ANALYSIS OF UNSTRUCTURED DATASET USING AN ENHANCED DESCRIPTIVE MINING ALGORITHM

Supported by tetfund

A. R. Ajiboye[1]*, C. Umezuruike[2] and R. E. Peter[2]
[1]Department of Computer Science, University of Ilorin, Kwara State, Nigeria
[2]School of Computing & Information Technology, Kampala International University, Uganda
*Corresponding author: ajibabdulraheem@gmail.com

Received: August 11, 2020 **Accepted:** October 14, 2020

Abstract: The grouping of large unstructured dataset is one of the main tasks in cluster analysis. A dataset is unstructured if it has a muddle of data types whose pattern makes it uneasy to search or partition. Unstructured dataset is difficult to classify because it does not have a defined schema. An Enhanced Descriptive Mining Algorithm (EDMA) proposed in this study was used to group the given instances in the input space into a number of clusters. The aim of this study is to partition and analyse a given unstructured dataset to its constituent's distinct features. In order to achieve this central objective, the proposed EDMA is implemented along with the data dictionary created within the program to support the analysis; the implementation was carried out using java programming language. The unstructured dataset taken as input was retrieved from an open repository and comprised of numeric, alphabetic and some special characters. The resulting output of this study shows a well clustered data that is partitioned according to their similarity features. Based on a number of metrics, the performance of the proposed technique is determined by evaluating its effectiveness in relation to some existing techniques: k-means and EM clustering techniques. Findings from this study showed that, the proposed technique is reliable, accurate, and very suitable for the clustering of unstructured dataset.

Keywords: Clustering algorithm, unstructured dataset, classification, descriptive mining, data dictionary

Introduction

There is huge accumulation of data that is being captured on regular basis through several means such as: E-mails, blogs, transaction data, and billions of Web pages. These create terabytes of new data and many of these data streams are unstructured, hence, creating a new challenge in analysing them (Jain, 2010). Data of huge volume like this calls for advances in methodology to automatically understand, process, and summarize them.

Clustering of large unstructured data sets is one of the main tasks in cluster analysis. The clustering technique uses unsupervised learning approach which seeks to identify objects that can be grouped together based on their similarity features. The clustering of dataset generally involve the grouping of elements into a number of clusters, on the basis of their common traits (Rodriguez & Laio, 2014). There is no consensus definition for clustering, however, the study reported in (Napoleon & Lakshmi, 2010), describes clustering as an automated search for dataset that share similar features. Clustering has also been described as an essential data analysis and visualization tool (Xie *et al.*, 2016).

The clustering of data can be achieved using a number of algorithms such the K-means algorithm (Suh, 2012). K-means clustering is a well-known method in machine learning, where a bunch of data is explored in order to find interesting clusters of things that is based on the attributes of the data itself (Kane, 2017). Another popular clustering technique is K-medoids algorithm (Berkhin, 2006), which also clusters data by determining the mode, whereas, K-means algorithm groups data based on distance measures; both uses partitioning approach. However, the k-means is the most reported clustering algorithms for partitioning of data sets into group of objects (Daiyan *et al.*, 2012). Also, survey studies reported in (Berkhin, 2006; Fahad *et al.*, 2014) showed that, this technique is faster and the technique was described as the leading distance-based clustering technique.

Apart from the distanced based clustering techniques, other common types of clustering include those that are based on grid, density and hierarchical clustering. Each of these clustering approaches come with a number of challenges (Jain, 2010). This study therefore, proposed an enhanced clustering technique that can be used effectively in lieu of the already known and well established partitioning clustering techniques. The proposed approach implement java for better exploration of unstructured data set, the method scan through each character of the unstructured data and map them into appropriate cluster based on their similarity features. The clustering technique which is otherwise referred to as descriptive mining technique is particularly useful in several areas, most especially for classification purposes. It is an unsupervised learning technique as the grouping of data does not requires any class labels (Han *et al.*, 2012b). The implementation of this technique automatically unveils the hidden features or patterns in the datasets.

