International Journal of Software Engineering and Computer Systems (IJSECS)
ISSN: 2289-8522, Volume 4 Issue 2, pp. 38-48, August 2018

DOI: https://doi.org/10.15282/ijsecs.4.2.2018.3.0047

PERFORMANCE EVALUATION OF SELECTED DISTANCE-BASED AND DISTRIBUTION-BASED CLUSTERING ALGORITHMS

Ajiboye, A. R. Olufadi, H. I.
Department of Computer Science
Faculty of Communication & Information Sciences
University of Ilorin, Ilorin, Nigeria.
ajibabdulraheem@gmail.com

ABSTRACT

Clustering is an automated search for hidden patterns in a datasets to unveil group of related observations. The technique is one of the viable means by which the patterns or internal structure of the data within the same collection can be revealed. Choosing the right algorithm to achieve clusters of good quality is usually a challenge, especially when the number of clusters cannot be pre-determined. This study focuses on evaluating a number of selected clustering algorithms in finding quality clusters in the data sets. To achieve the central objective of this study, prominent technique in both the distance-based and the distribution-based clustering algorithm, specifically k-means and EM clustering algorithm respectively are implemented in this study. The data sets on which the algorithms were implemented comprised of 1,309 records of passenger information that boarded a ship retrieved from rapidMiner open repository. Experiments were conducted and clusters were formed based on the number of chosen partitions, k. The qualities of the clusters formed are measured using the concept of external criterion, Normalized Mutual Information (NMI), to validate all the clusters formed. The resulting output of this study shows that, the distance-based algorithm find clusters of higher quality with NMI value of 0.912 out of a maximum achievable value of 1. The experiment further reveals the average execution time it takes each algorithm to form the cluster model. The findings of this study also unveiled some useful insight into the choice of clustering algorithm as regards their support for a particular data type and the ease of execution of each algorithm.

Keywords: clustering, data mining**,** k-means, EM-clustering, un-supervised learning.

INTRODUCTION

Clustering analysis is generally referred to as an unsupervised learning approach that seeks to identify or group objects based on their similarity features. The technique involve the grouping of elements into a number of clusters, on the basis of their common traits (Rodriguez et al., 2014). Clustering techniques using the K-means (Suh, 2012) and K-medoids algorithms (Berkhin, 2006), are typical distanced-based approaches. In distribution-based algorithms, there is an attempt to reproduce the observed realization of data points as a mix of predefined probability distribution functions (McLachlan et al., 2008). The descriptive technique is particularly useful in several areas especially classification purposes. It is an unsupervised learning technique as it group data object without consulting class labels (Han et al., 2012); It automatically unveils the hidden features or the patterns in the dataset.

There is no consensus definition for clustering, the study reported in (Napoleon et al., 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). Basically, the technique involves the division of data into groups, which most times also referred to as clusters. The object of the same cluster shares some features, and by the perspective of machine learning, clusters directly correspond to hidden patterns (Berkhin, 2006).

K-means algorithm groups data based on distance measures; it 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 describe it as the leading distance-based clustering technique. Therefore, this makes it the algorithm of choice in this study.

Also, the Expectation and Maximization (EM) Algorithm group data based on probability distribution and being the most reported algorithm in this category, it is selected for implementation for the clustering of data in this study. Apart from these two descriptive algorithms being considered, it is worth to note that clustering may also be hierarchical, density or grid-based, while hierarchical algorithm does its grouping as crystals grows, partitioning algorithms learn clusters directly.

The use of any type of clustering algorithms come with a number of challenges (Jain, 2010); typical among the challenges is the determination of the exact number of clusters to be found in the given data. In order to address these challenges, it is so paramount to determine the performance of a number of clustering algorithms. This study therefore, proposed the implementation of some prominent algorithms with a view to unveiling their strengths. The objective of this study specifically focuses on the implementation of the two leading algorithms that operate based on distance and distribution measures. The two selected algorithms are implemented using an online public dataset retrieved from rapidMiner open repository.

