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 clustering method.

* 1. Clustering From (Survey of Clustering Data mining Techniques)

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.

There is a close relationship between clustering techniques and many other disciplines. Clustering has always been used in statistics [Arabie & Hubert 1996] and science [Massart & Kaufman 1983]. The classic introduction into pattern recognition framework is given in [Duda & Hart 1973]. Typical applications include speech and character recognition. Machine learning clustering algorithms were applied to image segmentation and computer vision [Jain & Flynn 1996]. For statistical approaches to pattern recognition see [Dempster et al. 1977] and [Fukunaga 1990]. Clustering can be viewed as a density estimation problem. This is the subject of traditional multivariate statistical estimation [Scott 1992]. Clustering is also widely used for data compression in image processing, which is also known as vector quantization [Gersho & Gray 1992]. Data 3 fitting in numerical analysis provides still another venue in data modeling [Daniel & Wood 1980].

Categorization of clustering algorithms is neither straightforward, nor canonical. In reality, groups below overlap. For readers convenience we provide a classification closely followed by this survey. Corresponding terms are explained below.

1. Hierarchical Methods

A1. Agglomerative Algorithms

A2. Divisive Algorithms

1. Partitioning Methods

B1. Relocation Algorithms

B2. Probabilistic Clustering

B3. K-medoids Methods

B4. K-means Methods

1. Density-Based Algorithms

C1. Density-Based Connectivity Clustering

C2. Density Functions Clustering

1. Grid-Based Methods
2. Methods Based on Co-Occurrence of Categorical Data
3. Constraint-Based Clustering
4. Clustering Algorithms Used in Machine Learning
5. Gradient Descent and Artificial Neural Networks

H1. Evolutionary Methods

1. Scalable Clustering Algorithms

I1. Algorithms for High Dimensional Data

I2. Subspace Clustering

I3. Projection Techniques

I4. Co-Clustering Techniques

* 1. **K-means**

According to [7], K-mean is an algorithm of type Partitioning Relocation Clustering. The k-means algorithm [Hartigan 1975; Hartigan & Wong 1979] is by far the most popular clustering tool used in scientific and industrial applications. The name comes from representing each of k clusters C by the mean (or weighted average) c of its points, the so-called centroid. While this obviously does not work well with categorical attributes, it has the good geometric and statistical sense for numerical attributes. The sum of discrepancies between a point and its centroid expressed through appropriate distance is used as the objective function. For example, the -norm based objective function, the sum of the squares of errors between the points and the corresponding centroids, is equal to the total intra-cluster variance.

Two versions of k-means iterative optimization are known. The first version is similar to EM algorithm and consists of two-step major iterations that (1) reassign all the points to their nearest centroids, and (2) recomputed centroids of newly assembled groups. Iterations continue until a stopping criterion is achieved (for example, no reassignments Happen). This version is known as Forgy’ s algorithm [Forgy 1965] and has many advantages:

1. It allows straightforward parallelization [Dhillon & Modha 1999]
2. It is insensitive with respect to data ordering.

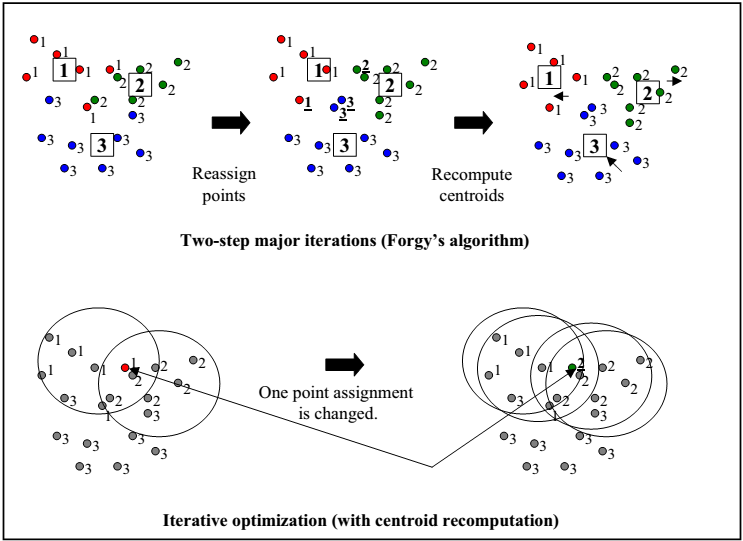
The second (classic in iterative optimization) version of k-means iterative optimization reassigns points based on more detailed analysis of effects on the objective function caused by moving a point from its current cluster to a potentially new one. If a move has a positive effect, the point is relocated and the two centroids are recomputed. It is not clear that this version is computationally feasible, because the outlined analysis requires an inner loop over all member points of involved clusters affected by centroids shifts. However, in case it is known [Duda & Hart 1973; Berkhin & Becher 2002] that all computations can be algebraically reduced to simply computing a single distance. Therefore, in this case both versions have the same computational complexity.

There is experimental evidence that compared with Forgy’s algorithm, the second (classic) version frequently yields better results [Larsen & Aone 1999; Steinbach et al.2000]

The wide popularity of k-means algorithm is well deserved. It is simple, straightforward, and is based on the firm foundation of analysis of variances. The k-means algorithm also suffers from all the usual suspects:

The result strongly depends on the initial guess of centroids (or assignments)

1. Computed local optimum is known to be a far cry from the global one
2. It is not obvious what is a good k to use
3. The process is sensitive with respect to outliers
4. The algorithm lacks scalability
5. Only numerical attributes are covered
6. Resulting clusters can be unbalanced (in Forgy’s version, even empty)



* 1. **Hierarchical Clustering.**

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) [Jain & Dubes 1988; Kaufman & Rousseeuw 1990]. An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved. Advantages of hierarchical clustering include:

1. Embedded flexibility regarding the level of granularity
2. Ease of handling of any forms of similarity or distance
3. Consequently, applicability to any attribute types

Disadvantages of hierarchical clustering are related to:

1. Vagueness of termination criteria
2. The fact that most hierarchical algorithms do not revisit once constructed (intermediate) clusters with the purpose of their improvement

In hierarchical clustering our regular point-by-attribute data representation is sometimes of secondary importance. Instead, hierarchical clustering frequently deals with the matrix of distances (dissimilarities) or similarities between training points. It is sometimes called connectivity matrix. Linkage metrics are constructed (see below) from elements of this matrix. The requirement of keeping such a large matrix in memory is unrealistic. To relax this limitation different devices are used to introduce into the connectivity matrix some sparsely. This can be done by omitting entries smaller than a certain threshold, by using only a certain subset of data representatives, or by keeping with each point only a certain number of its nearest neighbors. For example, nearest neighbor chains have decisive impact on memory consumption [Olson 1995]. A sparse matrix can be further used to represent intuitive concepts of closeness and connectivity.

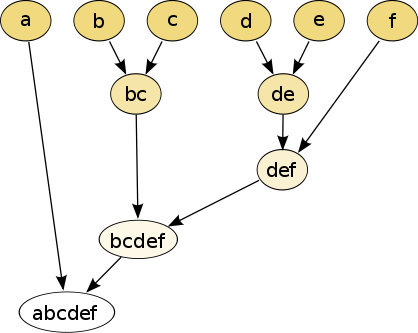
Notice that the way we process original (dis)similarity matrix and construct a linkage metric reflects our a priori ideas about the data model.

**Linkage Metrics**

Hierarchical clustering initializes a cluster system as a set of singleton clusters (agglomerative case) or a single cluster of all points (divisive case) and proceeds iteratively with merging or splitting of the most appropriate cluster(s) until the stopping criterion is achieved. The appropriateness of a cluster(s) for merging/splitting depends on the (dis)similarity of cluster(s) elements. This reflects general presumption that clusters consist of similar points. An important example of dissimilarity between two points is the distance between them.