The dataset explored in this research is unstructured. A dataset is said to be unstructured if it possess an undefined schema. Typical examples of data in this category includes: data stream, multimedia, web data, text data etc. This study will however be limited to text data only. The research in text mining has been very active, and its important goal is to derive high-quality information from text (Han *et al*., 2012b). One of the drawbacks of k-means algorithm is its inability to cluster text data and it appears to be very sensitive to outliers or missing values.

The goal of this study focuses on partitioning a given unstructured data set into a number of clusters based on their similarity features. The proposed technique which is implemented in java, addresses some of the glaring weaknesses that is peculiar to some existing partitioning algorithms. The proposed approach is found to support missing values, and it clusters numeric, alphabetic and special characters.

Descriptive mining algorithms

The most imperative thing in data mining is the model creation of high accuracy that can unveil useful information from data through prediction (predictive mining) or description (descriptive mining). The predictive and descriptive tasks has been identified in (Kumar, 2014) as the two major tasks performed in data mining. The focus of this research is on descriptive mining, otherwise referred to as clustering.

Clustering is of great importance in business intelligence. It plays a central role in customer relationship management, which groups customers based on their similarities (Han *et al.*,

2012a). Clustering techniques apply when there is no class to be predicted but the instances are to be divided into natural groups. Having a clear and accurate picture of how business is growing is obviously a massive benefit for any business. Some information can help to confidently judge what should be expected in the near future for proper planning and to prepare for the worst, redistributing or scaling up resources where needed, to allocate budget effectively.

During the clustering process, there is an attempt to group the data set into observation subsets, or *clusters*, this implies that the observations should share some similarities to those in the same cluster but differ in a number of ways to those from the observations that belong to other clusters. Also, according to Han *et al*. (2012), it is a bit tasking to provide a crisp categorization of clustering methods. This is due to the fact that some overlapping may be recorded; nevertheless, it is useful to present a relatively organized picture of clustering methods. There are four major fundamental clustering methods and they can be classified into: Partitioning, Hierarchical, Density-based and Grid-based.

Hierarchical methods: Clustering using this method is characterised by hierarchical decomposition. This is a multiple levels decomposition of the given set of data objects. This method cannot correct erroneous merges or splits and may incorporate other well established techniques for better performance. Hierarchical method can be classified as being either *agglomerative* or *divisive*, this is however, based on how the hierarchical decomposition is formed.

***Density-based methods*:** This approach has the characteristics clustering dense regions of objects especially in the space that are separated by low-density regions. It is a very useful technique for the filtering of outliers. The distance-based clustering approach is capable of finding only spherical-shaped clusters. It becomes very difficult when there is need to discover clusters of arbitrary shape. Clustering approaches that are developed based on the notion of density is the most suitable to cluster such object. The basic idea used in density-based approach is to continue growing a given cluster as long as the density or the number of objects or data points in the neighbourhood exceeds some threshold.

***Grid-based methods*:** Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All the clustering operations are performed on the grid structure, otherwise referred to as quantized space. One of the main advantages of using this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. The use of grids is often an efficient approach to many spatial data mining problems, including clustering. Therefore, grid-based methods can be integrated with other clustering methods such as density-based methods and hierarchical methods. Some clustering algorithms integrate the ideas of several clustering methods;this makes it a bit difficult to classify a given algorithm as uniquely belonging to only one clustering method category.

Partitioning methods: This involves partitioning of given n objects into k groups; each partition is usually referred to as a cluster. The division of the objects in *n* is such that each group must contain at least one object. The basic partitioning methods typically adopt *exclusive cluster separation*; this implies that each object must belong to exactly one group. Most partitioning methods are distance-based and in general, the criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects in different clusters are far apart or very differs greatly. Some of the characteristics of clustering using this method include: finding the mutual exclusive of spherical shape, it is distance-based, it uses mean or medoid in most cases to represents cluster centres. The clustering of data using the concept of k-means uses the algorithm illustrated in Fig. 1.

