or distributed memberships in all the regions (groups) identified.

Sharma et al [12] proposed efficient clustering technique for regionalization of a spatial database (RCSDB). This algorithm combines the 'spatial density' and a covariance based method to inductively find spatially dense and non-spatially homogeneous clusters of arbitrary shape. RCSDB takes into account spatial point distributions as well as the distribution of several non-spatial characteristics. RCSDB classify a database of geographical locations into homogeneous, planar and density-connected subsets called regions. It finds internally density connected sets.

A Comparative study [10] of the regionalization used in spatial data mining techniques is reported. Regionalization techniques can be divided into four parts: Conventional clustering method, maximization of regional compactness approach, an explicit spatial contiguity constraint approach, and density based approach.

III. CLUSTRING ALGORITHUM

The proposed model analyses three cross clustering algorithms which are blend of centroid based K-Means and three agglomerative algorithms such as single linkage, Ward's and DBSCAN for regionalization and brings out the uncertainties faced

A. Architecture

The analysis of spatial temporal variation of multiple pollutants can be actualized by not only recording the type and the value of the pollutants such as CO_2, NO_2 and CO but also tracking the location of the sensors during each measurement time.

1) *Data Cleansing:* The data collected from sensors is first cleansed. The database contains huge volume of data, which are to be normalized to equalize the size or magnitude and the variability of these features. This can also be seen as a way to adjust the relative weighting of the attributes. A uniform view of the data in this project is obtained by z-score normalization using Eq(1)

$$z = \frac{x - \mu}{\sigma} \qquad (1)$$

2) *Reverse Geocoding and Spatial Aggregation:* The data obtained from the sensors contain the local pollution information status including the co-ordinates, timestamp, vehicle id, CO_2 NO_2 and CO. Geo coding is the process of converting the place name into GPS coordinates whereas reverse geocoding is the coding of a point location to a readable address. The GPS coordinate of the vehicle is interpolated to the nearest address by integrating the data from the sensors with the geographical information system (GIS) of Chennai database. The pollution status of the city can be in a snapshot by aggregating the data points in spatial and temporal domains

B. Regionalisation using Hard Clustering Algorithms

Hard Clustering can be widely classified into two categories: Hierarchical and Partitioning cluster algorithm. Hierarchical clustering looks for grouping of clusters in a hierarchy either agglomerative (bottom up) or divisive (top down) approach. Partitioning algorithm partitions the entire data into distinct partitions with an initial assumption of the number of clusters and cluster centroids.

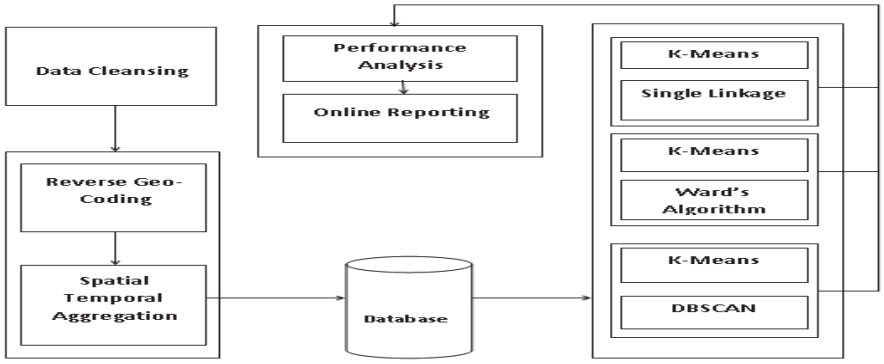


Fig.1. Functional Architecture for Regionalisation of Spatial temporal Dataset

Three cross clustering algorithms, which are a blend of connectivity based agglomerative algorithm and centroid based partitioning algorithm, are examined in this analysis. The centroid based K-Means algorithm is used to partition the feature space into K distinct clusters using the gas attributes such as CO, CO_2 and NO_2. In the second step, objects in the same cluster with no spatial adjacency or connectivity will be split, forming different regions [10]. This solution enables a quick evaluation of spatial dependence among objects. Regionalization of data with high degree of homogeneity dynamically helps in finding the contribution of various sources to the pollution levels. The aim of this module is to carry out a qualitative comparison of three different methods namely Single linkage, ward's algorithm and DBSCAN used in the context of the regionalization problem.