To merge or split subsets of points rather than individual points, the distance between individual points has to be generalized to the distance between subsets. Such derived proximity measure is called a linkage metric. The type of the linkage metric used significantly affects hierarchical algorithms, since it reflects the particular concept of closeness and connectivity. Major inter-cluster linkage metrics [Murtagh 1985, Olson1995] include single link, average link, and complete link.

Linkage metrics-based hierarchical clustering suffers from time complexity. Under reasonable assumptions, such as reducibility condition (graph methods satisfy this condition), linkage metrics methods have complexity O([Olson 1995]. Despite the unfavorable time complexity, these algorithms are widely used. When the connectivity matrix is sparsified, graph methods directly dealing with the connectivity graph G can be used. In particular, hierarchical divisive MST (MinimumSpanning Tree) algorithm is based on graph partitioning [Jain & Dubes 1988].



* 1. **Comparison Between Clustering Techniques**

In this section, strength and weakness of K-means and Hierarchical Clustering are discussed. Comparison K-means and Hierarchical Clustering are also discussed.

* Strengths of K-Mean

1. Simple: - Easy to understand and to implement.
2. Efficient: Time complexity: O(tkn), where n is the number of data points, k is the number of clusters, and t is the number of iterations.
3. Since both k and t are small. K-Means is considered a linear algorithm

* Weaknesses of k-means

1. The algorithm is only applicable if the mean is defined.
2. For categorical data, k-mode - the centroid is represented by most frequent values.
3. The user needs to specify k.
4. the algorithm is sensitive to outliers: Outliers are data points that are very far away from other data points.

* Weaknesses of k-means: To deal with outliers

1. One method is to remove some data points in the clustering process that are much further away from the centroids than other data points. To be safe, we may want to monitor these possible outliers over a few iterations and then decide to remove them.
2. Another method is to perform random sampling. Since in sampling we only choose a small subset of the data points, the chance of selecting an outlier is very small. Assign the rest of the data points to the clusters by distance or similarity comparison, or classification
3. The k-means algorithm is not suitable for discovering clusters that are not hyper-ellipsoids (or hyper-spheres).