The clusters formed based on each clustering algorithm are evaluated in order to determine the quality of the clusters using Normalized Mutual Information (NMI). The mutual information is an external validity measure and one of the recognized evaluation measures for estimating the clustering quality (Strehl et al., 2000). Evaluation using this technique shows the extent to which the cluster finds in the data matches with some external structure.

The findings of this study unveil the performance of each algorithm and gives a useful insight into data clustering using the techniques of distanced based and distribution based approaches. The rest of this paper is structured as follows: In section 2, the strategies used in grouping data objects using the two techniques are discussed. Some related studies are discussed in Section 3, while in Section 4, the method proposed for this study is presented. This is followed by Result and Discussion shown in Section 5, while the evaluation of the resulting outputs of this study is shown in Section 6. This study concludes in Section 7.

DISTANCE AND DISTRIBUTION-BASED CLUSTERING ALGORITHMS

Distance-based clustering: Prominent distance-based clustering algorithms are k-means and k-medoids (Jain, 2010); other well-known algorithms in this category include k-modes, PAM, CLARANS etc. But k-means being the most popular among the clustering algorithms that are distance-based (Andreopoulos et al., 2009), is the algorithm implemented in this category. Although, k-means algorithm was first proposed more than five decades ago, it is still very relevant and widely used for clustering till the present time. There are some reasons responsible for its widely acceptance, these include: its simplicity, efficiency, empirical success and its ease of implementation (Jain, 2010). It is important to carefully select

algorithm that appears most suitable for the clustering of a particular dataset. This is because clustering algorithms in most cases tend to find similarities in the data irrespective of whether or not any clusters are actually present.

K-means: This is a very common technique in machine learning where a bunch of data is explored to find interesting clusters of patterns in the data, based on the attributes of the data itself (Kane, 2017). When there is need to group data, the K-means method can only be applied for a set of objects whose mean can be computed (Han et al., 2012). The use of k-means clustering involves splitting out data into *k* groups, that is where the *k* comes from. It has to do with how many different groups the data is to be partitioned into, and this grouping is achieved by finding *k* centroids. For instance, suppose a data set, *D*, contains *n* objects in Euclidean space, the k-means techniques distribute the objects in *D* into *k* partitions, $C_1 \ldots . C_k$; $C_i \subset D$ and $C_i \cap C_j = \phi$ for $(1 \leq i, j \leq k)$. A distance-based clustering technique uses the centroid of a cluster C_i to represent that cluster. The difference between an object $p \in C_i$ and c_i which is the representative of the cluster is measured by dist. (p, c_i) where dist. (x, y) is the Euclidean distance between two points x and y. The clustering algorithm for k-means is represented in Figure 1.

Step 1. Place *k* points into the space *S*

Step 2. Compute the Euclidean distance, *d*, between the data points

$$d = \| P(x, y) - c_k \|$$

Step 3. Assign each object to the cluster that has the closest centroid

Step 4. After all the data points have been assigned, re-compute the positions of the *k* centroid using the relation:

$$c_k = \frac{1}{k} \sum_{y \in c_k} \sum_{x \in c_k} P(x, y)$$

Step 5. Repeat the steps until the centroids remain constant.

Figure 1. The k-means algorithm.

As shown in Figure 1, k-means algorithm basically iterates between two methods, as one step updates clusters in line with the minimum distance rules, the other step update centroids as the centre of gravity of clusters (Mirkin, 2012). After the grouping of the data, it recalculates the new centroid of each cluster. The k-means algorithm requires three user-specified parameters, these are: the number of partitions usually represented by k, cluster initialization, and the distance metric (Jain, 2010). The use of k-means requires making decision on the number of final clusters at the beginning; then clusters that appear to be poorly initialized can be corrected and reallocated on optimality criteria (Suh, 2012). The k-means algorithm is sensitive to empty values, outliers and support only data that can define Mean.

Distribution-based clustering: In this clustering approach, one attempts to reproduce the observed realization of data points as a mix of predefined probability distribution functions (McLachlan et al., 2007). The clustering model of this category is based on the

assumption that the data is generated by a mixture of underlying probability distributions (Fahad et al., 2014). In other words, this approach assume that the data is generated from a mixture distribution, where each cluster is described by one or more mixture components (Jain, 2010). One of the most prominent probability distribution-based algorithms is EM clustering algorithm, others in this category include Self Organizing Maps (SOMs), CLASSIT etc. This study therefore, implements EM clustering algorithm in this category and the algorithm is discussed briefly in the next sub-section.

The EM-clustering algorithm: This is a framework for approaching maximum likelihood or maximum posteriori estimates of parameters in statistical models (Han et al., 2012). It basically consist of two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step, given the current cluster centres, each object is assigned to the cluster with a particular centre that appears closest to the object. In other words, each object is expected to belong to the closest cluster. In the M-step, given the cluster assignment, for each cluster, the algorithm adjusts the centre in such a way that the summation of the distances from the objects assigned to this cluster and the new centre is minimized. In other words, Expectation step assigns objects according to the parameters of distribution clusters, while maximization step finds the new clustering or expected likelihood in distribution model-based clustering (Han et al., 2012). The algorithms in this category optimizes the fit between the given data and some predefined mathematical model (Fahad et al., 2014). The EM clustering is based on the assumption that generation of data is by a mixture of underlying probability distributions. The algorithm is represented in Figure 2.

1. Input the dataset (x), number of clusters (m), max no. of iteration (i)
2. **E-step:** Compute the expectation of the complete data log-likelihood

   $$Q(\theta,\theta^T) = E\left[\log p(x^g, x^m \middle| \theta) x^g, \theta^T\right]$$

3. **M-step:** Choose a new parameter estimate to maximize the Q-function

   $$\theta^{t+1} = \max Q\left(\theta, \theta^T\right)$$

4. **Iteration:** Increment $t = t + 1$; repeat steps 2 and 3 until the convergence criteria is met.

   where θ is an unknown hidden variable.

Figure 2. The EM clustering algorithm, adapted from (Fahad et al., 2014).

RELATED STUDIES

The distance-based clustering algorithm is well reported for the grouping of data sets. Typical algorithms in this category include k-means, k-medoids and k-modes. The application of clustering ranges from astronomy to bioinformatics, bibliometrics, and pattern recognition. These techniques have been applied for grouping of data in several other areas of research such as images (Lin et al., 2014), video object segmentation (Huang, 1998), document clustering (Forsati et al., 2013) etc. But sometimes, distance-based clustering technique may not be very suitable for the clustering of a particular data; probability distribution approach may be suitable for such data. Unlike distance-based techniques, some clustering algorithms are distribution-based and perform excellently well. A typical example of algorithm in this category is EM clustering. A number of studies that have applied this

technique for the exploration of data using the concept of unsupervised learning approach is discussed in this section.

Apart from using the clustering algorithms in several areas already mentioned, other important areas the algorithms have been found very useful include biomedical (Andreopoulos et al., 2009) and wireless sensors network (Jiang et al., 2009). The clustering algorithm proposed in (Cutting et al., 2017) was found to be useful in browsing large document collection. The study viewed clustering as an information tool and proposed a document browsing technique that employs the use of document clustering as its primary operation.

In order to achieve a better understanding of the patterns in the data, the study reported in (Reynolds et al., 2004) proposed the use of k-medoids to identify sets of similar rules with a view to unveiling the hidden patterns of the data. The algorithm was found to be very efficient and showed better tolerance than the k-means. The k-medoids is widely proposed in several other studies (Daiyan et al., 2012; Joshi et al., 2011; Reynolds et al., 2004; Zadegan et al., 2013). In order to optimize the performance of k-means, the algorithm was modified in (Daiyan et al., 2012), the modification resulted in a faster clustering and better cluster quality.