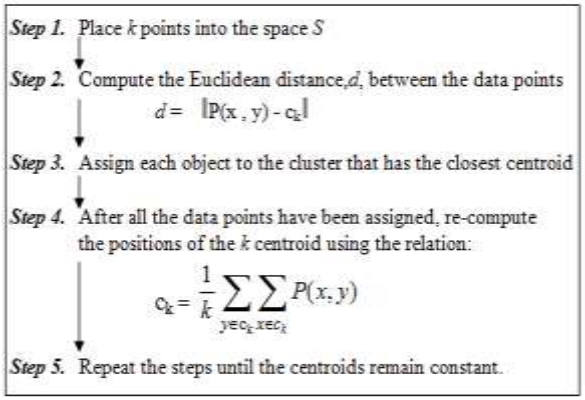

Fig. 1: The k-means algorithm, adapted from Mirkin (2012)

As shown in the algorithm, the k-means algorithm iterates between two main steps; one step basically concerns with the making updates of the clusters in conformance to the minimum distance rules, the second step update the centre of gravity of the clusters (Mirkin, 2012). As soon as the grouping is concluded, it recalculates the new centroid for each cluster formed.

A number of algorithms have been proposed in the literature for data clustering. Some of these algorithms are reviewed in this study. A deep belief network proposed in Wang *et al.* (2018) was based on k-means clustering approach. The study was aimed at predicting the wind power output; the techniques were found to be suitable for the problem since the target data was not known. This is a classification problem and several researchers have shown that this areas of study deserves attention; andin order to achieve a much better clustering of data, the study reported in Shafeeq & Hareesha (2012) saw the need to modify the existing k means algorithm with a view to achieving an improved cluster quality. Also in the study, substantial efforts were made at fixing the optimal number of cluster. Although, it is practically not feasible to fix the number of clusters that may be formed in advance. Similar modification was also done in order to ensure the performance of k-means is optimized. Also, in a related study reported in Daiyan *et al*. (2012), the modifications were found to have responsible for a faster clustering and a much better cluster quality.

Similarly, the use of k-means to improve classification of data is proposed in Ding & Li (2007). In general, the k-means algorithm is well reported for the creation of cluster models as a result of its capability of generating a cluster data in a faster way. The study achieved a better classification with the k-means as a result of combining the approach of linear discriminant analysis which in practice, adaptively selects the most discriminating subspace. The related study reported in Reynolds *et al.* (2004), digressed a bit from k-means by proposing a much closer algorithm, k-medoids. The technique clusters data using the mode concept and the technique appears to be more tolerance than the k means. The clustering method using k-medoids is widely used and reported in a number of studies proposed in the literature (Daiyan *et al*., 2012; Joshi *et al.*, 2011; Reynolds *et al*., 2004; Zadegan *et al.*, 2013).

The Expectation Maximization algorithm implemented in the study reported in Wu *et al.* (2014), is one of the well applied clustering algorithms. The study in addition to using the algorithm also developed a binning software, maxbin, which

automates the binning of assembled metagenomic scaffolds. The EM algorithm was the technique applied in Mustafa *et al.* (2011), to determine the missing values in Gaussian Bayesian network modelling for forest growth. The study proposed in Sharma *et al.* (2012), analysed and compared the performance of some clustering techniques with respect to the time it takes to create a cluster model. Although, the number of partition and the size of the data been clustered may influence the time it takes to cluster data, however, it takes some algorithms so much time to determine the similarity and dissimilarity of the data been explored.

In order to achieve the dataset of good quality, most especially while dealing with unstructured data, the use of k-means appears not to be suitable. Other related technique such as EM clustering technique also cannot meet expectations in this regards. Hence, the need to propose a technique that can deeply explore data for much better groupings. The proposed algorithm in this study, scans through the character that made up the data set in the input space. Each character is detected based on the information defined in the data dictionary; as the detected data type falls within the numeric, alphabetic or special character. This grouping shows what a true similarity features really implies, which several clustering techniques implemented on similar data fail to address, but in this study, the proposed technique takes a unique approach of scanning each character in the data set for better separation of data based on their similar features.

Material and Methods

Data collection

The data used for the implementation of the designed algorithm was retrieved from the samples of unstructured dataset from the data repositories freely available on the Web (Worldbankblog, 2020). The data used for the implementation essentially comprised of mixed text data. Usually, data in this format used to be of poor quality. The essence of using the clustering approach for analysis of this type of data is to automatically group the data, taking into cognizance the implicit similarities features that exist among them.

Unstructured dataset is sometimes classified as a qualitative data, and a type of data that cannot be easily processed using some of the well-known conventional tools and method (Pickell, 2018). Some of the common data that falls within the unstructured data include text, video, audio, mobile activity, social media activity, satellite imagery, surveillance imagery and several others to mention just a few. Generally, unstructured data is difficult to deconstruct, this is because there is no pre-defined model, as it cannot be organized in relational databases. Instead, such dataset is better managed using non-relational, or NoSQL databases. Fig. 3 graphically illustrates how the unstructured data differs from the structured data.

Procedures

In order to ensure that each cluster formed has a distinct feature and dissimilar to the data that belong to another cluster, a data dictionary was created. Each data in the input space was scanned and defined according to its type; the data of the same type consist of data that are of similar definition. This process continues until all the unstructured data in the input space are defined. In the course of clustering the data, only data of the same type are grouped together. Thus, numeric, alphabetic and data of special characters are grouped differently. There is also the need to establish that none of the data in the input space are left out, as each data must belong to a unique cluster. All the data in the input space must be assigned to an appropriate cluster based on their type, and prior tothe display of the generated output, unclustered data must be revisited, in case there may be any left over.

The proposed algorithm

In this study, the algorithm proposed for the clustering of unstructured dataset is represented in Fig. 2. The steps listed in the algorithm are implemented using java. Most of the well known partitioning algorithms such as k-means and EM clustering techniques requires the number of clustered to be indicated at the initial stage, however, the proposed technique automatically determines the number of clusters in the input space. During the clustering process, the proposed algorithm iterate between steps 2 and 6, until all the data in the input space are clustered. Finally, the data that are grouped together shares similar unique features, while these data are dissimilar to the data belonging to the other clusters. Fig. 4 shows the output of the clustered data based on the implementation of the proposed algorithm.

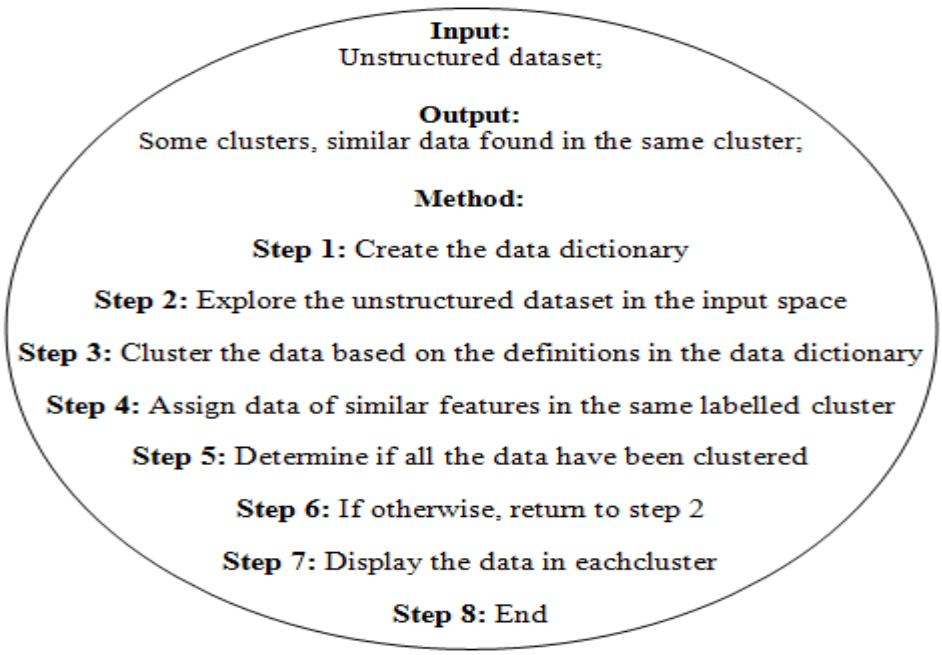

Fig. 2: Enhanced Descriptive Mining Algorithm (EDMA)

