Abstract

One of new challenges of our daily activity, is how to deal with uprising amount of unclassified and erratic data, to search through high volume of data, in order to point to specific category of data. Traditional statistical approach for analyzing, organizing and identifying categories among large volume of data has computational barriers and is a time-consuming task. Modern technology and solutions are applied on spatial data with the purpose of extracting knowledge autonomously. Advanced data mining techniques are potential candidates to reduce risk of being involved with dizzy, gloaming unclassified information which are proliferated increasingly. One of the unsupervised techniques to classification problem is clustering mechanism leveraging many aspect of big-data analysis. this paper presents a new approach of clustering technique with definition of new hybrid model outperform classic methods of clustering such as partitioned or hierarchical algorithms. It uses the systematic cooperation of two popular clustering algorithms: the AGlomerative NEStive (AGNES), as a hierarchical clustering method and κ-means, as a partitional clustering method. It inherits the aspects of two approaches: *low time complexity and closest relative neighbors’* mechanism in k-means and AGNES algorithms, respectively. The proposed method expects to be faster and more accurate than two classic methods. The paper evaluate the result using several popular evaluation criteria indicated in the evaluation section. The result reveals that the proposed algorithm performs faster with a higher quality of clusters regarding to inherited merits from parents (k-means, AGNES).

1. **Introduction**

With subtle growth in tendency among researchers to carry out data mining techniques on different type of data (Medicines, Geographical or Weather-related data), knowledge discovery is considered to be needed more frequently ever before. KDD[[1]](#footnote-1) is the higher level process of obtaining facts through data mining and distilling this information into knowledge or ideas and beliefs about the mini-world described by the data. This generally requires a human-level intelligence to guide the process and interpret the results based on pre-existing knowledge[1]. Exiting methods for exploratory spatial analysis and spatial datamining span across three main categories: computational, statistical, and visual approaches[2]. The paper subjects mainly on first category of spatial analysis. Computational approaches which resort to computer algorithms to search large volume of data in order to find specific type of patterns such as spatial clusters[3], spatial association rules[4] and spatial outliers[5].

In general, computational methods are able to search for structures in large datasets with great efficiency but lack the ability to interpret and attach meaning to patterns[2].Statistical methods are rigorous and verifiable but often assume a priorimodel which has been roughly predetermined by the analyzer[2]. Visualization techniques in spatial analysis is………. (source should indicates the topic)

This paper organized around topics to propose a solution in computational aspect of spatial analysis. Computational methods are able to search for structures in large datasets with great efficiency but lack the ability to interpret and attach meaning to patterns[2]. Presenting an effective method in order to cluster spatial data which are gathered from diverse sources is a challenging task. Employing clustering methods in order to discover related data in spatial analysis is a main purpose of our study. The proposed method utilizes a systematic hybrid approach by combining AGNES as a hierarchical and K-means as a partitional clustering algorithms. The paper will cover both the new proposed algorithm, and its function in crime incidents’ location data as spatial analysis. The case study assess clustering approach in different type of dataset…. (Crimes location data, weather data, or diabetes data) … Eventually, the method has been tested and evaluated through using different types of data (spatial data). Different types of data are selected in order to assess the accuracy and quality of the new proposed method.

The paper is consisted of several major sections. Section 2 is dedicated to discuss on background theories, related works and also recent researches around this subject. Next section (3) explains about the proposed hybrid clustering methods named HAK, its overall procedure and the parameters which effects the algorithm to output better results. Section 4,a discusses about clustering methods applied on spatial data (to classify them without supervise), in this section other types of dataset are utilized in order to justify the accuracy, validity and commensurate computational cost of our proposed algorithm. In section 6, some of the most popular evaluation criteria (Fisher’s separability criterion and minimum Total Distance), Davis Bolden, Silhouette coefficient, and other criteria are introduced, after which the proposed hybrid technique is evaluated on the basis of those criteria. The last section concludes our work and explains about possibilities and ideas to be done later.

1. **Literature Review**

**2.1) Clustering**

Cluster analysis or clustering is a task of assigning a set of objects into groups (called clusters) so that the object in the same cluster are more similar to each other base on selected features than to those on other clusters [6]. Comparison between different clustering methods, have been performed in order to present the quality, accuracy and efficiency of one algorithm comparing with other methods.