In a related study proposed in (Ding et al., 2007), the authors in the course of their research reported the use of k-means for the generation of class labels. Generally, k-means is widely used for the creation of cluster models due to its capacity to group data within a very short time. In the proposed study, in order to achieve an improved classification, the researchers combined the approach of linear discriminant analysis in such a way that it adaptively select the most discriminating subspace.

The use of EM clustering technique is also well reported in the literature. In order to have a good quality of clusters found in a big data set of high dimension, an enhancement of EM algorithm was proposed in (Ordonez et al., 2002). The EM Clustering techniques usually take longer times in creating cluster model and since the researchers were dealing with big data, the enhancement proposed was to speed up the time it takes for the formation of clusters. The study proposed in (Ambroise et al., 1997), also find the use if EM algorithm very useful for the clustering of spatial data.

Segmentation of images (Carson et al., 2002) proposed the use of EM algorithm to estimates a clustering model. The model shows the segmentation of the image into some regions; the resulting output of the study analysed a description of each region's in terms of colour and texture characteristics.

A study that compares the performance of different types of clustering techniques reported in (Sharma et al., 2012) focused on capturing the time it takes each algorithm to build the clustering model. Some algorithms take too much time to form cluster irrespective of the number of partitions. Also the slow pace of finding clusters in data sets may be as a result of the size of the data involved and the effectiveness of the algorithm being implemented.

This present study also measures the quality of each cluster formed in order to determine the accuracy of the model in addition to other information revealed. The reviewed literatures revealed how different clustering techniques have been implemented to explore data sets for descriptive purposes. In particular, a number of algorithms in both distance-based and distribution-based techniques are well reported. The present study therefore, focuses on evaluating the performance of the leading algorithms in these two categories of clustering types with a view to making comparison and unveils the strength of each technique in finding the quality clusters from the dataset.

MATERIAL AND METHODS

Data Collection: The data explored in the course of this study is a public dataset retrieved from rapidMiner open repository. The data are of high dimension and comprised mainly of passengers' information that boarded a ship. The attributes of interest in the dataset for the purpose of this research are the status of survivors, passenger class and their age bracket. The data comprised of 1,309 records and this dataset is explored in this research using the concept of un-supervised learning approach.

Data Transformation: The retrieved data are transformed from its present form to a format that can make it suitable for mining. It should be noted that most descriptive algorithms, especially k-means only support numeric values and it is sensitive to outliers. Also, since the focus of this study is on distance and distribution-based clustering; the datasets on which the algorithm is implemented must necessarily be transformed into a format that is appropriate or acceptable to these algorithms. In view of this, the attributes required for clustering is therefore transformed to numeric values. This was achieved through the conversion operator that converts nominal values to numeric as shown in the experimental setup. In order to achieve faster and accurate clusters; the numeric attribute given the clustering role is normalized. This makes the entire data in this field to fall between 0 and 1.

Algorithm implementation: The clustering algorithms implemented in this study are the k-means, which is a representative of distance-based algorithm and the EM-clustering, a representative of the probability distribution algorithm. Each of these algorithms finds cluster in the already pre-processed data and the experimental setup for both techniques is as shown in Figure 3. The first operator (import) in the setup reads the data to be explored, originally stored in an excel file. This is followed by a type conversion operator which is inevitable for the conversion of data source type, to its numerical equivalence. The next operator is the clustering operator. In the first experiment, k-means was used and in subsequent experiments this was replaced by EM clustering operator. The last operator shown in Figure 3, is for export. The clusters formed are exported to another excel file through the result port. This gives the opportunity to compute the cluster quality in relation to the data sets being explored.

The implementation is done in rapidMiner software environment. There are a number of reasons for its choice in this study. The software has all the basic algorithms required for data mining tasks; for predictive mining, descriptive mining and pre-processing of data to make the data suitable for mining. The software is also found to be very suitable in representing data visually. It has operators that is capable of reading several file formats, including Attribute-Relation File Format (ARFF), these are files known from the machine learning library in Weka.