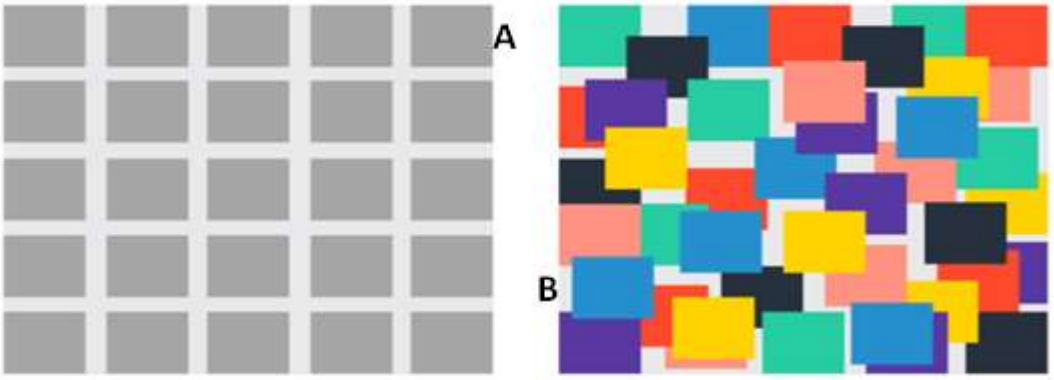

Fig. 3A&B: Structured and unstructured data adapted from (Pickell, 2018)

Implementation

The implementation of the proposed approach basically takes the input values and gives the corresponding cluster values as illustrated in Fig. 4. The algorithm proposed in this study is implemented using java programming language. The interface was designed in an interactive format. The proposed algorithm takes the unstructured data as input data, and after thorough exploration by the proposed algorithm, it displays the cluster in a distinct group based on their similarity features. The program written specifically identifies some letters, numbers and special characters as defined in the data dictionary; these are subsequently grouped accordingly. The data in the same cluster were found to be very similar to each other but different entirely with the data in another cluster. Fig. 3A&B illustrates graphically the views of the structured and unstructured data.

In the first part of the diagram, a relation (Fig. 3A) is well structured, this implies that there are some distinct rows and columns. A Dataset that is expected to go into a particular field knows ahead of time the exact domain value that can go into each cell as pre-defined in the table structure, and the types of data acceptable to those columns. The fact that it knows about the actual structure of that Dataset ahead of time, it makes optimization of data object to be more efficient.

Results and Discussion

Clustering improves the quality of the data as the process gives a resulting output that shows a high degree of relationship in the clustered data. One of the rationales for clustering dataset is to give a clear and accurate picture of the growing pattern of the business. The motive of this study was to cluster the unstructured dataset in the input spacebased on their similarity features.

The proposed techniques does not require initialization of desired number of clusters. This is determined automatically by displaying the approximate number of clusters at the implementation stage. The proposed technique supports a number of data type and not restricted to continuous values as in k-means.

The three clusters automatically generated in this study are labelled as cluster 1, cluster 2 and cluster 3. The proposed algorithm scan through each character in the input space with a view to identifying the different data types that exist in them. The detected data were subsequently grouped into clusters based on their type as shown in Fig. 4. The implementation was carried out in an interactive environment.

In the present study, evaluation was carried out by comparing the performance of the proposed technique with the performance of selected existing approaches as reported in the literature. Findings show that, the proposed EDMA in this study can be used in lieu of some state-of-the-art techniques, especially the k-means.

This study evaluates the performance metrics of the proposed technique, EDMA, to the two of the well-established clustering approaches that uses partitioning methods, specifically k-means and EM clustering. Findings from the evaluation are summarized in Table 1.