**2.2) Clustering Major Categories**

The major categories of clustering are partitioned clustering, hierarchical clustering, density based clustering and … we focus on the two most popular categories of clustering because each of them has some specific features. In partitioning clustering method, the partitions of cluster is selected and each member is assigned to a partition with lowest distance metrics. Example of partitioning algorithms are k-means, clustering (MacQueen 1967), k-medoid clustering, genetic k-means algorithm (GKA), Self-Organizing Map (SOM) and also graph-theoretical methods (CLICK, CAST).

One of the rigorous advantage of k-mean or other partitioned clustering methods is its lowest time complexity that is inherited from its preselected candidates (mean, mode or centroids) natures. Therefore, algorithms of this category are fast and are comfortable for high dimensional and large scale datasets due to low time cost allocated to clustering. On the other hand, hierarchical clustering, are more precise and clustered generated from this category are well-clustered. However, due to computation cost of calculating distances between members in different hierarchies. In this paper, we focus on two naïve algorithm of k-mean and accumulative hierarchical clustering, which merits main characters inherited from both approaches.

**2.3) Limitation of Traditional clustering methods**

According to the paper [7], k-means clustering requires a specified number of clusters in advance. It needs initial centroids to be selected, randomly. K-means is also sensitive to outliers. Randomly selecting initial start point might affects the quality of output clusters. Therefore, much iteration must be performed of the entire clustering process in order to identify best fitted-clusters [reference] (Shin et al, in preparation). On the other hand, hierarchical clustering cannot well-cluster data with similar pattern. When the size of clusters becomes larger, the cluster actual expression patterns become less relevant. Hierarchical clustering uses dendrogram that provides an easy understanding of the data but it decrease the quality of clusters as more quantity of data is increased.

Partition-based clustering techniques such as K-Means [7] and Clarans[8] attempt to break a data set into K clusters such that the partition optimizes a given criterion. These algorithms assume that clusters are hyper-ellipsoidal and of similar sizes. They can’t find clusters that vary in size, as shown in Figure A1, or concave shapes, as shown in Figure A2.



DBScan7 (Density-Based Spatial Clustering of Applications with Noise), a wellknown spatial clustering algorithm, can find clusters of arbitrary shapes. DBScan defines a cluster to be a maximum set of density-connected points, which means that every core point in a cluster must have at least a minimum number of points (MinPts) within a given radius (Eps). DBScan assumes that all points within genuine clusters can be reached from one another by traversing a path of density connected points and points across different clusters cannot. DBScan can find arbitrarily shaped clusters if the cluster density can be determined beforehand and the cluster density is uniform.

Hierarchical clustering algorithms produce a nested sequence of clusters with a single, all-inclusive cluster at the top and single-point clusters at the bottom. Agglomerative hierarchical algorithms[7] start with each data point as a separate cluster. Each step of the algorithm involves merging two clusters that are the most similar. After each merger, the total number of clusters decreases by one. Users can repeat these steps until they obtain the desired number of clusters or the distance between the two closest clusters goes above a certain threshold. The many variations of agglomerative hierarchical algorithms[7] primarily differ in how they update the similarity between existing and merged clusters. These schemes fail for data in which points

In a given cluster are closer to the center of another cluster than to the center of their own cluster. This situation occurs in many natural clusters, 4,8 for example, if there is a large variation in cluster sizes, as in Figure A1, or when cluster shapes are concave, as in Figure A2.

The single-link hierarchical method measures the similarity between two clusters by the similarity of the closest pair of data points belonging to different clusters. Unlike the centroid/ medoid-based methods, this method can find clusters of arbitrary shape and different sizes. However, it is highly susceptible to noise, outliers, and artifacts.

Researchers have proposed CURE[9] (Clustering Using Representatives) to remedy the drawbacks of both of these methods while combining their advantages. In CURE, a cluster is represented by selecting a constant number of well-scattered points and shrinking them toward the cluster’s centroid, according to a shrinking factor. CURE measures the similarity between two clusters by the similarity of the closest pair of points belonging to different clusters. Unlike centroid/ medoid-based methods, CURE can find clusters of arbitrary shapes and sizes, as it represents each cluster via multiple representative points. Shrinking the representative points toward the centroid allows CURE to avoid some of the problems associated with noise and outliers. However, these techniques fail to account for special characteristics of individual clusters. They can make incorrect merging decisions when the underlying data does not follow the assumed model or when noise is present.