During the experiments, similar configuration settings were followed to configure the setup for both algorithms based on the configuration settings represented in Table 1. Since the essence of clustering the data in the capsized ship was to reveal both the survived passengers and those that did not survive in relation to the passenger class; this is the reason for chosen k as 2, being the number of data partition.

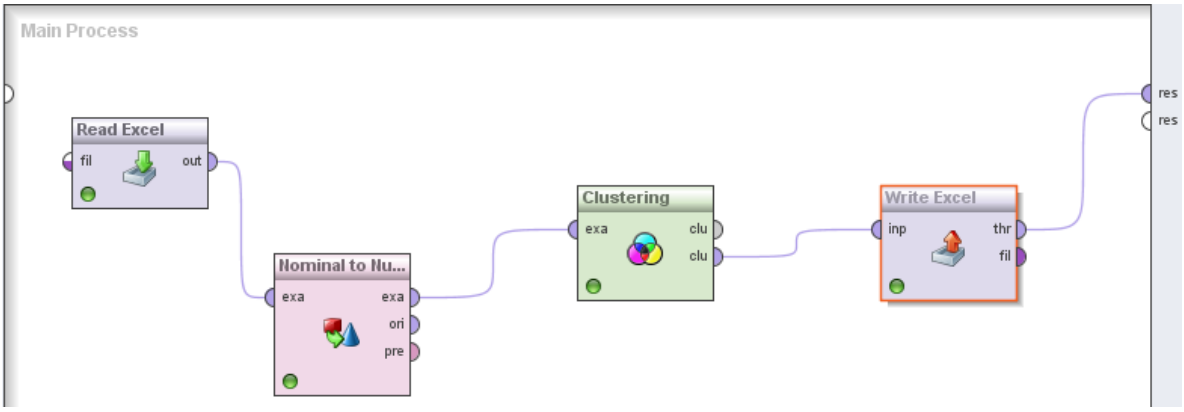

Figure 3. Experimental setup for data clustering

Table 1. Configuration settings for both K-means and EM clustering algorithms

Algorithm	Number of clusters	Max. runs	Numerical measure	Max optimization
Distance-based (k-means)	2	10	distance	100
Distribution-based (EM)	2	10	random	100

RESULTS AND DISCUSSION

This section illustrates the clustering results generated in the course of implementing each technique of distance-based and distribution-based algorithms. Figure 4, shows the clusters formed in the course of implementing k-means, while Figure 5, illustrates the clusters formed as a result of implementing the EM clustering algorithm. The partitions shown from k-means result appears more distinct, and comparing the quality of the clusters formed, it appears higher. The actual quality of the clusters formed is determined using the numeric values of the clusters formed as contained in the write excel operator as shown in the configuration setup. It is those numeric values that gives the visualized results in Figures 4 and 5. Two clusters were formed since the value of k in the configuration settings is 2. Based on the content of the data being explored, the blue colour indicate the survived passengers while the red colour indicate the cluster of passengers that did not survive. The example sets in each cluster comprised of cluster 0 (809) and cluster 1 (500). This analysis of the example set in each cluster is delivered through the export operator to the output file.