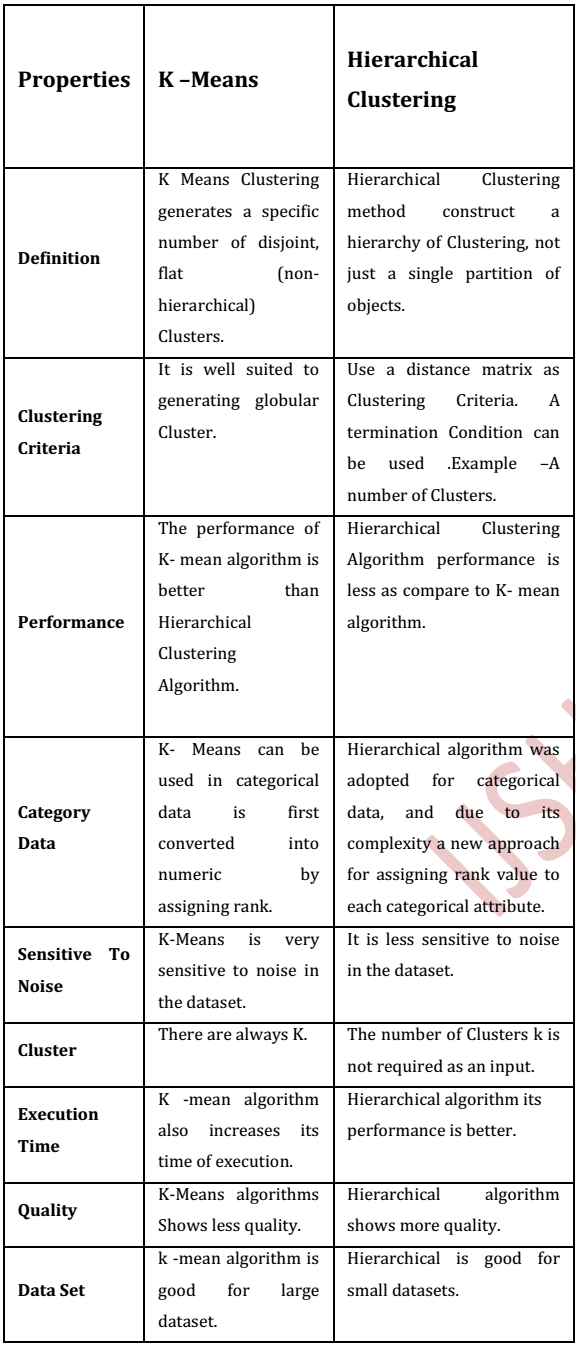
* Strengths of Hierarchical Clustering

1. Conceptually Simple.
2. Theoretical properties are well understood.
3. When Clusters are merged /split, the decision is permanent => the number of different alternatives that need to be examined is reduced.

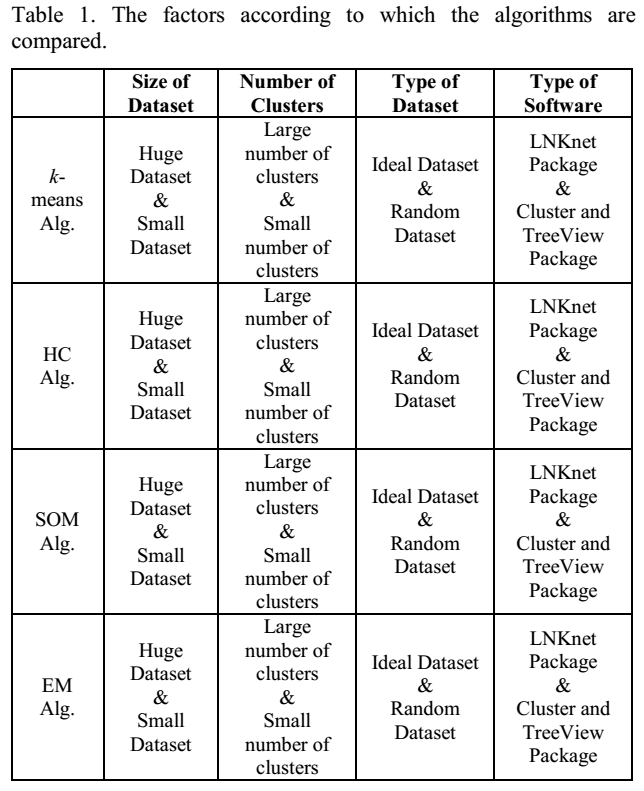
* Weakness of hierarchical Clustering

1. Merging /splitting of clusters is permanent => Erroneous decisions are impossible to correct later.
2. Divisive methods can be computational hard.
3. Methods are not (necessarily) scalable for large datasets.
4. It does not require the number of clusters k in advance.
5. It needs a termination/readout condition. The final mode in both Agglomerative and Divisive is of no use.

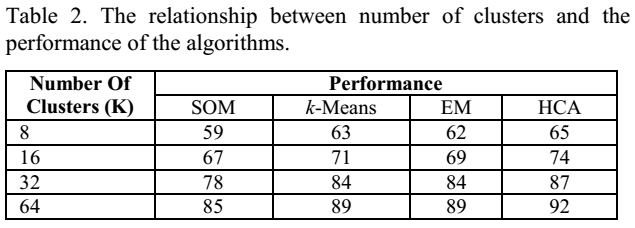
* Comparisons on K-Means and Hierarchical Clustering



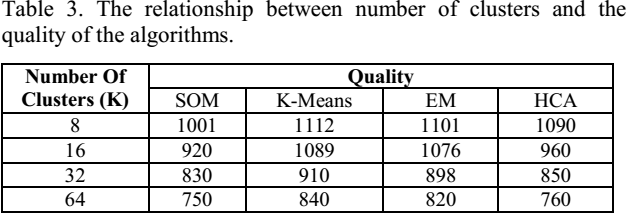
* According to the research [8], different factors were used to compare different clustering algorithms