Rock[10] (Robust Clustering Using Links), a recently developed algorithm that operates on a derived similarity graph, scales the **aggregate interconnectivity** with respect to a user-specified interconnectivity model.

In some algorithms, the similarity between two clusters is captured by the aggregate of the similarities (that is, the interconnectivity) among pairs of items belonging to different clusters. The rationale for this approach is that sub-clusters belonging to the same cluster will tend to have high interconnectivity. But the aggregate interconnectivity between two clusters depends on the size of the clusters; in general, pairs of larger clusters will have higher interconnectivity. Many such schemes normalize the aggregate similarity between a pair of clusters with respect to the expected interconnectivity of the clusters involved. For example, the widely used group-average method2 assumes fully connected clusters, and thus scales the aggregate similarity between two clusters by n ´m, where n and m are the number of members in the two clusters.

However, the major limitation of all such schemes is that they assume a static, user-supplied interconnectivity model. Such models are inflexible and can easily lead to incorrect merging decisions when the model under- or overestimates the interconnectivity of the data set or when different clusters exhibit different interconnectivity characteristics. Although some schemes allow the connectivity to vary for different problem domains (as does Rock[10]), it is still the same for all clusters irrespective of their densities and shapes.

**2.4) Recent Approaches on Clustering Improvement**

In 1999, a new method of hierarchical clustering using dynamic modeling have been proposed called Chameleon[11] . Chameleon is a new agglomerative hierarchical clustering algorithm. Chameleon uses an approach to model the degree of interconnectivity and closeness between each pair of clusters. Chameleon finds the clusters in the data set by using a two-phase algorithm. During the first phase, Chameleon uses a graph-partitioning algorithm to cluster the data items into several relatively small sub-clusters. During the second phase, it uses an algorithm to find the genuine clusters by repeatedly combining these sub-clusters.

Chameleon’s sparse-graph representation of the items is based on the commonly used k-nearest-neighbor graph approach. Each vertex of the k-nearest-neighbor graph represents a data item. An edge exists between two vertices v and u if u is among the k most similar points of v, or v is among the k most similar points of u.



Figure 1) Chameleon Algorithm

**2.3) K-Means Initialization Issue:**

**K-Means is extremely sensitive to cluster center initialization**

[1] H. J. Miller, “Geographic data mining and knowledge discovery,” *Handb. Geogr. Inf. Sci.*, pp. 352–366, 2008.

[2] D. Guo, “12 Multivariate Spatial Clustering and Geovisualization,” *Geogr. Data Min. Knowl. Discov.*, p. 325, 2009.

[3] H. Miller and J. Han, “Spatial clustering methods in data mining: a survey,” *Geogr. data Min. Knowl. Discov. Taylor Fr.*, 2001.

[4] J. Han, K. Koperski, and N. Stefanovic, “GeoMiner: a system prototype for spatial data mining,” in *ACM SIGMOD Record*, 1997, vol. 26, no. 2, pp. 553–556.

[5] S. Shekhar, C.-T. Lu, and P. Zhang, “A unified approach to detecting spatial outliers,” *Geoinformatica*, vol. 7, no. 2, pp. 139–166, 2003.

[6] M. Kaushik and B. Mathur, “Comparative Study of K-Means and Hierarchical Clustering Techniques,” pp. 93–98, 2014.

[7] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data.pdf*, vol. 355. 1988.

[8] R. T. Ng and J. Han, “Efficient and Effective Clustering Methods for Spatial Data Mining,” *Proc. 20th Int. Conf. Very Large Data Bases*, pp. 144–155, 1994.

[9] S. Guha, R. Rastogi, and K. Shim, “CURE: An efficient clustering algorithm for large databases,” *Inf. Syst.*, vol. 26, no. 1, pp. 35–58, 2001.

[10] S. Guha, R. Rastogi, and K. Shim, “Rock: a robust clustering algorithm for categorical attributes,” *Inf. Syst.*, vol. 25, no. 5, pp. 345–366, 2000.

[11] G. Karypis, E.-H. Han, and V. Kumar, “Chameleon: hierarchical clustering using dynamic modeling,” *Computer (Long. Beach. Calif).*, vol. 32, no. 8, pp. 68–75, 1999.

1. Knowledge Discovery of Data [↑](#footnote-ref-1)