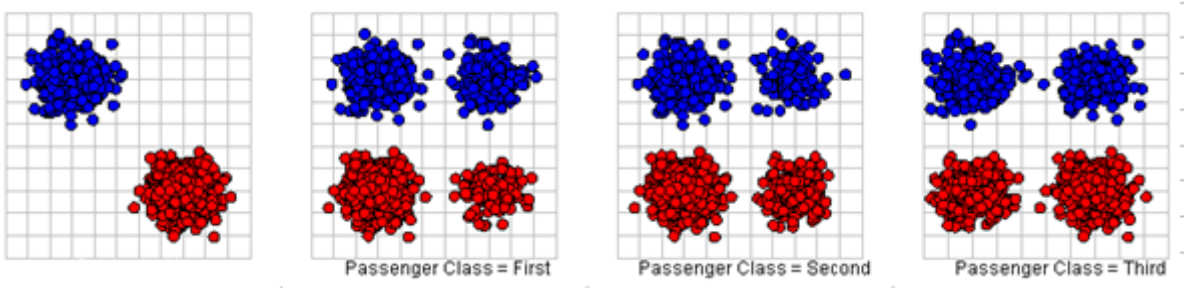

Figure 4. Clusters formed through k-means implementation

Similarly, the partitions in Figure 5, being the output of EM implementation also record a well distinct partition but with higher execution time compared to k-means. In this case, the red colour illustrates the cluster of passengers that survived; while the blue colour indicate the non-survived passengers. The numerical values of this graphical representations are delivered to the output file as shown in the configuration setup (see Figure 3) through the export operator.

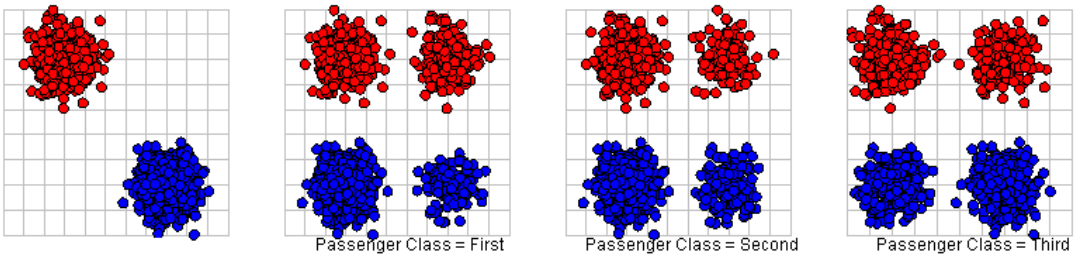

Figure 5. Clusters formed through EM Clustering implementation

To know the level of accuracy of the clusters formed, it is necessary to compute the quality of each cluster as some data may not be correctly clustered. A number of techniques have been reported in the literature. Prominent among these techniques include some statistical indexes such as Rand index, Calinski-Harabasz index, Davies-Bouldin index, Silhouette index, Dunn index, Bic index (Rendón et al., 2011). These clustering measures are usually referred to as internal validity index. The external validity measures include F-measure, NMI-measure, Purity, Entropy etc.

EVALUATION OF RESULTS

The results of evaluating the qualities of the clusters formed with regards to the clustering algorithms implemented in this study are presented in this section. In order to determine the qualities of the clusters formed, the Normalized Mutual Information (NMI) is computed. NMI is for measuring the clusters quality with respect to a given class labels of the data sets (Fahad et al., 2014). It is an external measure because it is mandatory to know the class labels of the instances. NMI-measure is widely used in information theory to measure the mutual independency of two random variables. Intuitively, it measures how much information a random variable tells about the other. NMI is usually measured between 0 and 1. The NMI of value 1 denotes strong similarity; but the similarity reduces as the value of NMI tends to 0. According to Han et al. (2012), given 2 frequent patterns, P_α and P_β, let $X = \{0,1\}$ and Y = {0,1}, be two random variables representing the appearance of P_α *and* P_β respectively. Mutual information $I(X,Y)$ is computed as:

$$\mathrm{I}(X;Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \qquad (1)$$

All logs are in base 2.

$$\text{where } P(x=1,y=1) = \frac{|D_\alpha \cap D_\beta|}{|D|} \qquad (2)$$

$$P(x=0, y=1) = \frac{|D_\beta| - |D_\alpha \cap D_\beta|}{|D|} \quad (3)$$

$$P(x=1, y=0) = \frac{|D_\alpha| - |D_\alpha \cap D_\beta|}{|D|} \quad (4)$$

$$P(x=0, y=0) = \frac{|D| - |D_\alpha \cup D_\beta|}{|D|} \quad (5)$$

where D is a basic object in the Database; α and β are random variables. The value of NMI computed for the clusters formed is represented in Table 2.