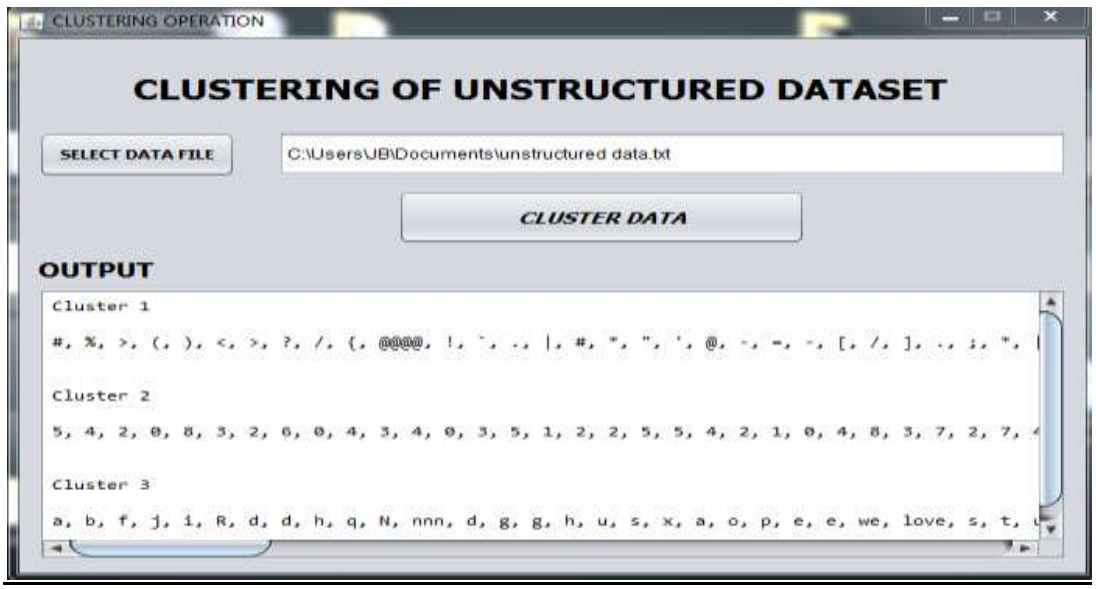

Fig. 4: Partitioning of the unstructured dataset into 3 clusters

Table 1: Performance metrics of k-means, EM clustering and the proposed technique

Metrics	k-Means	EM Clustering	The proposed technique
Number of Cluster	Number of cluster and maximum runs must be initially specified.	Number of cluster, maximum runs and maximum optimization must be initially specified.	No need for initial specification as approximate number of clusters is automatically determined.
Defined schema	Perform much better on structured data	Perform much better on structured data	It perform best on structured and Unstructured text data
Type of dataset	Clusters continuous variables	Clusters both continuous and categorical variables.	Clusters any type of text data
Data Tolerability	Does not tolerate outliers and missing values	Tolerate missing values.	Tolerate both outliers and missing values

1 **Number of clusters:** Most of the partitioning algorithms require that the number of clusters should be specified before the clustering begins. The proposed technique searches for similarities and dissimilarities in the dataset and put them in a separate cluster; it does not require initial specification of the number of clusters to be formed.

2 **Defined schema:** The k-means and EM clustering algorithm performs well on a structured dataset. The data to be explored using k-means should be properly organised and of numeric type. The EM clustering is capable of clustering alphanumeric data However, the proposed technique supports data of different types, including special characters. It also performs satisfactorily on unstructured dataset.

3 **Type of dataset:** k-means algorithm clusters only the numeric data. This is achieved by computing their mean; the mean of letters or special characters cannot be computed directly, the algorithm therefore, takes only numbers. This is a huge gap which the present study attempts to address. The proposed study clusters data regardless of the data type.

4 **Data tolerability:** k-means algorithm does not tolerate missing values or outliers. EM clustering tolerate missing and hidden values. The proposed technique tolerate both missing values and outliers.

The essence of comparing the performance of the said algorithms to the proposed technique based on a number of metrics was to establish its efficacy. The metrics used are: number of clusters, defined schema, type of dataset and data tolerability. The reason for choosing these two established algorithms is because, both algorithms cluster data using partitioning methods. The proposed technique also falls into this type of clustering techniques. The literature has also reveals some characteristics features of these algorithms based on the listed metrics. The performance of the three algorithms based on the listed metrics are summarized in Table 1

Conclusions

Clustering is an unsupervised learning approach that involves grouping of data based on their similarity features. Clustering approach is particularly useful or inevitable when there is no specific class to be predicted but the instances being explored are to be partitioned into natural groups.

Clustering is capable of grouping data that has similar features, and presents these data to the user in a more concise form. This paper presents the exploration of unstructured dataset using a descriptive mining algorithm otherwise known as clustering techniques. One of the benefits of clustering data is the tendency to solve classification problem.