C. K- Means with Single Linkage Algorithm

Single linkage, nearest neighbour or shortest distance is a method of calculating distances between clusters in hierarchical clustering the distance D(X,Y) between clusters X and Y is described by the expression,

$$D(X, Y) = \min d(x, y).$$
$$x \in X, y \in Y \qquad (2)$$

1) *Step 1*: Set the number of clusters in K –Means. Apply k- means algorithm to generate k – homogenous clusters with similar gas characteristics
2) *Step 2*: For each of the K- clusters. Plot the objects in *n*- dimensional space (where *n* is the number of attributes)
3) *Step 3:* Construct the initial dissimilarity matrix SJij and contiguity matrix
4) *Step 4:* Locate the minimum value of SJij in the dissimilarity matrix that is contiguous and merge clusters (i) and (j) to a single cluster
5) *Step 5:* Delete the rows i and j from the dissimilarity matrix
6) *Step 6:* Update the dissimilarity matrix by adding a row for the newly merged cluster
7) *Step 7:* Compute the distance between the merged cluster (i,j) and other cluster(k)

$$d[(k),(i,j)] = \min\left(d[(k),(i)], d[(k),(j)]\right) \qquad (3)$$

8) *Step 8:* Repeat the steps from step2 until there are no more contiguous pairs of clusters remaining.

D. K- Means with Ward's Algorithm

Ward's minimum variance criterion minimizes the total within-cluster variance. In ward's algorithm, at each step the pair of clusters with minimum cluster distance is merged. This is calculated using,

$$d_{ij} = d(\{X_i\}, \{X_j\}) = \| X_i - X_j \|^2 \qquad (4)$$

1) *Step 1:* Set the number of clusters in K –Means. Apply k- means algorithm to generate k – homogenous clusters with similar gas characteristics
2) *Step 2:* For each of the K- clusters Plot the objects in *n*- dimensional space (where *n* is the number of attributes).
3) *Step 3:* Construct the initial dissimilarity matrix SJij and contiguity matrix.
4) *Step 4:* Locate the minimum value of SJij in the dissimilarity matrix that is contiguous and merge clusters (i) and (j) to a single cluster
5) *Step 5:* Delete the rows i and j from the dissimilarity matrix
6) *Step 6:* Update the dissimilarity matrix by adding a row for the newly merged cluster
7) *Step 7:* Compute the distance between the merged cluster(i,j) and other cluster(k).
8) *Step 8:* Repeat the steps from step2 until there are no more contiguous pairs of clusters remaining.

$$d(C_i \cup C_j, C_k) = \frac{(n_i + n_j)}{(n_i + n_j + n_k)} d(C_i, C_j) + \frac{(n_j + n_k)}{(n_i + n_j + n_k)} d(C_j, C_k) - \frac{n_k}{(n_i + n_j + n_k)} d(C_i, C_j) \qquad (5)$$