Table 2. Clusters quality and time for cluster model formation

Algorithm	Execution time (s) for cluster model formation	NMI value
Distance-based (k-means)	1.0	0.912
Distribution-based (EM)	5.0	0.865

CONCLUSION AND FURTHER STUDIES

This paper presents the performance of selected clustering algorithms in finding the quality clusters in the data sets. The study specifically experiments on both the distance and distribution-based algorithms. Although, clustering algorithms in these categories are many, but the most prominent in each of the category is the focus of this study. This study also determines the quality of clusters produced through each algorithm with respect to the datasets explored. As illustrated in Table 2, the cluster model created as a result of implementing k-means has NMI value of 0.912, this implies that, the clusters formed is of good quality. The value is very close to the maximum achievable value of 1; this study also determines the execution time it takes for each algorithm to create a cluster model.

In grouping the data sets into categories or clusters, the objects that shares similar features are expected to be in the same cluster, while data in different clusters are expected to show dissimilar features. This therefore, makes clustering a good classification technique. This study has shown the strength of each algorithm implemented and also illustrates a standard approach of determining the quality of the clusters produced.

This study recommends k-means algorithm for the clustering of data sets whose Mean can be defined as the algorithm does not tolerate missing values; unlike the EM clustering technique. Also, the k-means is found to show a good strength in the grouping of large set of data within a very short time. This work can be extended in the near future to encompass other known clustering techniques that cluster data based on a grid, hierarchical and density approaches.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their useful suggestions.

REFERENCES

Ambroise, C., Dang, M., & Govaert, G. (1997). Clustering of spatial data by the EM algorithm *geoENV I—Geostatistics for environmental applications* (pp. 493-504): Springer.

Andreopoulos, B., An, A., Wang, X., & Schroeder, M. (2009). A roadmap of clustering algorithms: finding a match for a biomedical application. *Briefings in Bioinformatics, 10*(3), 297-314.

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

Carson, C., Belongie, S., Greenspan, H., & Malik, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(8), 1026-1038.

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (2017). Scatter/gather: A cluster-based approach to browsing large document collections. Paper presented at the ACM SIGIR Forum.

Daiyan, G. M., Al Abid, F. B., Khan, M. A. R., & Tareq, A. H. (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.

Forsati, R., Mahdavi, M., Shamsfard, M., & Meybodi, M. R. (2013). Efficient stochastic algorithms for document clustering. *Information Sciences, 220*, 269-291.

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

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery, 2*(3), 283-304.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*.

Jiang, C., Yuan, D., & Zhao, Y. (2009). Towards clustering algorithms in wireless sensor networks-a survey. Paper presented at the Wireless communications and networking conference, WCNC 2009. IEEE.

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.

Lin, C.-H., Chen, C.-C., Lee, H.-L., & Liao, J.-R. (2014). Fast K-means algorithm based on a level histogram for image retrieval. *Expert Systems with Applications, 41*(7), 3276-3283.

McLachlan, G., & Krishnan, T. (2007). *The EM algorithm and extensions* (Vol. *382*): John Wiley & Sons.

McLachlan, G., & Krishnan, T. (2008). *The EM Algorithm and Extensions.* Whiley Series in Probability and Statistics. John Wiley & Sons, New York, USA, second edition.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine, 13*(6), 47-60.

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

Napoleon, D., & Lakshmi, P. G. (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.

Ordonez, C., & Omiecinski, E. (2002). FREM: fast and robust EM clustering for large data sets. Paper presented at the Proceedings of the eleventh international conference on Information and knowledge management.

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *International Journal of computers and communications, 5*(1), 27-34.

Reynolds, A. P., Richards, G., & Rayward-Smith, V. J. (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.

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

Strehl, A., Ghosh, J., & Mooney, R. (2000). Impact of similarity measures on web-page clustering. Paper presented at the Workshop on artificial intelligence for web search (AAAI 2000).

Suh, S. C. (2012). *Practical Applications of Data Mining:* : Jones & Barlett Learning.

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

Zadegan, S. M. R., 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.