A Comparative Study of Clustering Algorithms for Mixed Datasets

Saad Harous[1], Maryam Al Harmoodi[2], Hessa Biri[3]

[1,2,3]*College Information Technology, UAE University, Al Ain, UAE*
harous@uaeu.ac.ae

Abstract: Clustering groups, a set of elements in a manner that elements in the same categoryhave more common characteristics (based on a given set of attributes) among themthan to elements in other categories. Each group is called a cluster. Clustering is used in many areas: sensor networks, social networks, health, business and other applications. There are many different clustering algorithms with different parameters. The appropriate clustering algorithm and parameter settings depend on data set and the problem being solved. Some work only on numerical data and other on mixed data. Our aim is to do a comparative study of these algorithms.

Keywords: Clustering, Mixed Data, K Means, Similarity measure.

I. INTRODUCTION

Data mining is an area in computer science where discovering knowledge from large dataset. Different types of clustering are used in data mining. Clustering is the process of dividing the data into categories where elements in cluster have common characteristics. Some of these clustering techniques are implemented in WEKA which is an open source [1].

The authors in [1] compared between three clustering methods: partition method [2], hierarchical method [2] [3] and density-based method [2] [3] [4]. Using an example of each one of the methods, for example simple K-means is example of partition method and so on. WEKA tool is used to run for example simple K-means on forest fire dataset and other algorithms to evaluate the clustering algorithms. The author used different parameters to decide which one is the best. The parameters used are: number of clusters, number instances, number of repetition, within clusters the addition of squared errors, time used to construct model and un-clustered instances. To choose the best algorithm, we should consider how it takes to construct a model, how cluster instances are distributed, number of repetitions and squared error [1].

This paper[5] compares four clustering algorithms: K-mean algorithm, HC algorithm, SOM algorithm and EM algorithm. The comparison is done based on four factors: the dataset's size, how many clusters, kind of dataset, and what software is used. For each parameter four tests are performed, one for each clustering techniques. It starts by using a large amount of data then by usingless data. According to the first factor which is the size of dataset in the huge dataset 600 rows and 60 columns are used while in the small dataset 200 rows and 20 columns are used. The small dataset is a subset of the large dataset. HC and SOM show better results when the dataset is small but the other two algorithms K-mean and EM becomes better when using a huge dataset. In the number of the cluster k will be chosen as (8,16,32 and 64) to make the comparison easier. Both random dataset and ideal dataset are used. The last factor is the type of software two packages are used here (LNKnet package [6] "UNIX environment" and Cluster and TreeView package [7] "Windows environment"). These two packages do not have an effect on the performance of the algorithm since the same is is obtained using either one of these packages [5].

The authors in [8] considered the following algorithms: K-prototype, KMCMD, K-centers, Improved K-prototype, KHMCMD, DH, SBAC, TMCM, M-ART, CAVE, MSOINN, BILCOM, AUTOCLASS, SVM Clustering, GFCM, KL-FCM-GM, Fuzzy K-means, Fuzzy K-prototype, MixSOM, GMixSOM, FMSOM and UFLA. They compared these algorithms based on certain criteria's such as the scalability, shape of cluster, sensitivity to noise/outlier, high-dimensional data, input data in any order, interpretation of results, depending on prior knowledge and user-define parameter, data structure, type of cluster, representation of cluster and lastly the time complexity [8].

Clustering algorithms for numerical data are very well studied. There are many clustering algorithms for mixed (numerical and categorical) data but to our knowledge there is no compressive study available that compares these algorithms to each other. It would be very useful if we are able to know which one of these algorithms is more suitable for which application. The aim of our work is to choose the well-known algorithms from the available clustering algorithms and do a theoretical and experimental comparisons among them. Based on this study we will give recommendations for developers which algorithms and what parameters setting are suitable for which application.

Clustering Algorithms Considered

In our study we considered the following clustering algorithms:

978-1-5386-9346-9/19/$31.00 ©2019 IEEE

1. *K-Means Clustering Algorithm*

K-means algorithm [9] is a simple algorithm (does not use supervise learning) that solves the general clustering problem. The technique uses a simple approach to divide a set of data elementsinto a number of groups called clusters (assume K clusters) fixed before we start the clustering. The main idea is to specify K centers, one for each cluster. Then, select randomly K points as cluster centers. The elementsin the data set are made members of the cluster with the closest center based on to their Euclidean distance to the centers. The centroid or mean of all objects in each cluster is calculated. These steps are repeated until there is no changes in other words each cluster contains the same points in consecutive phases. K-means is a clustering techniqueused very often but it does not work for categorical attributes.

2. *K-Modes Clustering Algorithm*

K-modes [10] solves K-means problem by removing the limitation of being used for only numerical datasets while having the same effectiveness. The K-modes technique is a modification of K-means technique that "uses a simple matching dissimilarity measure for categorical variables, modes instead of means for clusters" [10]. Also it uses"a frequency-based method to update modes in the clustering process to minimize the clustering cost function with the matching dissimilarity measure for categorical variables" [10]. These modifications when applied to the K-means approach enable the K-modes technique to group a huge amount of non-numerical data elements that represent real world data in an efficient manner.

3. *K-Prototype Clustering Algorithm*

This algorithm is used for mixed data types [11]. In this algorithm, new definitions for distance and clusters centers are given. Mode values are used for categorical data and mean values are used for numerical data. The approach used to calculate the distance between the center and a point (categorical data) is Hamming distance, which may result in inaccurate value. A new cost function and a distance measure is proposed in [12] to overcome this shortcoming.

4. *Unsupervised Feature Learning Clustering Algorithm*

Unsupervised Feature Learning clustering algorithm [13] is designed to handle mixed data type. It uses fuzzy adaptive resonances theory to clusterbig and small data. This technique performs efficient even when used in an unsupervised learning. Also, there is no need to specify how many clustersare needed from the start. This algorithm generatesa more suitable clustering set.

5. *TMCM Learning Clustering Algorithm*

The two-step clustering method (TMCM) [14] is designed to handle very large data sets. It can handle numeric and categorical attributes. The similarity or relationships amongelements that have categorical attributes is structured based on the ideas of co-occurrence. Using these computed relationships, all categorical attributes are transformed into numeric attributes. Most clustering algorithms handle every attribute as a single entity and ignore the relationships between them. But this algorithm studies the relationships between elements to compute the similarity between pairs of objects. A sensible numeric value can be assigned to non-numerical objectsbased to the computed relationship.

6. *Affinity Propagation Clustering Algorithm*

Affinity propagation clustering [15] is an "exemplar based clustering method" that has shown efficient performance on many kinds of data. Traditional clustering algorithms start by selecting N initial points as clusters centers, but this algorithm considers all data points as possible clusters centers. This method is appropriate for exploratory data analysis because of its ability to take non-metric similarities as input. This algorithm uses a more efficient mixed similarity measure to calculate distances between each twoelements. This measure will compute the weight coefficients for numeric attributes. According to the authors [15] this work not only clusters mixed data well, it does the same for pure numeric or categorical data.

7. *Unsupervised learning with mixed numeric and nominal data*

The authors in [16] present a "Similarity-Based Agglomerative Clustering (SBAC) algorithm" that handles mixed numeric data and nominal features. The similarity measure is suggested by Goodall [17] for biological taxonomy. This technique assigns higher weight for value that represents characteristics that are not common but have the same similarity computations. The feature values'underlying distributions is made up with no assumptions. It is adopted to describe the similarity measure between pairs of elements. The goal of this scheme is to build a dendrogram and simple distinctness heuristic used to find a grouping of the data. The common way to cluster analysis (numerical taxonomy) that symbolize data items as points in a multidimensional metric space and use distance metrics such as: Euclidean and Mahalanobis measures, so we are able to define similarity between objects. In contrast, theoretical clustering techniques use conditional probability valuation as a wayto define clusters' relationship [16].

8. *Clustering heterogeneous data with k-means*

Unsupervised Feature Transformation (UFT)-k-means, can change non-numerical attributes into numerical attributes, with traditional k-means algorithm for heterogeneous data grouping. The values of changed

features can be normalized. Mixed data may be unified to have only numerical attributes by using a combination of UFT and k-means. Once the data is unified, it clustered in an efficient way. They use Gaussian mixture models (GMM) [16], with UFT for clustering heterogeneous data, these models have some problems to connect with initialization instability and reliance [17]. The UFT technique is based on the fact that a distance is added to every value of the original non-numerical features. What make the Gaussian mixture models (GMM) less stable that the transformed numerical values provide more choices of initialization. Gaussian has the following features: (1) gives a description of the common probability distributions of data element inapplications applicable in real life; (2) has characteristics that may express mutual information (MI) and entropy in a simple way;(3) its distribution parameters can be estimated withoutknowing the data range in advance. Also, we can use hierarchical clustering algorithms with UFT for heterogeneous data clustering. It is highly influenced by outliers and the clustering structure is not updated when the data is being processed. But it has the following advantages: (1) during the clustering process no parameter is required except the number of clusters represented by k; (2) UFT for non-numerical attributes is robust because it is based on mutual information (MI); (3) UFT can assignsuitable numerical values to the original non-numerical attributes. This way it does not have to use the Hamming distance in the dissimilarity measurement for grouping the data; and (4) UFT enables principle component analysis (PCA) to visualize heterogeneous data in meaningful way [18].

9. Automated variable weighting in k-means type clustering

The main problem of using k-mean algorithm is the selection of variables. Since k-mean algorithm treats all variables as having the same weight in the clustering process it cannot select variables automatically. In practical situation, attributes, for business related problem clustering such as customer segmentation, are selected according tothe compression of the business problem being treated and what kind of data is involved. Many variables (in the range of hundreds) are mostly extracted from the data set in the initial phase which results to a space with very high dimension. In general, an interesting clustering structure typically is formed in a subspace that is defined by a subset of the selected variables during initial phase. Identifying this subset of variables is essential to be able to find this clustering structure. Weighting in k-Means (W-k-mean) algorithm can automatically assigned weight to the involved variables based on the influence of these variables in the clustering process. A new weight for each variable based is assigned based on the difference of the cluster distances. The new weights are used in choosing the cluster memberships of objects in the next iteration. The weights are used to identify important variables for clustering process. The variables, that may cause noise to the clustering process, can be removed. W-k-mean improves the k-means algorithm by adding a new step which updates the variable weights based on the current partition of data. The weights assigned to the variables by W-k-means measure the influence of these variables during the clustering process. Noisy variables are assigned very small weights which eliminate/reduce their influence. “The weights can be used in variable selection in data mining applications where large and complex real data are often involved” [19].

10. Coupled interdependent attribute analysis on mixed data

Coupled Interdependent Attribute is functional couplings that only rely on the given data without any domain knowledge. There are some solutions when modeling the interdependence for mixed data, one of these solutions is to convert discrete values into numerical values or discretize numerical attributes into discrete attributes. This approach leads to an information loss and it is very complicated. It uses couplings via frequency, co-occurrence and correlation. Individual couplings will effectively capture certain hidden information from the data. Several attempts have been made to model the interdependence within categorical attributes [20] and within numerical attributes [21] individually. Interdependence of heterogeneous attributes is divided into two categories. Firstly, the Attribute Coupling for discrete data. (1) Coupling in discrete attribute. It represents the discrete data by converting the original space spanned by discrete attribute into a new space, and it is based on pairwise similarity between values of each discrete attribute. (2) it is related to coupling context and weight, it is not fair to simply treat every pair of attributes to have an equal coupling weight, because some cases are related, and some are not, and it is based on relevance and redundancy. Also, they must satisfy two conditions.

11. A better k-prototypes algorithm for mixed numeric and categorical data

The k-prototypes scheme is one of the methods for clustering mixed numeric and categorical attributes data objects. A more efficient k-prototypes approach for mixed numeric and categorical data uses a new concept named the distribution centroid to represent the prototype of categorical attributes in a cluster. Then it integrates the mean with distribution centroid to represent the prototype of a cluster where the objects have mixed data. Also, it uses a new dissimilarity measure that takes into consideration the importance of each attribute in evaluating the dissimilarity between data elements and prototypes. The goal of this algorithm is to have an efficient representation for the categorical attribute in a

mixed prototype because the mean is good enough for the numeric attribute part. But it should take into account the importance of different attributes towards the clustering process. This scheme captures the properties of clusters with mixed attributes more accurately than other methods. It considers the important effect of different attributes on the clustering process by making use of the new dissimilarity measure. It computes automatically the important effect on the clustering process based on the partition of data elements [22].

12. Feature weighting in k-means clustering

A set of scalar features is heterogeneous if after clustering, the interpretation of clusters based on one set of attributes is independently of the clusters based on another set of attributes [23]. It is unusual to ignore the various different types of features. To summarize the work in [23] and results, their first contribution is, defining a spin between data objects, and weighted sum of appropriatespin on single component. Their second contribution is to use a convex optimization formulation; they "generalize the classical Euclidean k-means algorithm to employ the weighted distortion measure" [23]. The third contribution [23] is to generalize the Fisher's discriminant analysis. Theypresented an optimal feature weighting that results in grouping objects together. Optimal feature aims to get the minimum average within cluster spin and maximize the average among clusters scuttle along all feature spaces.

13. Geometrical codification for clustering mixed categorical and numerical databases

This method assigns a pair of polar coordinates in the unit circle for every category. Each circle is divided in the same number (number of categories) of regularly distanced points. The polar codifications have two advantages: 1) all attributes have the same effect in a clustering process even if there are many categories. 2) all categories have the same features so that the converted set of points is balanced. In general, this is a good method but there is a problem in the case where the number of points to split the circle is large which leads to a large codification error. Spherical coordinates can be used to reduce the error. [24]. Distributing regularly points in 3D is not as simple as the 2D. The number of points is chosen carefully to be used to codify a specific attribute with N categories in the regular polyhedra should be used with the number of node higher than or same as the number of categories.

In Table 1, we are comparing between the 13 algorithms: K-means, K-modes, K-prototype, Unsupervised feature learning clustering algorithm, A two-phaseapproach for grouping mixed and numeric data, Affinity propagation clustering algorithm for mixed numeric and categorical datasets, Unsupervised learning with mixed numeric and nominal data, Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation, Automated variable weighting in k-means type clustering, Coupled interdependent attribute analysis on mixed data, A more efficient k-prototypes clustering techniques for mixed data, Feature weighting in k-means clustering and Geometrical codification for clustering mixed categorical and numerical databases. We choose some specific criteria to compare between the 13 algorithms such as, the ability to predict the value of k, scalability and shape of cluster, the dataset type that the algorithm deal with, sensitivity to noise, the category that the algorithm belongs to. To predict the value of k for all algorithms are random except the UFL which uses visual assessment of tendency to predict the number of k. K-means is difficult to predict the value of k but still random. TMCM and SBAC are not scalable and Automated variable weighting in k-means type is not scalable for large dataset. Most of the clustering algorithms can deal with mixed dataset except K-means and K-modes.

II. CONCLUSION

In this paper we studied and compared 13 clustering algorithms. K-means algorithm is the first designed algorithm for this purpose. But it has few shortcomings such as: choosing a suitable value for K, does not work for mixed data. The other twelve clustering algorithms choose the value of K randomly and work for mixed data.

ACKNOWLEDGMENT

This work is supported by a grant from the UAE University (31T111- 07_9_SURE+2018).

REFERENCES

[1] C. Shah, A. Jivani, "Comparison of Data Mining Clustering Algorithms", Nirma University International Conference on Engineering (NUiCONE), pp. 1-4, Nov. 28-30, 2013.

[2] J. Han, M. Kamber, "Data Mining Concepts and Techniques", third edition, Morgan Kaufmann Publishers an imprint of Elsevier.

[3] S. Pande, S. Sambare, V. Thakre, "Data clustering using data mining techniques", International journal of advance research in computer and communication engineering Vol1. 1, Issue 8, October-2012

[4] M. Ester, H. Kriegel, J. Sander, X. Xu "A density-based algorithm for discovering clusters in large spatial databases with noise", Second International Conference on Knowledge Discovery and Data Mining (1996).

[5] O. Abu Abbas, "Comparison Between Data Clustering Algorithm", The International Arab Journal of Information Technology, Vol. 5, No. 3, pp. 320-325, 2008.

[6] L. Kukolich and R. Lippmann, "LNKnet User's Guide", MIT Lincoln Laboratory, 1999

[7] M. Eisen, "Cluster and Tree View Manual", Stanford University, 1998

[8] K. Balaji, K. Lavanya. "Clustering Algorithms for Mixed Datasets: A Review", International Journal of Pure and Applied Mathematics, Vol. 118, No. 7, pp. 547-556, 2018.

[9] J. MacQeen "Some methods for classification and analysis of multivariate observations", 5th Symposium on Mathematical Statistics and Probability, pp. 281-297, 1967.
[10] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Kluwer Academic Publishers, Vol. 2, No. 3, pp. 283-304, 1998.
[11] Z. Huang, N.K. Ng, H. Rong, Z. Li, "Automated variable weighting in k-means type clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 657-668, 2005.
[12] A. Ahmad, L. Dey, "A K-mean clustering algorithm for mixed numeric and categorical data", Data Knowledge Engineering, 63, 219-228, 2015.
[13] M. Wei, D. Wunsch, "Clustering Data of Mixed Categorical and Numerical Type with Unsupervised Feature Learning", IEEE Access, Vol. 3, pp. 1605-1613, 2015.
[14] S. Ming-Yi, J. Jar-Wen, L. Lien-Fu, "A Two-Step Method for Clustering Mixed Categorical and Numeric Data", Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 11-19, 2010.
[15] K. Zhang, X. Gu, "An Affinity Propagation Clustering Algorithm for Mixed Numeric and Categorical Datasets", Mathematical Problems in Engineering. Vol. 2014, No. 14, pp. 1-8, 2014.
[16] G. Biswas, C. Li, "Unsupervised Learning with Mixed Numeric and Nominal Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 4, pp. 672-690, 2002.
[17] D. Goodall, "A New Similarity Index Based on Probability", Biometrics, vol. 22, pp. 882-907, 1966.
[18] M. Wei, T. Chow, R. Chan, "Clustering Heterogeneous Data with k-Means by Mutual Information-Based Unsupervised Feature Transformation", entropy, Vol. 17, No. 3, pp. 1535-1548, 2015.
[19] J. Huang, M. Ng, H. Rong, Z. Li, "Automated Variable Weighting in k-Means Type Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 5, pp. 675-668, 2005.
[20] C. Wang, L. Cao, M. Wang, J. Li, W. Wei, Y. Ou, "Coupled nominal similarity in unsupervised learning", In CIKM, 973–978, 2011.
[21] C. Wang, Z. She, L. Cao, "Coupled attribute analysis on numerical data", IJCAI '13 Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 1736–1742, Beijing,China, August 03-09, 2013.
[22] T.Ji, C. Zhou, C. Ma, Z. Wang, "An improved k-prototypes clustering algorithm for mixed numeric and categorical data", Neurocomputing, Vol. 120, pp. 590-596, 2013.
[23] D. Modha, and W. Spangler, "Feature Weighting in k-Means Clustering", Kluwer Academic Publishers, Vol. 52, No. 3, pp. 217-237, 2003.
[24] F. Barcelo-Rico, J. Diez, "Geometrical codification for clustering mixed categorical and numerical databases", Journal of Intelligent Information Systems, Vol. 39. No. 1, pp. 167-185, 2011.
[25] M. Ankerst, M. Breunig, H. Kriegel, J. Sander, "OPTICS: Ordering points to identify the clustering structure", in proceedings of ACM SIGMOD Conference, 1999 pp. 49-60.

TABLE 1: Comparison characteristic of 13 clusteringalgorithms

	Predict the value of K	Scalability	Shape of Cluster	Dataset type	Sensitivity to Noise	Category
K-means	Difficult	Yes	Convex	Numerical only	Yes	Partition
K-modes	Random	Yes	Convex	Categorical	Yes	Partition
K-prototype	Random	Yes	Convex	Mixed	Yes	Partition
UFL	Use visual assessment of tendency	Yes	Arbitrary	Mixed	No	Hierarchical
TMCM	Random	No	Arbitrary	All	Yes	Hierarchical
AP	Random	Yes	Arbitrary	Mixed		
SBAC	Random	No	Arbitrary	Mixed features	No	Hierarchical
UFT-k-means	Random		Convex	Numerical and non-numerical	Yes	
Automated variable weighting in k-means type	Random	Not scalable for large dataset	Convex	Mixed	Weight small (No) Weight big (Yes)	Partition
Coupled interdependent attribute analysis on mixed data	Random	Yes	Any shape	Mixed		Partition
Improved k-prototypes	Random	Yes	Convex	Mixed numeric, categorical	No	Partition
Feature weighting in k-means	Random		Arbitrary	Mixed		Partition
Geometrical codification	Random		Polar, spherical	Mixed		Hierarchical