

Efficiency of Random Sampling Based Data Size Reduction on Computing Time and Validity of Clustering in Data Mining

Zeynel Cebeci¹, Figen Yildiz²

INFO

Received 19 Nov. 2015

Accepted 6 Mar. 2016

Available on-line 30 Apr. 2016

Responsible Editor: M. Herdon

Keywords:

data reduction, random sampling, cluster analysis, external validity indices, big data, k-means clustering.

ABSTRACT

In data mining, cluster analysis is one of the widely used analytics to discover existing groups in datasets. However, the traditional clustering algorithms become insufficient for the analysis of big data which have been formed with the enormous increase in the amount of collected data in recent years. Therefore, the scalability has been one of the most intensively studied research topics for clustering big data. The parallel clustering algorithms and the Map-Reduce framework based techniques on multiple machines are getting popular in scalability for big data analysis. However, applying the sampling techniques on big datasets could be still alternative or complementary task in order to run the traditional algorithms on single machines. The results obtained in this study showed that the data size reduction by the simple random sampling could be successfully used in cluster analysis for large datasets. The clustering validities by running K-means algorithm on the sample datasets were found as high as those of the complete datasets. Additionally the required execution time for cluster analysis on the sample datasets was significantly shorter than those obtained for the complete datasets.

1. Introduction

Although there are dozens of definitions, big data is a broad term that defines the very large datasets that are complex, diverse and massive to manage and analyze with the traditional data analysis tools and methods. With the wide spread use of the internet, the new developments in the information technologies and the recently enormous growth in the use of social networks in addition to the increased applications of the internet of things, the data has rapidly become very big. Big data is now so big, and continues to bigger day by day. For instance, the size of spatial and temporal data recorded for a thousand acres of field can reach to several hundred terabytes in some of agricultural practices like precision farming. In data mining, the need for new data storage architectures, data processing techniques and data analysis algorithms has led to the new challenges in knowledge discovery from big data.

However, sometimes it is used as the synonym of Knowledge Discovery in Databases (KDD), in fact, Data Mining (DM) is a suit of high-end analyses in KDD involving a sequentially applied group of the data processing stages such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. In this context, DM is an emerging multidisciplinary field of study borrowing the methods and techniques from many research areas such as database and data warehouse architecture, high performance computing, machine learning, artificial intelligence, information retrieval, data visualization, mathematics, statistics and a series of related fields. DM is therefore actually a fusion of the techniques and methods which has been developed to discover hidden knowledge such as the meaningful patterns, relationships, changes, trends and useful structures generally in large datasets which are stored in databases, data warehouses and other kind of repositories. These DM methods and techniques can be usually classified in three groups as descriptive, predictive and prescriptive analytics.

Cluster analysis, one of the descriptive analytics in DM, is used to reveal priori unknown cluster tendency (or grouping structure) in datasets. Although there are many clustering algorithms, the

¹ Zeynel Cebeci

Div. of Biometry & Genetics, Faculty of Agriculture, Çukurova University, 01330 Adana - Turkey
zcebeci@cukurova.edu.tr, cebeciz@gmail.com

² Figen Yildiz

Div. of Biometry & Genetics, Faculty of Agriculture, Çukurova University, 01330 Adana - Turkey
yildizf@cukurova.edu.tr

majority of them have some disadvantages in analyzing large datasets because the system resources may be insufficient to execute them in the memory of a single computer, or to finish analysis within a short time on a single computer (Jain & Verma 2014, Kothari & Kale 2014, Shirkorshidi *et al* 2014, Zerhari *et al* 2015). Some of the datasets can be so big that they cannot be loaded to a single computer memory as a whole set, and hence they are also called “unloadable data” (Hathaway & Bezdek 2006). Since the volume of unloadable data beyond the ability of analyzing by traditional tools and algorithms within a reasonable execution time, the scalability is required to perform clustering algorithms on high volume and high dimensional datasets stored in databases and data warehouses. The scalability approaches which have been proposed for big data clustering could be gathered in two groups as follows (Shirkorshidi *et al* 2014):

- Single machine techniques
 - o Sampling based techniques
 - o Dimension reduction techniques
- Multiple machine techniques
 - o Parallel clustering techniques
 - o Map-Reduce framework based techniques

While the single machine techniques are performed on a single machine, the multiple machine techniques are executed on more than one machine. Although the majority of clustering algorithms have been developed to run on a single machine, today they are being improved in scalability with parallel computing extensions to analyze the large datasets on multiple machines. In general, the parallel computing is a common choice to discover clusters on large data sets (Nagesh *et al* 2000). For instance, various parallel K-means algorithms (Huang 1998, Stoffel & Belkoniene 1999, Dhillon & Modha 2000), and the parallel Fuzzy C-means algorithm (Kwok *et al* 2002) and the fuzzy extension called as the Generalized Extensible Fast Fuzzy C-means (geFFCM) algorithm by Hathaway *et al* (2006) have been proposed to perform partitioning of big data. Beyond these extensions, the clustering algorithms in Map-Reduce framework are based on task distribution among multiple machines. In the Map-Reduce framework, a clustering task is firstly divided into small tasks and assigned to the distributed systems in the mapping stage, and then the obtained results are merged in reduction phase. K-means, Expectation Maximization (EM) and DBCURE algorithms are frequently being used in this kind of scalability approaches.

Some algorithms have been proposed and their efficiencies with parallel computing have been proven, yet they are not widely in use. The use of them by individual researchers is not easy task due to their complexity, difficulties in task organization and the required resources to run on multiple machines. Therefore, there is still a strong need for the easy-to-use approaches to cluster big data which are unloadable into memory but storable on single machines.

In DM, data reduction is mainly done for the dimension reduction purposes. In order to lower computing costs, the number of dimensions (features) in data is decreased with some well-known statistical methods such as multidimensional scaling, principal component analysis, factor analysis and the others (Napoleon & Pavalakodi 2011). Another purpose of data reduction is to make unloadable data loadable. In this way, it is expected that the use of smaller datasets drawn from unloadable datasets by some sampling techniques may shorten the computing time for clustering.

Although the research works in data reduction have mostly been focused on high dimensionality issues in order to select the most important features to be analyzed, there are also a few research works on data size reduction by sampling techniques. The scalability for the clustering algorithms on big data can be applied by sampling and some other reduction techniques (Hathaway *et al* 2006). Shirkorshidi *et al* (2014) evaluated that the sampling takes at the first place of progress of the developments in clustering algorithms to deal with big data. Zaki *et al* (1997) revealed that sampling reduced the computing costs in the mining of association rules. Cormode & Duffield (2014) stated that the use of sampling motivates to generate summary datasets that support rapid queries, and prolong the useful life of the data in storage. Palguna *et al* (2015) studied on some sampling formulations for Twitter’s big data, and found that “random sampling can be used as a first step before using other sophisticated algorithms to form human readable summaries or for mining social opinion”. Tong & Kang (2013) reported that the randomized techniques reduce the complexity of the input data size to speed up/scale up a data clustering algorithm.

Earlier studies suggest that running the clustering algorithms on smaller datasets by the random sampling may remarkably shorten the required execution time of clustering task for large datasets. For that reason, in this study, we aimed to empirically determine the effect of the random sampling on computing time and validity of clustering for data mining purposes. According to our goals, we generated the artificial datasets of different cluster structures, and analyzed them by using the K-means algorithm in different sampling rates.

2. Simple random sampling for data size reduction

Sampling is to draw n samples representing characteristics of an entire population with N size. The aim of sampling is to describe or to make inferences about a larger population with the statistics computed from a smaller sample. In general, however there are the probability and the nonprobability based sampling methods; the most commonly used sampling methods are simple random sampling, systematic sampling, stratified sampling and cluster sampling.

The simple random sampling (SRS) is the most commonly applied method in which each unit in a population has the same chance ($1/N$) of being drawn. SRS can be applied with replacement or without replacement of drawn samples. In SRS without replacement, when an object is drawn it cannot be sampled once again because it is not given back to the population. Because of its simplicity and the proven efficiency in traditional applications we used SRS without replacement method for sampling datasets in this study.

3. External validity indices and clustering quality

The success of sampling can be determined with the clustering similarity of smaller sample datasets and large complete dataset from which the samples are drawn. Thus, clustering results from sample datasets are compared to clustering results on complete datasets by using the cluster validity indices. The indices measuring clustering validity can be external, internal and relative depending on the information available prior to the clustering task (Dudoit & Fridlyand 2002, Kovács *et al* 2005, Thalamuthu *et al* 2005, Rendón *et al* 2011). If the cluster labels of objects in a dataset are already known they can be used as external information. So the external indices compare the results of a clustering to the cluster labels of a reference clustering. The internal indices do not use any external information because they measure clustering validity by using some metrics obtained in clustering. The relative indices are based on comparison of the clustering results from successive runs of a clustering algorithm on the same dataset by changing the input parameters in each run.

Since the data is mostly unlabeled in clustering, the internal indices are commonly used to determine the validity of a partitioning task. On the other hand, most of the internal indices might be sensitive to the density of clusters in datasets, and have a tendency to ignore small clusters and clusters with low density (Žalik & Žalik 2011). According to Desgraupes (2013) “the external indices of comparison are indices designed to measure the similitude between two partitions. They take into account only the distribution of the points in different clusters and do not allow to measure the quality of this distribution”. Hence, for our research purpose the external indices could be more robust in comparison of clustering concordance between sample and complete datasets. In their research aiming to carry out the effect of sampling, de Souto *et al* (2012) also preferred to use the external validity indices for assessing the partitions for highly imbalanced datasets. In our study, since we expect that the cluster densities can be changed by the sampling rates we also assumed that the external indices would be more informative in comparison of the partitions obtained on different sample datasets.

For a given dataset X with n objects ($X = \{x_1, x_2, \dots, x_n\}$), let $C1$ be any partitioning of X obtained as a result of a clustering algorithm ($C1 = \{C1_1, C1_2, \dots, C1_k\}$), and $C2$ be a reference partitioning of X built with the use of an external information, namely class labels ($C2 = \{C2_1, C2_2, \dots, C2_k\}$). In order to check the agreement of these clustering, $C1$ could be compared to $C2$ as the “ground truth” or “actual class structure” by using the external indices. The external indices measuring the magnitudes of this agreement are calculated from contingency tables which contain the frequencies of objects in the same and different clusters in $C1$ and $C2$.

Table 1. Structure of a contingency table

| | Clustering on the reference dataset (C1) | Clustering on sample dataset (C2) |
|--|--|-----------------------------------|
| Clustering on the reference dataset (C1) | <i>a</i> | <i>b</i> |
| Clustering on sample dataset (C2) | <i>c</i> | <i>d</i> |

As shown in Table 1, a contingency table has the four elements, namely *a*, *b*, *c* and *d* which are the frequencies of objects for the following situations:

- *a*: Number of objects in same clusters in both C1 and C2,
- *b*: Number of objects in same clusters in C1 but in different clusters in C2,
- *c*: Number of objects in different clusters in C1 but in same clusters in C2,
- *d*: Number of objects in different clusters in both C1 and C2.

Although there are many more, the external validity indices used in this study are listed in Table 2. They can be computed by using the above described elements of a contingency table with the formulas given in Table 2.

Table 2. The formulas of some well-known external indices

| Indices | Formula |
|-----------------------|--|
| Jaccard (J) | $\frac{a}{a+b+c}$ |
| Czekanowski-Dice (CD) | $\frac{2a}{2a+b+c}$ |
| Fowlkes-Mallows (FM) | $\frac{a}{\sqrt{(a+b)(a+c)}}$ |
| Kulczynski (K) | $\frac{1}{2} \left(\frac{a}{a+b} + \frac{a}{a+c} \right)$ |
| Rand (R) | $\frac{a+d}{a+b+c+d}$ |
| Sokal-Sneath 1 (SS1) | $\frac{a}{a+2(b+c)}$ |

The values of indices are computed in the range [0, 1] for the most of external validity indices. An index value becomes higher while it approaches to 1. Whereas the quality of clustering decreases while it approaches to 0. In other words, if an index value is 1 it means that there is an excellent agreement between ground truth and clustering results, but a value of 0 means no similarity at all. Thus the magnitude of an external index value indicates the quality of different cluster analysis on different datasets sampled from a complete dataset. Although there is no a commonly agreed scheme for describing the strength of index values, a classification scheme as shown in Table 3 may be helpful to describe it.

Table 3. Quality classification of clustering validity according to index values

| Index value | Quality of clustering |
|-------------|-----------------------|
| 1.00 | excellent |
| 0.90 - 0.99 | very high |
| 0.70 - 0.89 | high |
| 0.50 - 0.69 | moderate |
| 0.30 - 0.49 | low |
| 0.01 - 0.29 | very low |
| 0.00 | none |

3. Datasets and tools for analysis

In order to determine the efficiency of sampling based on the quality of clustering, six complete artificial datasets were used in this study. The characteristics of these datasets are listed in Table 4, and their scatter plots are given in Figure 1. Because the arbitrary shapes of clusters in large datasets may differently respond to the sampling, we generated the rectangular, circular and elliptical shaped

clusters in the first 3 datasets in order to see whether the shape effect is important or not. In order to observe the performance of clustering for noisy datasets we also generated the last three datasets containing the varying degree of noises for only circular shaped clusters. All the datasets were generated with 2 dimensions (2 features) by an R script using the functions from the packages *stats*, *ellipse* (Murdoch & Chow 2015), *splancs* (Bivand *et al* 2015), and *mlbench* (Leisch & Dimitriadou 2015) in R environment (R Core Team, 2015). In Table 4, k , N and n_k stand for the number of clusters in labeled complete dataset, the size of dataset and the size of clusters in each dataset, respectively. The number of clusters, k was taken as 9 in order to create a centered cluster in all the datasets. For using as the ground truth, the data points in each cluster were labeled with their cluster numbers.

Table 4. Characteristics of the experimental datasets

| Datasets | k | N | n_k | Shape of clusters | Scattering pattern of clusters | Separation status of clusters |
|----------|-----|------|-------|-------------------|--------------------------------|-------------------------------|
| DS1 | 9 | 1800 | 200 | rectangle | tiled | well separated |
| DS2 | 9 | 1800 | 200 | ellipse | tiled | well separated |
| DS3 | 9 | 1800 | 200 | circle | tiled | well separated |
| DS4 | 9 | 1800 | 200 | circle | circular | well separated |
| DS5 | 9 | 1800 | 200 | circle | circular, noisy | not well separated |
| DS6 | 9 | 1800 | 200 | circle | circular, too noisy | not well separated |

The simple random sampling without replacement (*srswor*) method in the *sampling* package by Tillé and Matei (2013) in R environment was applied to form the sample datasets from the complete datasets. DS_s -1%, DS_s -5%, DS_s -10%, DS_s -25%, DS_s -50% and DS_s -75% datasets (s denotes the subscripts of datasets) were generated from the relevant complete datasets (DS1-T to DS6-T) with varying sampling rates of 1.66%, 5%, 10%, 25%, 50% and 75% respectively. With the sampling rate of 1.66% we aimed to generate the sample datasets as small as of 30 samples. This is the smallest size which is considered as reliable for many statistical methods.

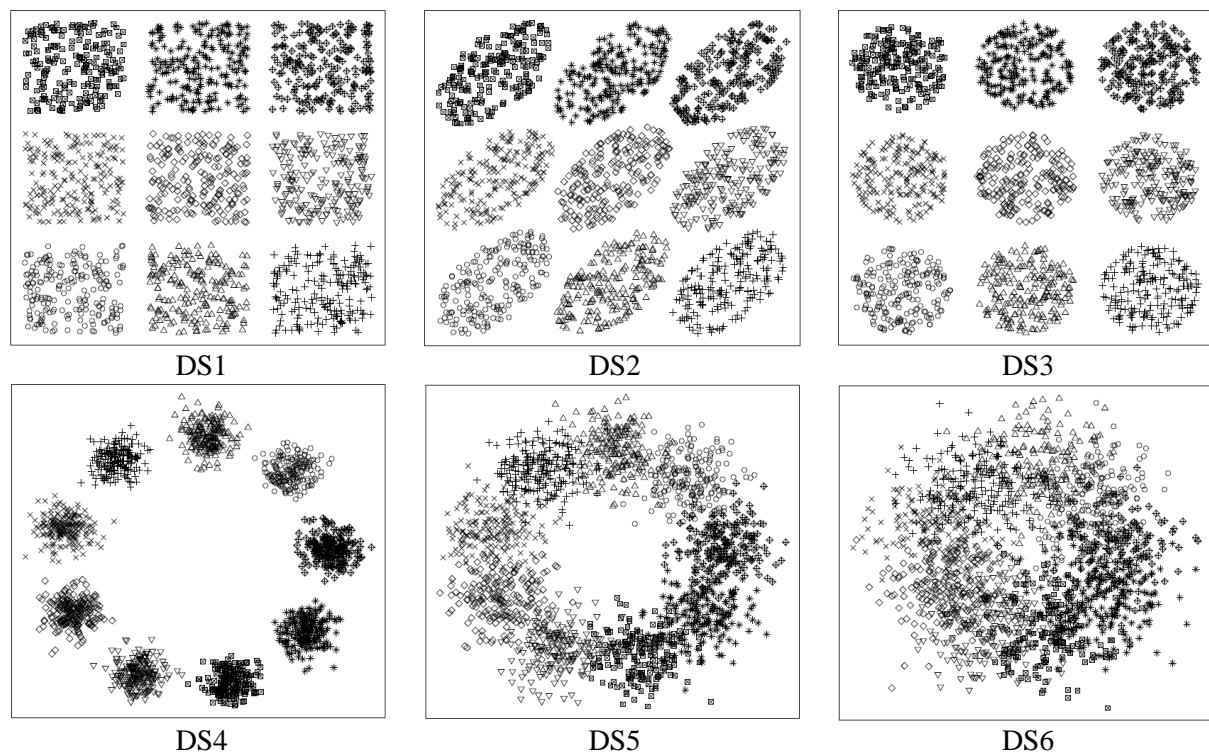


Figure 1. Scatter plots of the artificial datasets (DSs-T)

After sampling completed, the sample datasets were plotted for inspection of probable anomalies about the densities and distributions. An example of these scatter plots for the sample datasets drawn from DS3 was given in Figure 2.

In this study, K-means as one of the most common clustering algorithms was used by running `kmeans` function of the `stats` package of R. It was executed with the input parameters as of `nstart=10` and `iter.max=100`. For each sample dataset and related complete dataset, the k parameter, number of clusters were set to the number of reference clusters which are listed in Table 4 (`centers=k`).

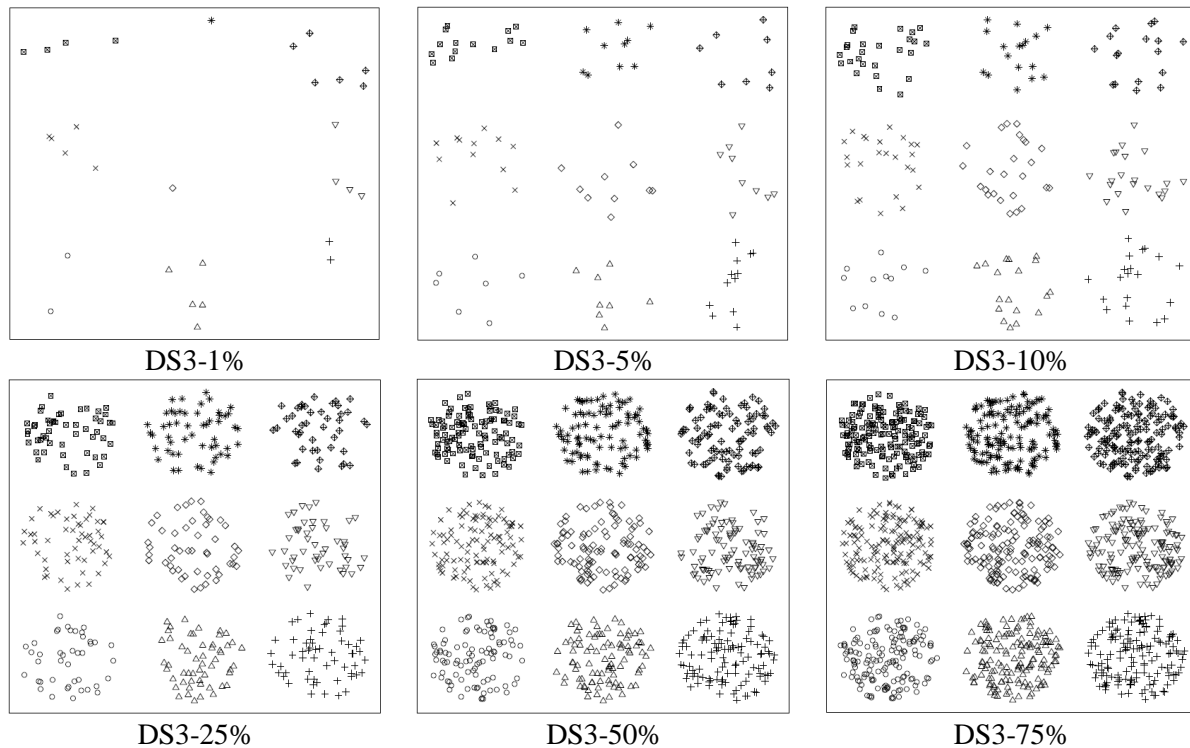


Figure 2. Scatter plots of the sample datasets sampled from DS3-T dataset.

The external indices were computed by using the results of K-means clustering from the `extCriteria` function of `clusterCrit` package developed by Desgraupes (2013). The execution time for each K-means run was computed `Sys.time` function as follows:

```
ctime <- Sys.time()
kmeans(dataset, centers=k, iter.max=100, nstart=10)
ctime <- round(Sys.time() - ctime, 12)
```

The analysis was performed on a PC with i7 microprocessor, 4 MB RAM and 1TB HDD running 64-bit Windows 10 operating system, and R was only active application during execution of R script.

5. Results and discussion

As seen in Table 5, DS1 having the rectangular shaped and tiled spreading clusters, the result of clustering with 1.66% sampling rate was very highly similar to the actual cluster structure (ground truth) according to the similarity scheme given in Table 3. Although SS1 index with the value of 0.73 was the lowest when compared to the values of other indices, it was still high enough. For DS1, the clustering similarities between the sample datasets and the reference datasets were the same according to all the indices in 5% and other higher sampling rates.

In DS2 dataset having the elliptical clusters with tiled spreading pattern there was a small amount of decrease in the values of all the indices when compared to those of DS1. However, according to all the external indices, the clustering results of sample datasets and reference

dataset were found to be similar with the index values over 0.60 except SS1 index values for DS2-5% dataset.

Table 5. The values of external indices for different datasets*

| Datasets | J | CD | FM | K | R | SS1 |
|----------|------|------|------|------|------|------|
| DS1-1% | 0.84 | 0.91 | 0.92 | 0.92 | 0.98 | 0.73 |
| DS1-5% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS1-10% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS1-25% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS1-50% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS1-75% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS1-T | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | | | | |
| DS2-1% | 0.61 | 0.76 | 0.76 | 0.77 | 0.95 | 0.44 |
| DS2-5% | 0.60 | 0.75 | 0.75 | 0.75 | 0.94 | 0.43 |
| DS2-10% | 0.96 | 0.98 | 0.98 | 0.98 | 1.00 | 0.93 |
| DS2-25% | 0.95 | 0.97 | 0.97 | 0.97 | 0.99 | 0.90 |
| DS2-50% | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 |
| DS2-75% | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 |
| DS2-T | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | | | | |
| DS3-1% | 0.69 | 0.82 | 0.82 | 0.82 | 0.96 | 0.53 |
| DS3-5% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS3-10% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS3-25% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS3-50% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS3-75% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS3-T | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | | | | |
| DS4-1% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS4-5% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS4-10% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS4-25% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS4-50% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS4-75% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DS4-T | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | | | | |
| DS5-1% | 0.61 | 0.76 | 0.76 | 0.76 | 0.95 | 0.44 |
| DS5-5% | 0.75 | 0.86 | 0.86 | 0.86 | 0.97 | 0.60 |
| DS5-10% | 0.77 | 0.87 | 0.87 | 0.87 | 0.97 | 0.63 |
| DS5-25% | 0.82 | 0.90 | 0.90 | 0.90 | 0.98 | 0.70 |
| DS5-50% | 0.78 | 0.88 | 0.88 | 0.88 | 0.97 | 0.64 |
| DS5-75% | 0.77 | 0.87 | 0.87 | 0.87 | 0.97 | 0.63 |
| DS5-T | 0.76 | 0.87 | 0.87 | 0.87 | 0.97 | 0.62 |
| | | | | | | |
| DS6-1% | 0.20 | 0.34 | 0.34 | 0.34 | 0.86 | 0.11 |
| DS6-5% | 0.28 | 0.44 | 0.44 | 0.44 | 0.88 | 0.17 |
| DS6-10% | 0.27 | 0.43 | 0.43 | 0.43 | 0.87 | 0.16 |
| DS6-25% | 0.28 | 0.43 | 0.43 | 0.43 | 0.87 | 0.16 |
| DS6-50% | 0.29 | 0.45 | 0.45 | 0.45 | 0.88 | 0.17 |
| DS6-75% | 0.28 | 0.44 | 0.44 | 0.44 | 0.87 | 0.16 |
| DS6-T | 0.27 | 0.43 | 0.43 | 0.43 | 0.87 | 0.16 |

(*) : The abbreviations of external indices are the same as in Table 2.

In DS3, however the clustering similarity of the reference dataset and DS3-1% was moderate according to J and SS1 indices, yet it was high for the remaining indices. The values of clustering validity indices obtained with 5% and higher sampling rates were found to be the same as 1 according to all the indices.

It was interesting that the clustering results for all the sample datasets drawn from DS4 having the circular shaped and circularly spreading clusters were fully similar with each other with an index value of 1 for all the indices.

Although DS5 dataset had the circular shaped clusters like DS4, the values of all the indices were relatively smaller when compared to those of DS4. This result was due to the noisy clusters in DS5 and the sample datasets drawn from it. However, even in DS5-1% sample dataset which was smallest among the sample datasets of DS5, the values of the indices except J and SS1 were high with the values over 0.76. Another interesting result for DS5 was that the index values increased until 25% sampling rate but slightly started to decrease in 50% and in higher sampling rates. This behavior was probably due to the increase in separation of the clusters because of noise reduction in the lower sampling rates. This finding indicates that sampling on large datasets may give more compact clusters if they are noisy and partly overlapped with each other.

In DS6, the values of indices were low for all the indices except R index. There were no apparent differences between the clustering validities for different sampled datasets and the complete dataset. This means that if a complete dataset is too noisy the reduced datasets with lower sampling rates could be recommended in favour of decrease in computing time. As a result we could say that “a small number of samples does not necessarily result in a poor classification, nor does a relatively large number of samples guarantee the best classification” as stated by Wharton (1987).

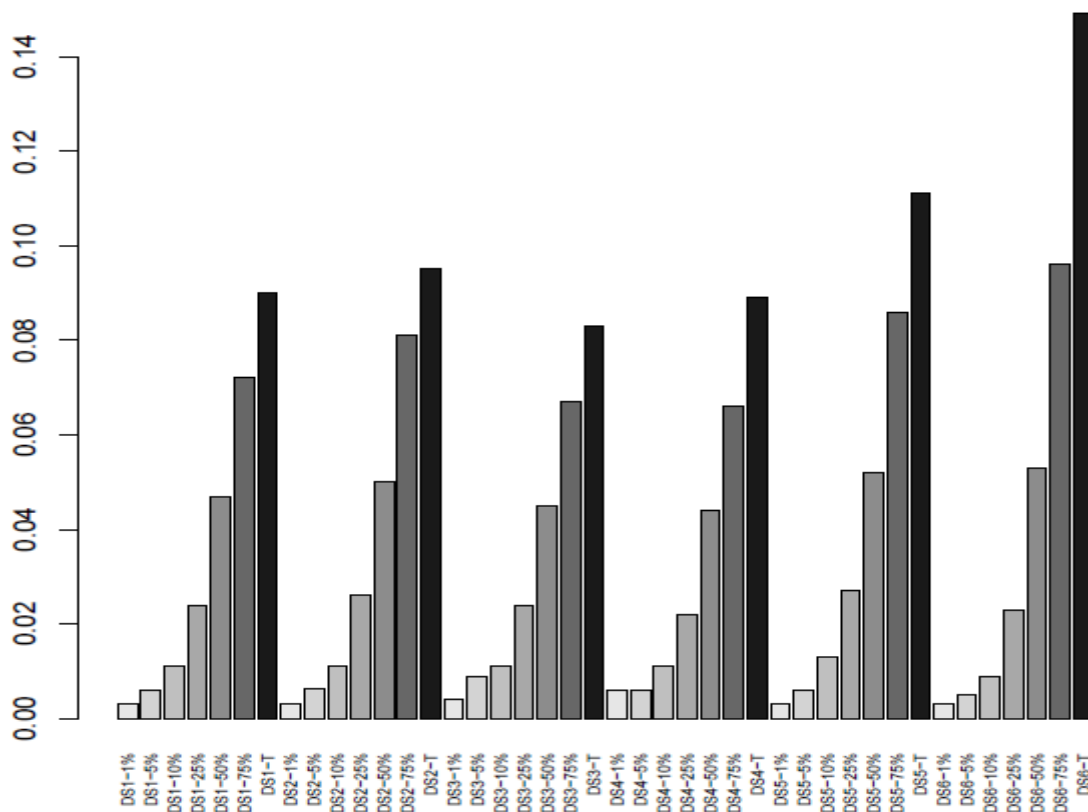


Figure 3. Computing time by the datasets (seconds)

Concerning the computing time required for clustering, there was no any significant difference between the datasets ($p > 0.05$). On the other hand, as clearly seen from Figure 3 the sampling rate increases more the computing time linearly increases. A positive and very strong correlation was obtained between the computing time and the sampling rates for all the analyzed datasets. The magnitude of this Pearson's correlation was significantly high as 0.97, and the relationship was

significantly linear with the model of $computing\ time = 0.000795 + 0.101263 * sampling\ rate$ ($t = 23.433, p < 2e-16$).

Table 6. The changes in computing time by the datasets

| Datasets | Computing time (sec) | Change-1 (%) | Change-2 (%) |
|----------|----------------------|--------------|--------------|
| DS1-1% | 0.003001 | - | 2901 |
| DS1-5% | 0.006007 | 100 | 1399 |
| DS1-10% | 0.011008 | 83 | 718 |
| DS1-25% | 0.024016 | 118 | 275 |
| DS1-50% | 0.047031 | 96 | 92 |
| DS1-75% | 0.072054 | 53 | 25 |
| DS1-T | 0.090065 | 25 | - |
| DS2-1% | 0.003002 | - | 3067 |
| DS2-5% | 0.006364 | 112 | 1394 |
| DS2-10% | 0.011007 | 73 | 764 |
| DS2-25% | 0.026017 | 136 | 265 |
| DS2-50% | 0.050033 | 92 | 90 |
| DS2-75% | 0.081053 | 62 | 17 |
| DS2-T | 0.095068 | 17 | - |
| DS3-1% | 0.004003 | - | 1975 |
| DS3-5% | 0.009011 | 125 | 822 |
| DS3-10% | 0.011010 | 22 | 654 |
| DS3-25% | 0.024019 | 118 | 246 |
| DS3-50% | 0.045029 | 87 | 84 |
| DS3-75% | 0.067046 | 49 | 24 |
| DS3-T | 0.083058 | 24 | - |
| DS4-1% | 0.006004 | - | 1383 |
| DS4-5% | 0.006005 | 0 | 1383 |
| DS4-10% | 0.011012 | 83 | 709 |
| DS4-25% | 0.022015 | 100 | 305 |
| DS4-50% | 0.044031 | 100 | 102 |
| DS4-75% | 0.066048 | 50 | 35 |
| DS4-T | 0.089054 | 35 | - |
| DS5-1% | 0.003002 | - | 3600 |
| DS5-5% | 0.006007 | 100 | 1749 |
| DS5-10% | 0.013014 | 117 | 754 |
| DS5-25% | 0.027024 | 108 | 311 |
| DS5-50% | 0.052040 | 93 | 113 |
| DS5-75% | 0.086061 | 65 | 29 |
| DS5-T | 0.111079 | 29 | - |
| DS6-1% | 0.003002 | - | 4867 |
| DS6-5% | 0.005003 | 67 | 2880 |
| DS6-10% | 0.009006 | 80 | 1556 |
| DS6-25% | 0.023016 | 156 | 548 |
| DS6-50% | 0.053035 | 130 | 181 |
| DS6-75% | 0.096069 | 81 | 55 |
| DS6-T | 0.149103 | 55 | - |

The second column (Change-1) of Table 6 shows the increase in computing time for a dataset from the preceding adjacent dataset. There was no increase (0%) from DS4-1% to DS4-5% while the highest increase (125%) was calculated from DS3-1% to DS3-5%. On the other hand, the change varied between 17% and 55% between the sample datasets formed with 50% and 75% sampling rate. This finding pointed out that the changes in computing time between contiguous datasets were significantly different ($t=-2.869$, $p=0.00703$).

Using the computing time for a complete dataset as the base, the decrease as the percentages in the computing time between the sample datasets and relevant complete datasets are listed in the third column (Change-2) of Table 6. The computing time required with the use of sampled datasets was significantly less for the sample datasets built in lower sampling rates ($t=-8.419$; $p=7.87e-10$). For instance, the computing time required for the sample datasets with 1.66% sampling rate were as less as 1383% and 4867% when compared to the time required for the relevant complete datasets. On the other hand, the decrease in computing time between the sample datasets with 5% sampling rate and the related complete datasets were changed between 822% and 2880%.

6. Conclusions

Sampling for data reduction is an important research question in DM. This study revealed that the clustering will be of high quality for the sample datasets which are generated by using simple random sampling. The values of external clustering validity indices computed on the sample datasets and the relevant complete datasets were found to be highly similar in lower sampling rates and almost same in higher sampling rates. This finding showed that clustering on sample datasets handled with data size reduction by sample random sampling can provide information as much as those gained from complete datasets. So, when sampling applied to large datasets, clustering tasks would be performed faster with no or a negligible loss of information about data structures.

Concerning the values of external validity indices obtained in this study, they were almost same for all the datasets with sampling rate of 10% when compared to those computed for the complete datasets. Therefore, this sampling rate can be evaluated good enough to discover existing clusters in datasets since it can provide high validity clustering in a relatively less computing time. Besides, in the applications in which near real-time processing is required, it could be possible to use the sample datasets drawn with a sampling rate of 1% without a remarkable information loss. This sampling rate can especially work well for the datasets which having noises and not well separated clusters.

However, the algorithms scaled for parallel processing or Map-Reduce techniques are being popular for cluster analysis in DM, we suggest that data reduction by sampling should at least be useful to determine the parameter k needed as an input by partitioning and discretization algorithms. In this way, before going to a clustering task on large datasets, k value could be determined with use of the simple random sampling within a short time.

In this study, the datasets were small in size even they were considered to be large enough for an average scientific research work. On the other hand, in a real-world application of DM we usually go to deal with really big data in a size of several hundred of terabytes or even more. In this case we recommend applying the multistage sampling techniques for data size reduction again. Nevertheless, working with a sampling rate around 1% could be still informative for cluster analysis of big data containing noisy and not well separated clusters. Finally, we suggest further studies examining the efficiency of this kind of size reduction on very big data with some improvements of random sampling algorithms.

References

- Bivand, R, Rowlingson, B, Diggle, P, Petris, G & Eglen, S 2015 'Package splancs'. R package version 2.01-38. URL <https://cran.r-project.org/web/packages/splancs/splancs.pdf>.
- Cormode, G & Duffield, N. 2014, 'Sampling for big data: A tutorial', Proc. of 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1975-1975. doi: [10.1145/2623330.2630811](https://doi.org/10.1145/2623330.2630811)
- De Souto, MCP, Coelho, ALV, Faceli, K, Sakata, TC, Bonadia, V & Costa, IG 2012 'A comparison of external clustering evaluation indices in the context of imbalanced datasets', Proc. of Brazilian Symposium on Neural Networks (SBRN), pp. 49-54. doi: [10.1109/sbrn.2012.25](https://doi.org/10.1109/sbrn.2012.25)

- Desgraupes, B. 2013 'Clustering indices', URL <http://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf>.
- Dhillon, IS & Modha, DS 2000 'A data-clustering algorithm on distributed memory multiprocessors', Revised papers from Large-Scale Parallel Data Mining, Workshop on Large-Scale Parallel KDD Systems (SIGKDD). Springer Berlin, Heidelberg, pp. 245-260. doi: [10.1007/3-540-46502-2_13](https://doi.org/10.1007/3-540-46502-2_13)
- Dudoit, S & Fridlyand, J 2002 'A prediction-based resampling method for estimating the number of clusters in a dataset', *Genome Biology*, vol. 3, no. 7, pp. 1-21.
- Hathaway, RJ, Bezdek, JC & Huband, JM 2006 'Scalable visual assessment of cluster tendency', *Pattern Recognition*, vol. 39, no. 6, pp. 1315-1324. doi: [10.1016/j.patcog.2006.02.011](https://doi.org/10.1016/j.patcog.2006.02.011).
- Hathaway, RJ & Bezdek, JC 2006 'Extending fuzzy and probabilistic clustering to very large data sets'. *Computational Statistics & Data Analysis* vol. 51, no. 1, pp. 215-234. doi: [10.1016/j.csda.2006.02.008](https://doi.org/10.1016/j.csda.2006.02.008)
- Huang, Z 1998 'Extensions to the K-means algorithm for clustering large datasets with categorical values', *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283-304.
- Jain, M & Verma, C 2014 'Adapting k-means for clustering in big data', *Int. J. of Computer Applications*, vol. 101, no.1, pp. 19-24. doi: [10.5120/17652-8457](https://doi.org/10.5120/17652-8457)
- Kothari, K & Kale, O 2014 'Survey of various clustering techniques for big data in data mining', *Int. J. of Innovative Research In Technology*, vol. 1, no.7., pp. 68-71.
- Kovács, F, Legány, C & Babos, A 2005 'Cluster Validity Measurement Techniques', Proc. of 6th Int. Symposium of Hungarian Researchers on Computational Intelligence, Nov 18-19, 2005, Budapest, Hungary.
- Kwok, T, Smith, K, Lozano, S & Taniar, D 2002 'Parallel fuzzy c-means clustering for large datasets', in B Monien & R Feldmann (Eds.): Euro-Par 2002, LNCS 2400, pp. 365-374. Springer-Verlag Berlin, Heidelberg. doi: [10.1007/3-540-45706-2_48](https://doi.org/10.1007/3-540-45706-2_48)
- Leisch, F & Dimitriadou, E 2015 'Package mlbench'. R package version 2.1-1. URL <https://cran.r-project.org/web/packages/mlbench/mlbench.pdf>.
- Murdoch, D & Chow, ED 2015 'Package ellipse'. R package version 0.3-8. URL <https://cran.r-project.org/web/packages/ellipse/ellipse.pdf>.
- Napoleon, D & Pavalakodi, S 2011 'A new method for dimensionality reduction using K-means clustering algorithm for high dimensional dataset', *Int. J of Computer Applications*, vol. 13, no.7, pp. 41-46. doi: [10.5120/1789-2471](https://doi.org/10.5120/1789-2471)
- Nagesh, HS, Goil, S & Choudhary, A 2000 'A scalable parallel subspace clustering algorithm for massive datasets', *Proc. of Int. Conf. on Parallel Processing*. IEEE, pp. 477-484. doi: [10.1109/icpp.2000.876164](https://doi.org/10.1109/icpp.2000.876164)
- Palguna, D, Joshi, V, Chakaravarthy, V., Kothari, V, Subramaniam, LV 2015 'Analysis of Sampling Algorithms for Twitter'. *Proc. of the 24th Int. Joint Conf. on Artificial Intelligence (IJCAI 2015)*, pp. 967-973.
- R Core Team 2015 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org>.
- Rendón, E, Abundez, I, Arizmendi, A & Quiroz, EM, 2011 'Internal versus External Cluster Validation Indexes' *Int. J. of Computers and Communications*, vol. 5, no. 1, pp.27-34.
- Shirkhorshidi AS, Aghabozorgi, S, Wah, TH & Herawan, T 2014 'Big data clustering: A review', *Computational Science and Its Applications – ICCSA 2014*, vol. 8583 of Lec.Notes in Comp. Sci.. pp. 707-720. doi: [10.1007/978-3-319-09156-3_49](https://doi.org/10.1007/978-3-319-09156-3_49)
- Stoffel, K & Belkoniene, A 1999 'Parallel k/h-means clustering for large datasets', Proc. of the 5th Int. Euro-Par Conference on Parallel Processing. Springer-Verlag London, pp. 1451-1454. doi: [10.1007/3-540-48311-x_205](https://doi.org/10.1007/3-540-48311-x_205)
- Thalamuthu, A, Mukhopadhyay, I, Zheng, X & Tseng, GC 2006, 'Evaluation and comparison of gene clustering methods in microarray analysis', *Bioinformatics*, vol. 22, no. 19, pp. 2405-2412. doi: [10.1093/bioinformatics/btl406](https://doi.org/10.1093/bioinformatics/btl406)
- Tillé, Y & Matei, A 2013 'sampling: Survey sampling'. R package version 2.6. URL <http://CRAN.R-project.org/package=sampling>.
- Tong, H & Kang, U 2013 'Big data clustering' Ch 11 in *Data Clustering: Algorithms and Applications*, CC

Aggarwal & CK Reddy (eds). CRC Press, pp. 259-276.

Wharton, SW 1984 'An analysis of the effects of sample size on classification performance of a histogram based cluster analysis procedure', *Pattern Recognition*, vol. 17, no.2, pp. 239–244. doi: [10.1016/0031-3203\(84\)90062-1](https://doi.org/10.1016/0031-3203(84)90062-1)

Zaki, MJ, Parthasarathy, S, Li, W & Ogihara, M 1997 'Evaluation of sampling for data mining of association rules', *Proc. of 7th Int Workshop on Research Issues in Data Engineering*, pp. 42-50. doi: [10.1109/ride.1997.583696](https://doi.org/10.1109/ride.1997.583696)

Žalik, KR & Žalik, B 2011 'Validity index for clusters of different sizes and densities', *Pattern Recognition Letters*, vol. 32, no.2, pp. 221–234. doi:[10.1016/j.patrec.2010.08.007](https://doi.org/10.1016/j.patrec.2010.08.007).

Zerhari, B, Lahcen, AA & Mouline, S 2015 'Big data clustering: Algorithms and challenge', *Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA'15)*, URL http://www.researchgate.net/publication/276934256_Big_Data_Clustering_Algorithms_and_Challenges.