Datasets Size: Effect on Clustering Results

Adeleke Ajiboye[1], Ruzaini Abdullah Arshah[2], Hongwu Qin[3]

Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang

[1]{ajibraheem@live.com}
[2, 3] {ruzaini, qinhongwu@ump.edu.my}

Abstract. The recent advancement in the way we capture and store data pose a serious challenge for data analysis. This gives a wider acceptance to data mining, being an interdisciplinary field that implements algorithm on stored data with a view to discovering hidden knowledge. Most people that keep records, however, are yet to reap the benefits of this tool, this is due to the general notion that a large datasets is required to guarantee reliable results. However, this may not be applicable in all cases. In this paper, we proposed a research technique that implements descriptive algorithms on numeric datasets of varied sizes. We modeled each subset of our data using EM clustering algorithm; two different numbers of partitions (k) were estimated and used for each experiment. The clustering results were validated using external evaluation measure in order to determine their level of correctness. The approach unveils the implication of datasets size on the clusters formed and the impact of estimated number of partitions.

Keywords: Data mining, Algorithms, Datasets, Partitions and Clustering.

1 Introduction

Clustering is a descriptive data mining task that group objects into classes based on similarity features that exist in them. Clustering is an operation that is fundamental in the field of data mining [1]. The improvement on technology has giving rise to availability of massive data being captured daily; several data are also being retrieved or transferred through the internet. The internet enables message transfer in form of emails, voice mail and other form of communications. Further exploration on these massively stored data using data mining techniques has indeed, brings about information generation which is a necessity for efficient decision making.

Clustering is an unsupervised learning technique as there are no predefined classes that would show what kind of desirable relations should be valid among the data [14]. Its capability of doing natural grouping or partitioning makes it indispensable in data mining.

Clustering deals with finding structure in a collection of unlabeled data [5]. A good clustering technique is expected to produce an intra-class similarity that is very high and inter-class similarity that is very low. Clustering of datasets can be achieved through the use of several algorithms; however, each algorithm differs in their notion of what constitutes a cluster. There are groups with small distance among cluster of the same members, the dense areas within the data space and certain statistical distribution. Methods used in clustering can be categorized into partition-based [2,11], hierarchy-based [10, 8], other known methods are based on density, grid and fuzzy.

Apart from classification, clustering also have several other applications, these include: image processing, analysis of spatial object, pattern recognition, data summarization, fraud detection and general data reporting task.

In this paper, we proposed a research technique that implements descriptive algorithms on numeric datasets of varied sizes. This research work was carried out to determine if the size of datasets has any implication on clustering results.

The remainder of the paper is organized in this order: In the next section, we reviewed some existing clustering approaches; most of the reviewed works were done on numeric datasets. In section 3, we modeled our dataset using Expectation Maximization (EM) clustering algorithm and the clusters generated were evaluated for correctness. All the results were discussed in section 4 and we concluded in section 5 by summarizing our findings.

2 Existing Clustering Approaches

There have been several publications on mining of big datasets; only few cases of data mining using small datasets have been reported in literature. This research mainly focused on determining the effect of size of datasets on clustering results. In this paper, we briefly reviewed some closely related works. Clustering of data has been successfully implemented with several clustering algorithms most especially when datasets is of moderate size. Several validity measures have also been reported.

In [1], Genetic k-means algorithm was proposed for clustering of numeric and categorical datasets. The proposed work assumes a given pre-classified data and measures the 'overlap' between clustering achieved and the ground truth classification.

In [3], divide-and-conquer technique was presented for clustering mixed numeric and categorical data. The method used involves dividing the original mixed datasets to numeric and categorical, existing algorithm were implemented on each sub datasets and later combined.

The submission by Yang and Fong [6], does not support the use of small dataset for mining purposes, they said "sampling technique is not suitable any more, the full data will tell the truth". The study by Salperyck and Lemaire [4], however have a different opinion. In their study on Learning with few examples: An empirical study on leading classifiers. The duo maintained that in certain situations, learning do start with only few

dataset as active and incremental learning were cited as the two main learning problems where a learning machine could learn well with just few data.

In [7], a streaming algorithm that effectively cluster large data stream was proposed, the work also provide empirical evidence of the algorithm's performance on synthetic and real data streams.

A Scaling Expectation-Maximization (EM) Clustering framework was proposed in [9]. The framework admits varying degrees of data membership in multiple clusters. The approach was reported to have operated within the confines of a limited memory buffer and the framework was also extended to update multiple clustering models simultaneously.

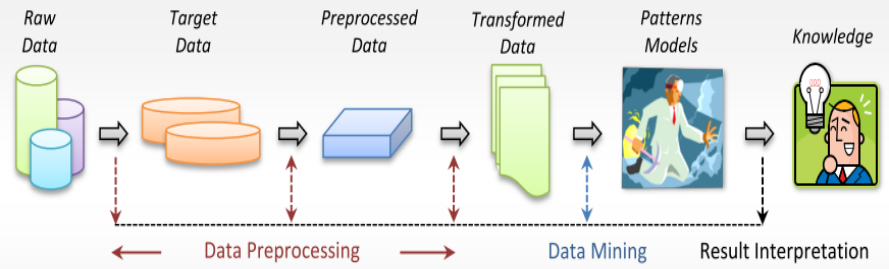


Fig. 1: Steps in a knowledge discovery process adopted from [12]

3 Segmentation of Datasets for EM Clustering

3.1 Datasets

The datasets we used for this experiment were sample of the original data collected from Joint Admission and Matriculation Board in Nigeria, West Africa. The Board is responsible for evaluating the suitability of all candidates seeking admission to tertiary institutions in that country. In the process of capturing data, the intention was not usually for mining purposes; we therefore, performed several pre-processing task prior to construction of the required models.

We divided the datasets to five subsets and we implemented EM clustering algorithm on each. Results from each model was carefully studied and further analyzed.

The implementation was done in RapidMiner environment. In the first experiment, five datasets of different sizes were clustered to 2, we clustered the same dataset to 5 in the second set of the experiment and their results were shown in Fig. 2. The purity of each cluster was computed using the formula:

$$\text{Purity}(D, C) = 1/N \sum_{K} \max_{j} |d_k \cap c_j|$$

where $D = \{d_1, d_2, \ldots, d_k\}$, the set of clusters and $C = \{c_1, c_2, \ldots, c_j\}$, the set of classes and N is the total number of data points. The purity value ranges between 0 and 1. A good cluster has a purity value close to 1, while a poor cluster has a purity value close to 0.

The EM clustering algorithm works as follows:

1. Initialize i to 0 and choose θ_i arbitrarily.
2. Compute $Q(\theta \mid \theta_i)$
3. Choose θ_i+1 to maximize $Q(\theta|\theta_i)$
4. If θ_i != θ_i+1, then set i to i+1 and return to Step 2

where step 2 usually referred to as the expectation step while Step 3 is called the maximization step[13] and θ is an unknown hidden variable.

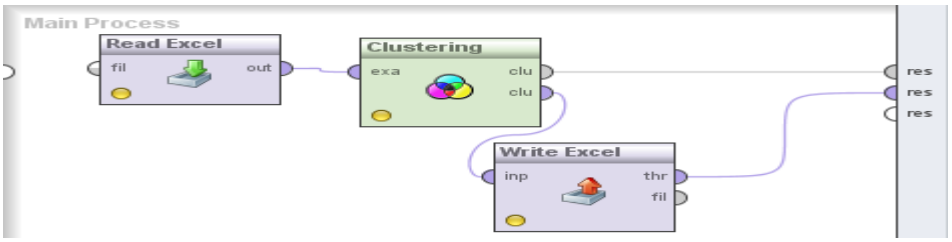


Fig. 2 : Design of EM clustering algorithm model

The following parameters were set:

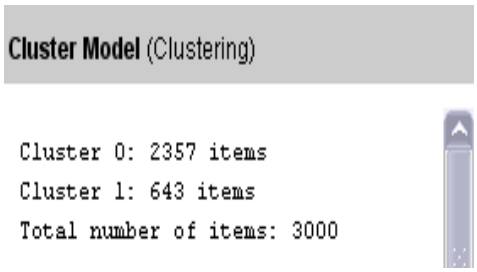
- **k**: The number of clusters: 2
- **max runs**: The maximal number of runs: 5
- **max optimization steps**: The maximal number of iterations performed for one run: 100
- **quality**: The quality that must be fulfilled before the algorithm stops: 1.0E-15-0.1
- **initial distribution**: Indicates the initial distribution of the centroids. default: k-means run

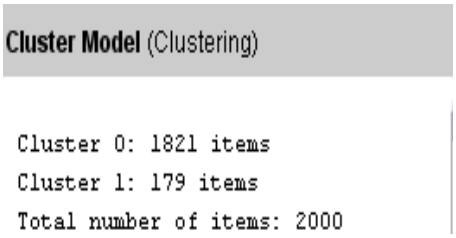
Fig. 3: Parameter settings of EM Clustering Algorithms

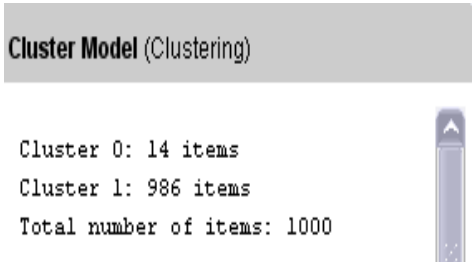
4 Experimental Results and Discussion

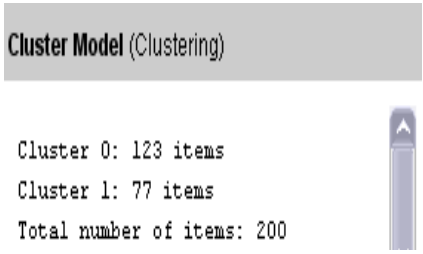
Due to the limited number of pages, the datasets used for the experiment could not be shown here; only the clustering results obtained from the experiment were displayed. The clustering results when k = 2 and k = 5 in each subset of dataset are shown in the following cluster models:

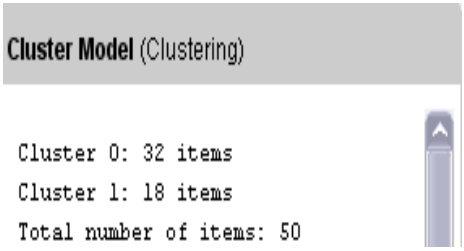
when k = 2 :











when k = 5 :