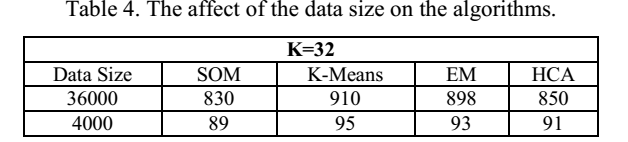
The result shows that when number of cluster increases, the performance of SOM algorithm becomes lower. The performance of k-means and E-M becomes better then hierarchical clustering algorithm.



The result of SOM based Clustering algorithm are more accurate than other algorithm in advance. However, when the number of clusters becomes greater, hierarchical clustering the accuracy becomes better than others.



The quality of K-means and EM becomes very good working with large datasets. On the other hand tow other algorithms shows more quality when size of dataset is small.



* 1. **Clustering Performance and Test Data**

In this section we briefly discusses the evaluation methods and dataset which are used in related researches to evaluate and compare different clustering algorithms.

***Test Data***

According to [8] dataset of (kdd cup) is obtained from [www.kdd.ics.uci.edu](http://www.kdd.ics.uci.edu). This is good dataset to test time series clustering algorithms because Euclidean distance will not be able to achieve perfect accuracy. Another good dataset in order to run on evaluation phase is IRIS and DIABETS dataset which are popular and also available at UCI repository of machine learning[9]. There is also a dataset containing earthquake phenomena which have occurred in Iran in the year of 2008, was selected from collection of datasets of geophysics Institute of Tehran University[10]. This dataset is elected as a type of geo-spatial category of data in order to analyze different algorithms. However, the data set contains the accurate coordinates of Iran’s earthquake events which have been collected by seismographs established across the country.

***Evaluation Metrics***

In general view, clustering evaluation criteria are categorized in three major classes. External Index, Internal Index and Relative Index. External Index is used to measure the extent to which cluster labels match externally supplied class labels. An example of this category is entropy metric. Internal Index is used to measure the goodness of clustering structure without external provided information. The last category is used to compare different clustering or clusters. External and Internal indexes can be used on this type of metrics.

Two of the most popular evaluation criteria on clustering algorithms are*: Inter-cluster distance and Intra-cluster distance*[11]. They are also known as Cohesion and Separation criteria, respectively. Two rules must be applied on above criteria that are maximization of intra cluster distance and minimization of inter cluster distance. There are also other metrics for cluster evaluation such as *Fisher’s separability criterion*, *Minimum Total Distance, Silhouette coefficient, Davies–Bouldin index, Daunn Index. The implementation of each of criterion are ignored and curious researchers might refer to references provided to each criteria.* Fisher criterion provides the ratio of inter-cluster variance to intra-cluster variance. By combining the two mentioned criteria, a more generic criterion is created, which is the simplified form of it [12].

* 1. Outlier Detection

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 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**

**2.2) Clustering Major Categories**

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.

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.

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

How clustering algorithms are implemented:

[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] P. Berkhin, “Survey of clustering data mining techniques,” *Group. Multidimens. Data Recent Adv. Clust.*, vol. 10, pp. 25–71, 2006.

[8] O. A. Abbas, “Comparisons Between Data Clustering Algorithms,” *inernational Arab J. Inf. Technol.*, vol. 5, no. 3, pp. 320 – 325, 2008.

[9] M. C. and M. P, “No Title.” UCI Repository of machine learning Databases.

[10] “earthquake dataset.” [Online]. Available: http://www.geophysics.ut.ac.ir.

[11] C. Raskutti, B. and Leckie, “An Evaluation of Criteria for Measuring the Quality of Clusters,” *Morgan Kaufmann Publ. Inc*, pp. 905–910, 1999.

[12] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *J. Intell. Inf. Syst.*, vol. 17, no. 2–3, pp. 107–145, 2001.

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