Data of huge volume captured through sources such as email, xml, web pages are referred to as unstructured data since they don't normally have any defined schema. Data of these categories and others calls for advances in the techniques to be used in order to automatically understand, process, and summarize them.

This study proposed an Enhanced Descriptive Mining Algorithm (EDMA). Attempts made to evaluate the performance of the proposed technique in relation to some existing partitioning algorithm such as k-means shows a significant enhancement. In addition, the proposed algorithm can effectively be used in lieu of the other well established clustering techniques that uses partitioning approach for the clustering of text data.

Conflict of Interest

Authors have declared that there is no conflict of interest reported in this work.

References

Berkhin P 2006. A survey of clustering data mining techniques: Grouping multidimensional data. Springer, pp. 25-71.

Daiyan GM, Al Abid FB, Khan MAR & Tareq AH 2012. An Efficient Grid Algorithm for Faster Clustering Using K Medoids Approach. Paper presented at the *Computer and Information Technology* (ICCIT), 2012.

Ding C & Li T 2007. Adaptive Dimension Reduction Using Discriminant Analysis and k-means Clustering. Paper presented at the *Proceedings of the 24th International Conference on Machine Learning*.

Fahad A, Alshatri N, Tari Z, Alamri A, *et al*. 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3): 267-279.

Han J, Kamber M & Pei J 2012a. *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufman, Elsevier, USA.

Han J, Kamber M & Pei J 2012b. *Data Mining Concepts and Techniques:* (3rd Edition ed.): Morgan Kaufmann.

Jain AK 2010. Data Clustering: 50 Years Beyond K-means. *Pattern Recognition Letters*.

Joshi R, Patidar A & Mishra S 2011. Scaling k-medoid Algorithm for Clustering Large Categorical Dataset and its Performance Analysis. Paper presented at the *Electronics Computer Technology* (ICECT), 2011.

Kane F 2017. *Hands-on Data Science and Python Machine Learning* (Vol. *Birmingham, UK*): Packt Publishing Ltd.

Kumar TS 2014. *Introduction to Data Mining* (First ed.): Pearson.

Mirkin B 2012. *Clustering: A Data Recovery Approach*: CRC Press, 2012.

Mustafa YT, Tolpekin VA & Stein A 2011. Application of the expectation maximization algorithm to estimate missing values in Gaussian Bayesian network modeling for forest growth. *IEEE Transactions on Geoscience and Remote Sensing*, 50(5): 1821-1831.

Napoleon D & Lakshmi PG 2010. An efficient K-means clustering algorithm for reducing time complexity using

uniform distribution data points. Paper presented at the *Trendz in Information Sciences & Computing* (TISC), 2010.

Pickell D 2018. Structured vs Unstructured Data – What's the Difference. Retrieved from https://learn.g2.com/structured-vs-unstructured-data

Reynolds AP, Richards G & Rayward-Smith VJ 2004. The application of k-medoids and pam to the clustering of rules. Paper presented at the *International Conference on Intelligent Data Engineering and Automated Learning*.

Rodriguez A & Laio A 2014. Clustering by fast search and find of density peaks. *Science*, 344(6191): 1492-1496.

Shafeeq A & Hareesha K 2012. Dynamic clustering of data with modified k-means algorithm. Paper presented at the *Proceedings of the 2012 Conference on Information and Computer Networks*.

Sharma N, Bajpai A & Litoriya MR 2012. Comparison the various clustering algorithms of weka tools. *Facilities*, 4(7): 78-80.

Suh SC 2012. *Practical Applications of Data Mining.* Jones & Barlett Learning.

Wang K, Qi X, Liu H & Song J 2018. Deep belief network based k-means cluster approach for short-term wind power forecasting. *Energy*, 165: 840-852.

Wu YW, Tang YH, Tringe SG, Simmons BA, *et al.* 2014. MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2(1): 26.

World Bank Blogs 2020. Retrieved from https://blogs.worldbank.org/opendata on 18/01/20.

Xie J, Girshick R & Farhadi A 2016. Unsupervised Deep Embedding for Clustering Analysis. Paper presented at *the 33rd International Conference on Machine Learning*, NY, USA.

Zadegan SMR, Mirzaie M & Sadoughi F 2013. Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowledge-Based Systems*, 39: 133-143.