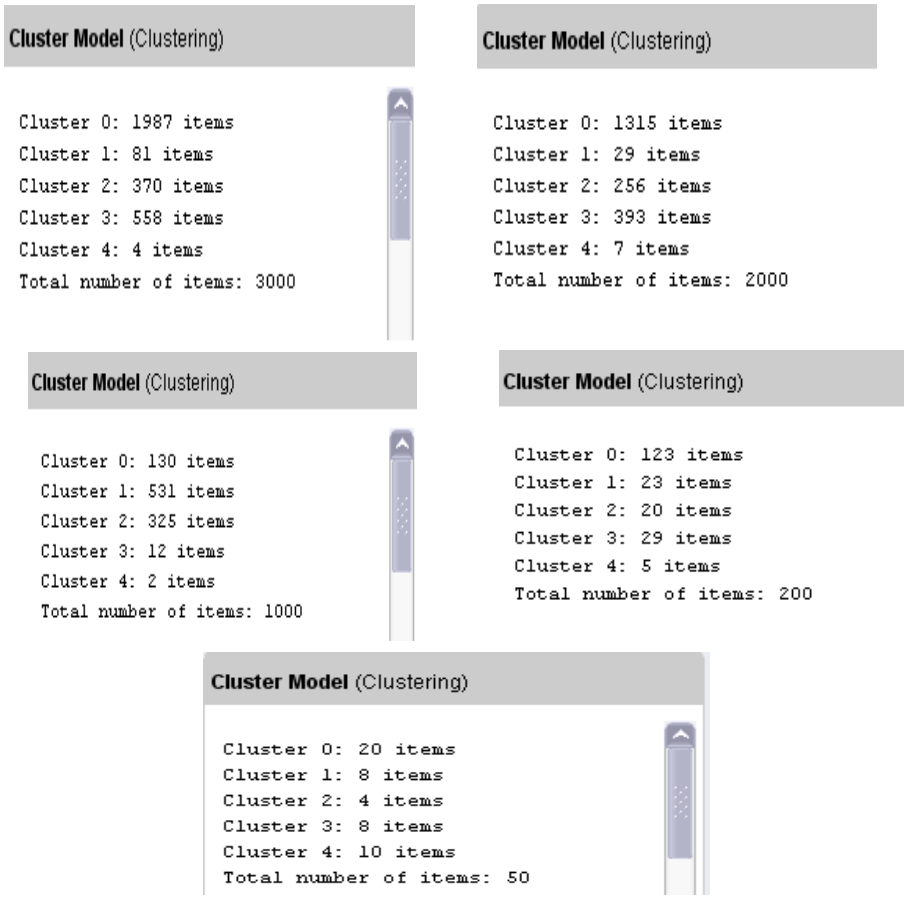


Table 1: Purity of formed clusters for k = 2 and k = 5

Dataset size	**Purity (p)**	
	# of clusters = 2	# of clusters = 5
3000	0.245	0.379
2000	0.206	0.413
1000	0.172	0.368
200	0.280	0.765
50	0.340	0.640

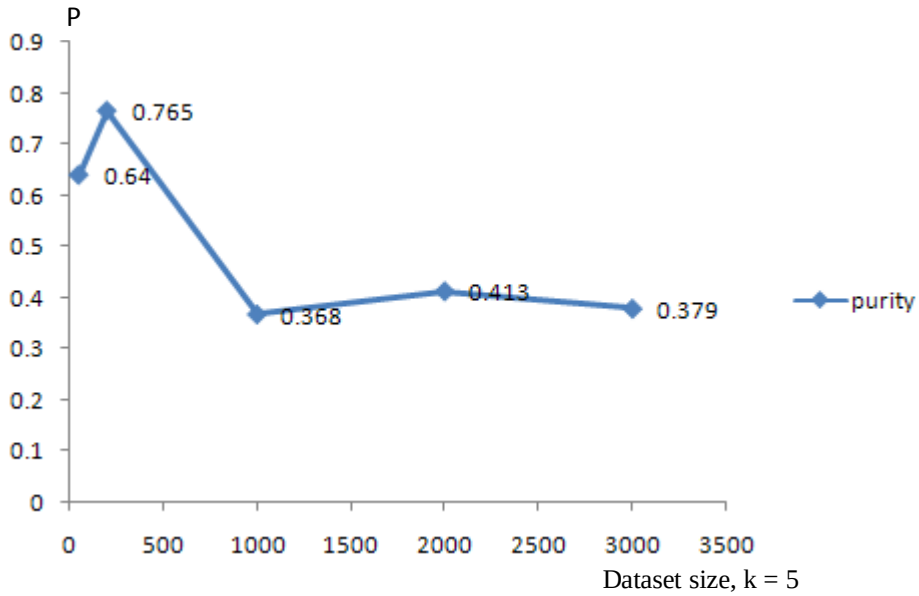

P
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.765
0.64
0.368
0.413
0.379
purity
0
500
1000
1500
2000
2500
3000
3500
Dataset size, k = 5

Fig. 4a

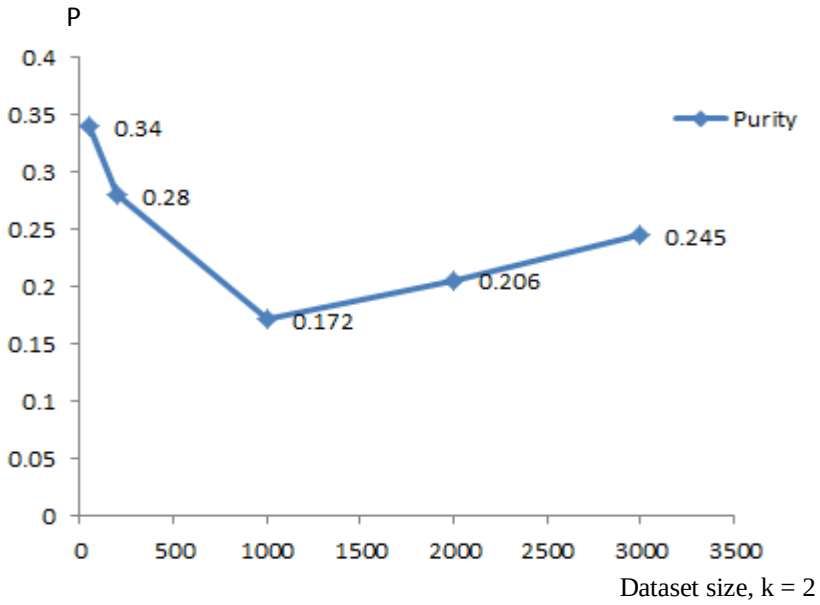

P
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.34
0.28
0.172
0.206
0.245
Purity
0
500
1000
1500
2000
2500
3000
3500
Dataset size, k = 2

Fig. 4b

Table 1, shows the summary results of purity of formed clusters. The datasets clustered were of sizes 3000, 2000, 1,000, 200 and 50. Clustering each of the dataset with a partition of 2, results to very poor clusters similarities, while an increase in the number of partition to 5, revealed some slight improvement. Dataset with 200 data points has the highest value of purity as shown in Fig. 4a. The graph descends due to an increase in dataset while the number of partition did not change. The composition of dataset with 2000 data points was responsible for a slight rise in value noticed in Fig. 4a and 4b, the purity value was generally low in Fig. 4b due to the choice of partition k that was set at 2, this value was not enough to group similar object that has very strong features together to form a cluster, especially in a data point of thousand.

5 Conclusion

This paper revealed the effect of size of datasets on clustering results. The process of clustering involves grouping physical or abstracts data (or objects) into classes of similar or dissimilar clusters. To determine how data are distributed within the data space, we conducted experiment on datasets of different sizes. We modeled our data with EM clustering algorithm and clusters were generated based on the value of k specified. We measure the level of correctness of the clusters formed using external validity measure, purity. Our experiment shows that, a number of factors do affect the clustering results. Among them is the choice of cluster partition, the nature of the dataset also play important determinants and size of the data to be clustered. Our results also revealed that, partitioning big dataset into 2 clusters would result into poor similarities among the data point in each cluster, most especially when the dataset is large, an increase in the number of cluster would lead to better results, however, the datasets has to be of reasonable size, but big dataset is not a guarantee to having a good clustering results.

References

1. Roy, D.K. and Sharma, L.K.: Genetic k-means clustering algorithm for mixed numeric and categorical datasets; International Journal of Artificial intelligence & Applications, vol. 1, No. 2. (2010)
2. Hung, L.K., Yang, D.L.: An efficient fuzzy C-means clustering algorithm, In Proceedings of the IEEEInternational Conference on Data Mining , pp. 225–232. (2001)
3. Murala, D.K.: Divide-and-conquer technique for clustering mixed numeric and categorical data, International Journal of Computer Science and Information Technologies, vol. 4, No.1, (2013)

4. Salperyck, C. and Lemaire, V.: Learning with few examples: An empirical study on leading classifiers, International Joint Conference on Neural Network, 2011.
5. M.D. Boomija: Comparison of a Partition Based Clustering Algorithms; Journal of Computer Applications, Vol. 1, No. 4 (2008)
6. Yang, H. and S. Fong: Improving Adaptability of Decision Tree for Mining Big Dataset, New Mathematics and Natural Computation vol. 9(1), pp77-95, 2013.
7. Guha, S. et al.: Clustering Data Stream: Theory and Practice; IEEE on knowledge and Data Engineering, http://www.computer.org/csdl/trans/tk/2003/03/k0515-abs.html (2003)
8. Guha, S. Rastogi, and Shim, K.: CURE: An efficient clustering algorithm for large databases, In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 73–84. (1998)
9. Bradley, P.S., Fayyad, U., Reina, C.: Scaling EM (Expectation-Maximization) Clustering to Large Databases, (1999) http://ftp.research.microsoft.com/pub/tr/originals/tr-98-35.pdf
10. Estivill-Castro, V., Lee, I: Hierarchical clustering based on spatial proximity using delaunay diagram, In Proceedings of the 9th International Symposium on Spatial Data Handling, 7a.26–7a.41. .(2000)
11. Ng, R. T., Han, J.,: Efficient and effective clustering methods for spatial data mining, In Proceedings of the 20th VLDB Conference, pp.144–155. (1994)
12. Knowledge discovery in database, International Hellenic University, http://vlabs.ihu.edu.gr/fileadmin/labsfiles/decision_support_systems/lessons/ClTrees_AssocRules/ClTrees_AssocRules-IHU.pdf
13. Expectation Maximization: http://cs.brown.edu/research/ai/dynamics/tutorial /Documents/
14. Halkidi, M., Y. Batistakis and M. Vazirgiannis: Clustering Validity Checking Methods: Part II, ACM Sigmod Record, (2002).