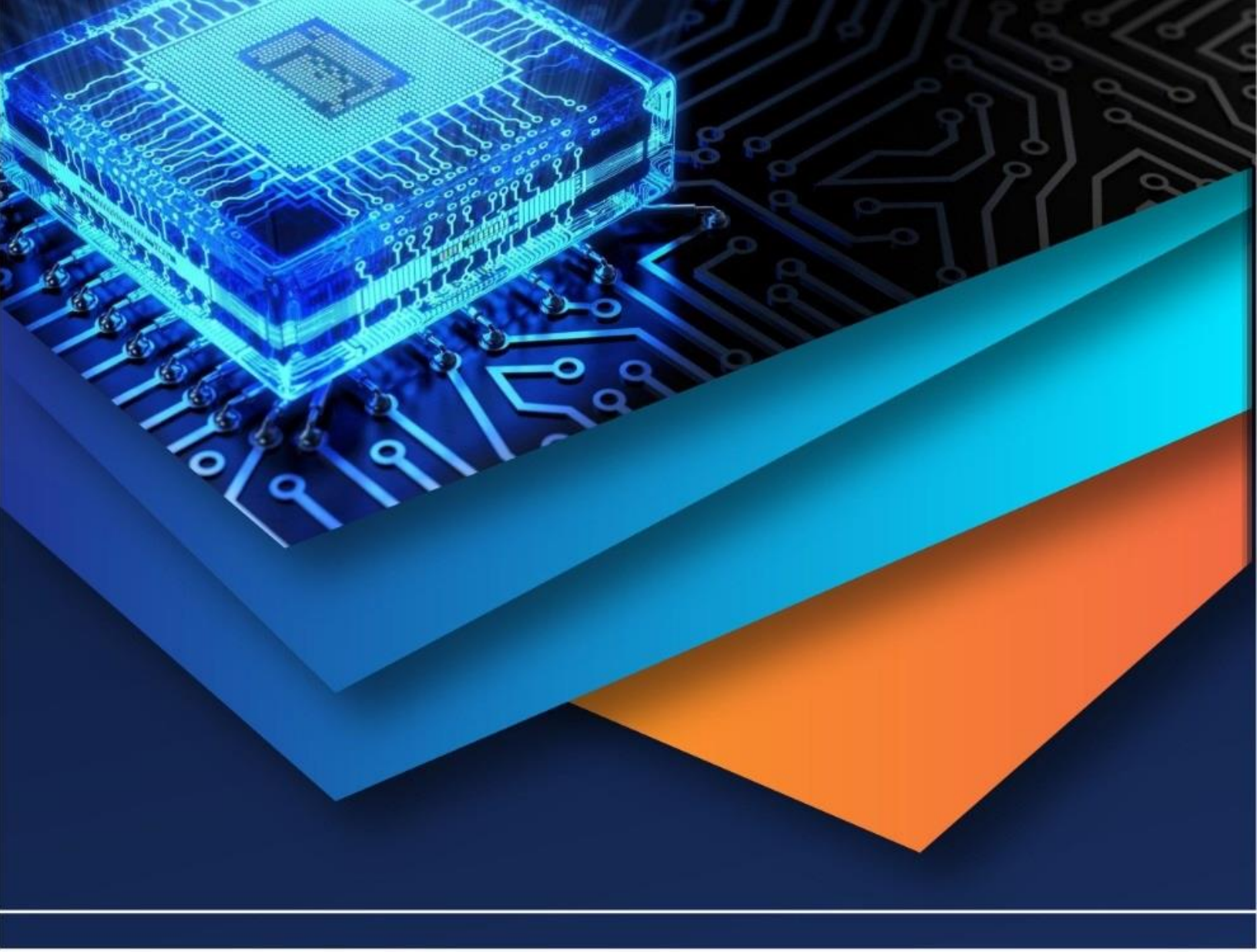
IV. CONCLUSIONS

The base paper, which we have taken into consideration, is to develop a system that applies data mining techniques to study air quality distribution using vehicular networking and map the distribution to geographic locations for effective policymaking in

Chennai. In their work, they used different hybrid-cluster methods for grouping sites into non-overlapping, contiguous and homogeneous regions and proof that ward's algorithm gives best results as compare to other algorithm. Hence, in our work, we have enhanced this work by using hybrid algorithm based on k Means clustering and ward's clustering algorithm with the help of HBO technique. The findings are more efficient and less time consuming.

REFERENCES

[1] J.Christina, Dr.K.Komathy, "Analysis of Hard Clustering Algorithms Applicable to Regionalization" Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).
[2] Dipika kalyani, Prof. Setu Kumar Chaturvedi, "A Survey on Spatio-Temporal Data Mining" International Journal of Computer Science and Network (IJCSN), Volume 1, Issue 4, August 2012
[3] G.Kiran Kumar, P.Premchand, T.Venu Gopal, " Mining Of Spatial Co-location Pattern from Spatial Datasets" International Journal of Computer Applications (0975 – 8887) Volume 42– No.21, March 2012.
[4] Sheng-Tun Li and Shih-Wei Chou, Jeng-Jong Pan "Multi-Resolution Spatio-temporal Data Mining for the Study of Air Pollutant Regionalizationl" Proceedings of the 33rd Hawaii International Conference on System Sciences – 2000.
[5] Diansheng Guo a,1, Jeremy Mennis, "Spatial data mining and geographic knowledge discovery—An introduction" Computers, Environment and Urban Systems 33 (2009) 403–408. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
[6] In-So0 Kang, Tae-wan Kim, and Ki-Joune Li, "A Spatial Data Mining Method by Delaunay Triangulation" http://cimic.rutgers.edu/~adam/mmis03/MMIS/spatial_dm.pdf.
[7] Rongqin Lan a, *, Wenzhong Shi b, Xiaomei Yang c, Guangyuan Lin, "Mining Fuzzy Spatial Configuration Rules: Methods and Applications" ISPRS Workshop on Service and Application of Spatial Data Infrastructure, XXXVI (4/W6), Oct.14-16, Hangzhou, China.
[8] http://ieeexplore.ieee.org/document/7823259/?part=1
[9] O. A. Mohamed Jafar, R. Sivakumar, "A Comparative Study of Hard and Fuzzy Data Clustering Algorithms with Cluster Validity Indices" http://icecit.sit.ac.in/images/Single%20column%20-%20Sample-elsevier.pdf



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY