Journal of Big Data

RESEARCH

Open Access

OsamorSoft: clustering index for comparison and quality validation in high throughput dataset

Ifeoma Patricia Osamor[1] and Victor Chukwudi Osamor[2*]

*Correspondence: vcosamor@gmail.com; victor.osamor@covenantuniversity.edu.ng
[2] Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria
Full list of author information is available at the end of the article

Abstract

The existence of some differences in the results obtained from varying clustering k-means algorithms necessitated the need for a simplified approach in validation of cluster quality obtained. This is partly because of differences in the way the algorithms select their first seed or centroid either randomly, sequentially or some other principles influences which tend to influence the final result outcome. Popular external cluster quality validation and comparison models require the computation of varying clustering indexes such as Rand, Jaccard, Fowlkes and Mallows, Morey and Agresti Adjusted Rand Index (ARI_{MA}) and Hubert and Arabie Adjusted Rand Index (ARI_{HA}). In literature, Hubert and Arabie Adjusted Rand Index (ARI_{HA}) has been adjudged as a good measure of cluster validity. Based on ARI_{HA} as a popular clustering quality index, we developed *OsamorSoft* which constitutes *DNA_Omatrix* and *OsamorSpreadSheet* as a tool for cluster quality validation in high throughput analysis. The proposed method will help to bridge the yawning gap created by lesser number of friendly tools available to externally evaluate the ever-increasing number of clustering algorithms. Our implementation was tested alongside with clusters created with four k-means algorithms using malaria microarray data. Furthermore, our results evolved a compact 4-stage *OsamorSpreadSheet* statistics that our easy-to-use GUI java and spreadsheet-based tool of *OsamorSoft* uses for cluster quality comparison. It is recommended that a framework be evolved to facilitate the simplified integration and automation of several other cluster validity indexes for comparative analysis of big data problems.

Keywords: *Clustering index*, *Algorithms*, *OsamorSoft*, *Validation*, *Rand*, *Automation*

Introduction

Given dataset points X_n as genes, $x_1, x_2, x_3, \ldots, x_n$, in d dimensional space say R^d, clustering process can be clearly stated as thus:

We are required to find partition subsets $X_1, X_2, X_3, \ldots, X_k \; \forall \; x_i, \; i = 1,2,3,\ldots,.n$, such that every gene falls into one of the subsets and no x_i falls into two or more subsets.

Partitions $X_1, X_2, X_3, \ldots, X_k$ satisfy the following: $X_1 \cup X_2 \cup X_3 \ldots \cup X_k = X$ and $X_i \cap X_j = 0 \; \forall \; i \neq j$, where $\cup$ represents union and $\cap$ represents intersection.

In addition, we cluster to form subsets with the goal that data points x_i that are similar as much as possible belongs to same group. This require a similarity measure (or dissimilarity

Springer Open

measure) usually given in form of values to represent the degree of resemblance or natural association between one data and another [1–5]. The converse indicates dissimilarity measure ρ which satisfies the following condition:

$\rho(x_i,x_i) = 0$ and $\rho(x_i,x_j) \neq 0 \ \forall \ i \neq j$.

The use of clustering algorithms covers all aspects of studies ranging from Science and Engineering to Social Sciences and Humanities. In genomics, clustering is applied in data analysis to put genes into groups called clusters so that the degree of similarity or level of association is strong between gene membership of one cluster or same cluster and weak between gene membership of dissimilar or different clusters. Several genomic studies find usefulness in the application of hierarchical clustering and variants including the agglomerative and divisive clustering [6]. However, in partitional clustering, we have QT (Quality Threshold) clustering [7], Self Organising Map (SOM) [8] and Standard k-means [1, 4] which continue to evolve in recent years [9, 10] for analysis of high dimensional data. In k-means clustering, given a set of n data points in d-dimensional space R^d and k, the task is to determine a set of k points in R^d, called centers, so as to minimize the mean squared distance from each data point to its nearest center [4, 11].

The k-means method has been shown to be effective in producing good clustering results for many practical applications such as data compression and vector quantization and pattern classification [12]. Clustering is applicable in data mining and knowledge discovery [13, 14], and it is useful in image segmentation [15, 16]. In bioinformatic analysis of gene expression [17] and other high throughput studies, the application of clustering date as far back as old as bioinformatics [18, 19] itself and currently the use of clustering seems inevitable [20]. For efficient and effective analysis of microarray data, Osamor et al. [21] recently developed a novel Pearson correlation-based Metric Matrices k-means (MMk-means) with a better run-time $O(nk^2)$ than the Traditional k-means and other variants of k-means algorithm like Overlapped and Enhanced k-means algorithms developed by Fahim et al. [11].

Big data now forms an important part of life sciences, technological developments in omics have brought about them been considered in the big data domain. The word big data implies an immense amount of unstructured or structured data which can not be processed using conventional technology and which is characterized by rapid production speed, great variety, and extreme heterogeneity. High throughput data which encompass a great deal of omics data generate large volume of data which can be analyzed to yeild useful information, these data volume can be characterized with big data production [22].

Different assumptions and statistics are made by different investigators in the development of available clustering algorithms and this perhaps account for some conflicting clustering result obtained from several algorithms over same data. This can also be a function of several choices available to the users at the time of clustering [23]. To this end, clustering results may be rendered unreliable unless they are passed through a cluster quality validation check.

Revisiting cluster index model

There is no specific definition of a cluster that has been widely agreed upon. But objects in a cluster should be identical or compatible and objects should be dissimilar in different clusters. This means maximizing the similarity within a cluster and reducing similarities between clusters [24, 25]. Clustering is among the methods used, specifically in genomics, for gaining insight into biological methods. Clustering can of course be seen in other fields of biological data processing [26]. Clustering algorithms is intended to classify clusters or groups of objects, which are more like one another than other clusters. This type of data analysis approach is closely linked to the task of generating a data model, that entails establishing a simple collection of properties that provides an understandable explanation of crucial aspects of a dataset [24]. In general, clustering approaches give further insight into complex data.

An early cluster index model was given by Rand [27] and this was followed by the work of Fowlkes and Mallows [10]. Rand statistic was first adjusted by Morey and Agresti [28, 29] with some level of error [30] before a more accurate adjustment was later done by Hubert and Arabie [31] to what is now popularly known as Hubert and Arabie Adjusted Rand Index ($\mathrm{ARI}_{\mathrm{HA}}$) [31]. Several authors in clustering community seem to believe that $\mathrm{ARI}_{\mathrm{HA}}$ is comparatively the preferred index to most other measures [30, 32]. Krieger and Green [23] proposed a model named SEGWAY based on Rand criterion measure to provide single clustering for a set of separate partitions in a database while Tseng and Kao [9] proposed a clustering algorithm christened, correlation search technique (CST) for analysis of data from gene expression.

Several algorithms have evolved [33] and more are still being implemented at increasing pace with unequal matching pace in the development and implementation of validity measures to ensure that appropriate results are obtained [34]. Few works done on cluster validity are either in the internal cluster validity measure [35, 36] or external validity measure which is our emphasis in this article. However, tools that provide easy application of these cluster indexes are scarce and the few created are either not freely available or the programming language used for implementation are obsolete and no longer compatible or supported by current software architectures. Some common examples includes ANSI COBOL program, *Clustisz* for IBM 370 DOS created in 1975 by McClain and Rao [37], and the 1996 BASIC program of Saltstone and Stange [38] whose cluster quality validation implementation are now obsolete and publicly unavailable online. These tool in their current state, are no longer compatible with current software environment hence the need for a new tool that is compatible, platform independent, easy-to-use, friendly, and publicly available as proposed in this work. We were able to improve the reliability of clustering results through the implementation and use of a Java and Spreadsheet-based cluster validity tool called *OsamorSoft* using the clustering index model of Hubert and Arabie Adjusted Rand Index ($\mathrm{ARI}_{\mathrm{HA}}$). The proposed solution constitutes *DNA_Omatrix* and *OsamorSpreadSheet* used to respectively generate a matching matrix and evaluate the level of agreement between the cluster output of two k-means algorithms. It could also be useful in determining the cluster quality of each algorithm by comparing it with clusters of known structure as benchmark or gold standard algorithm.

Methodology

Basic concept

- If there are two clustering algorithms 1 and algorithm 2, each with its cluster output to be compared to one another, then the output of these algorithms represents the cluster partitions with *cluster* i $(i=1,2,3 \ldots, I)$ and cluster j $(j=1,2,3 \ldots, J)$.

- If there is a gold standard cluster upon whose benchmark is to be used to determine the quality of clustering of another cluster algorithm, then the gold standard cluster structure should be known.
- Each of these partitions has clusters(subsets/groups) and the union of these subsets is the total number of genes *N* being clustered.
- The intersection of any two subsets may or may not be 0 for the number of matching genes in a pair of subsets which helps to generate the matching matrix *H.*
- Matching matrix is a text file which when opened using a Spreadsheet shows clearly the number of elements that are found in same intersecting clusters forming the square matrix.
- $\mathrm{ARI}_{\mathrm{HA}}$ is Hubert and Arabie Adjusted Rand Index which is an improved form of Rand Index, adjudged as more accurate and improved Rand Index is also computed by four defined parameters represented by a, b, c, and d.
- *OsamorSoft* constitutes the entire software comprising *DNA_OMatrix* and *OsamorSpreadSheet.*
- DNA_OMatrix is a Java module that generates a matching matrix that constitute the combination of algorithm 1 vs algorithm 2.
- *OsamorSpreadSheet* is an integrated Spreadsheet embedded with formula to automatically compute $\mathrm{ARI}_{\mathrm{HA}}$ without human aid.

Clustering validation model

One interesting clustering validation task is comparative method validity for determining the quality of clusters between two algorithms [5, 27, 30, 31, 39]. Assuming there are two partitions of *h* data points being clustered by two algorithms and we are interested to know how closely their resultant output agree, then we need to obtain a matching matrix *H.*

The matching matrix $H=\{h_{ij}\}$ representing the number of genes in cluster i $(i=1,2,3 \ldots, I)$ where i represents subsets from algorithm 1 and cluster j $(j=1,2,3 \ldots, J)$ where j represents subsets from algorithm 2 [32].

The details of the formula and variables of Hubert and Arabie Adjusted Rand Index presented in Eq. 1, 2, 3, 4, 5, 6 are further described in Hubert and Arabie [31], Yeung and Ruzzo [40], Steinley [30], Warren [32], Santos and Embrechts [41]

$$a = \sum_{r=1}^{R}\sum_{c=1}^{C}\binom{t_{rc}}{2} = \left(\sum_{r=1}^{R}\sum_{c=1}^{C}t_{rc}^{2} - n\right)/2 \tag{1}$$

$$b = \sum_{r=1}^{R}\binom{t_{r}}{2} = \left(\sum_{r=1}^{R}t_{r}^{2} - \sum_{r=1}^{R}\sum_{c=1}^{C}t_{rc}^{2}\right)/2 \tag{2}$$

$$c = \sum_{c=1}^{C}\binom{t_{c}}{2} = \left(\sum_{c=1}^{C}t_{c}^{2} - \sum_{r=1}^{R}\sum_{c=1}^{C}t_{rc}^{2}\right)/2 \tag{3}$$

$$d = \binom{n}{2} - a - b - c = \binom{n}{2} - \sum_{r=1}^{R}\binom{t_r}{2} - \sum_{c=1}^{C}\binom{t_c}{2} + a \tag{4}$$

$$d = \left(\sum_{r=1}^{R}\sum_{c=1}^{C} \mathrm{t}_{rc}^{2} + n^{2} - \sum_{r=1}^{R} t_r^{2} - \sum_{c=1}^{C} t_c^{2}\right)/2 \tag{5}$$

$$ARI_{HA} = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{N}{2}^{Z} - [(a+b)(a+c) + (c+d)(b+d)]} \tag{6}$$

.

Microarray data

The clustering of genomic data has been an outstanding approach to dealing with high-dimensional data provided by high-throughput technologies, such as microarrays of gene expression. With the advent of new technologies, such as next-generation sequencing, microarrays and eQTL mapping, large volume of high throughput biological data in terabytes are now produced [26]. The DNA microarray is one of the widely used techniques for evaluating millions of gene expressions simultaneously and microarray data were stored for any further analysis in publicly accessible repositories such as the Gene Expression Omnibus (GEO) [26].

Owing to its large number of features and the limited sample sizes, microarray data classification is a challenging task for many researchers [42, 43]. This type of data is used to collect tissue and cell sample information about the variations in gene expression that may aid in the diagnosis of a disease or to differentiate certain tumor types [42]. The microarray databases used most frequently in the literature were studied in [42] and the data characteristics issues such as their complexity and data imbalances were highlighted, hence, microarray data arising from image analysis output was used in the course of the research.

Design and implementation

Due to the need to have a clearer definition of the computed $\mathrm{ARI}_{\mathrm{HA}}$ values so that outcomes can be unambiguously related to specific biological meaning and interpretation, we designed and evolved a 4-stage cluster validation value categorization. This is a modification of the Kappa statistics categorization which has some challenges in relating its outcome to biological interpretation. We programmatically counted the number of cluster membership obtained from clustering experiment of malaria microarray data using four algorithms namely Traditional k-means, Enhanced k-means, Overlapped k-means [11] and MMkmeans as described in Osamor et al. [21], [44] and use it to observe a trend. This guided us to set heuristics scale similar to Steinley [30] with stages namely Excellent, Moderate, Fair and Poor for description of $\mathrm{ARI}_{\mathrm{HA}}$ as in Table 1.

Our method accepts cluster output of high-throughput data from two algorithms to be compared i.e. Algorithm 1 and Algorithm 2. The *DNA_Omatrix* tool accepts each

Table 1 Set of Heuristics for validating clustering technique

ARI_{HA} values	4-stage outcome
≥ 0.90	Excellent
≥ 0.80	Good
≥ 0.65	Moderate
< 0.65	Poor

cluster as a single file, truncates it into individual files equivalent to the number of clusters created. *DNA_Omatrix* processes it further to generate a matching matrix by finding the intersection of every 2 combination of clusters from the algorithms. This matrix is in turn pasted into the *OsamorSpreadSheet* which automatically computes the sum of squares of the matching matrix. Initially, the values of ARI_{HA} obtained were not very informative, hence we proposed the need to set a threshold for appropriate performance category. Our heuristics threshold are four categories ranging from Excellent with greater than or equal to 0.90 to less than 0.65 for poor level of agreement between the algorithms being compared. Despite differences in techniques, the value description is a bit similar to Kappa statistic proposed by Cohen [45, 46] but differs by having 4 sets instead of 6 sets of heuristics as present in Kappa statistics.

DNA_Omatrix module

This module is implemented in Java which is platform-independent. This is an improvement of an earlier implementation where Saltstone and Stange [38] used the older basic programming language. Our solution uses two buttons "first browse button" to select the cluster file from Algorithm 1 and the "second browse button" to select another cluster file from Algorithm 2. Pressing the Compute Matrix button invokes the sorting and intersection sub-module to collate genes of all possible paired combination of Algorithm 1 vs Algorithm 2 clusters and generate a matching matrix. Matching matrix is a text file which when opened using a Spreadsheet bears the number of items in same clusters being compared.

OsamorSpreadSheet

OsamorSpreadSheet is an integrated Spreadsheet embedded with formula to automatically compute ARI_{HA} without human aid. This module automatically computes the sum of squares, columns and rows totals. It accepts matching matrices from *DNA_Omatrix* and displays result as soon as the matrix is placed on it. Its computation is in 4 phases which can be described by pasting of matching matrix(Phase1), computation of sum of squares (Phase 2) and completion of other statistics computation for variables a, b, c, and d represented by Phase 3. These inputs are used to finally compute and display the final value of ARI_{HA} and its validity description in Phase 4.

Results and discussion

Several packages and programming languages have been used to implement clustering algorithms. Optimization and changes in existing methods are normally introduced during the development and implementation of these codes, this in turn leads to new versions of the original methods with diverse improvements, to make such codes more effective [47]. For assessing the similarities in clustering algorithms, several measures have been identified, [24] compared the normalized mutual information, Fowlkes-Mallows index, Adjusted Rand index and Jaccard index. Furthermore, the work of Shirkhorshidi et al. [48] studied the influence of Similarity and Dissimilarity Measures on clustering while Zhang and Fang [49] investigated missing data and imputation on clustering validity. In these cases, emphasized was not focused on $\mathrm{ARI}_{\mathrm{HA}}$. These problems happen when the expected value of the RI of two random partition does not take a constant value (zero for example) or the Rand statistic approaches its upper limit of unity as the number of cluster increases.

To test our method, firstly, we used example data of T_1 and T_2 matching matrix as described in Steinley [30]. This data of T_1 and T_2 fits dynamically into a 4 × 4 matrix of *OsamorSpreadSheet.* We may optionally hide the unused empty rows and columns using the spreadsheet's "hide" and "unhide" feature to obtain a good sizable interface which is optional. At the completion of the experiment, the results of $\mathrm{ARI}_{\mathrm{HA}}$ as obtained for $\mathrm{T1} = 0.245588$. Similarly, we also obtained $\mathrm{ARI}_{\mathrm{HA}}$ for $\mathrm{T}_2 = 0.7401$ following same approach. Our result is in line with the output manually generated by Steinley [30]. This implies that the T_2 has moderate validity compared to T_1 matrix with very poor validity.

Secondly, in order to ensure that further evaluation is attained using a high-throughput dataset, we employed the use of microarray data of Bozdech et al. [50, 51] containing 4915 genes by 49 time points and that of Le Roch et al. [52] containing 5119 genes by 17 time points. Following the description as specified in Osamor et al. [21], we clustered these datasets independently using MMk-means, Enhanced k-means, Traditional k-means and Overlapped k-means algorithms. *DNA_Omatrix* was used to correctly obtain the matching matrices and this was subsequently followed by the use of *OsamorSpreadSheet* to compute the $\mathrm{ARI}_{\mathrm{HA}}$ and upon evaluation, we obtained similar result in line with Osamor et al. [21]. This result further confirm that our method as described in *OsamorSoft* is appreciably good to serve as a useful tool for comparing clustering qualities of algorithms especially where matching matrices can be obtained. Interestingly, the method is also able to give clue to the suitable number of k clusters to be applied from various checks of $\mathrm{ARI}_{\mathrm{HA}}$ and ultimately make biological results easier to analyse and interpret. This is achieved by comparing the various k-cluster runs against a gold standard clustering for cluster quality checks as judged by the $\mathrm{ARI}_{\mathrm{HA}}$ value. It is worthy to recall that Xu et al. [53] recently tried to formulated a unified validity index framework but their work is restricted to only hierarchical clustering having proposed two synthetical clustering validity (SCV) indexes where minimum spanning tree was utilized to compute the compactness of intra-cluster with a view to minimise deficiencies in accuracy for existing validity indexes. We emphasize here that the methodology described represents a big data solution as could be encapsulated in the methodology pipeline shown in Fig. 1.

It is important to state clearly again that the methodology involves the use of big data because the dataset used is a large microarray data originating from imaging technology for gene expression in the biological domain. This is to say that what constitutes what can be classified as a big data varies from domain to domain but the

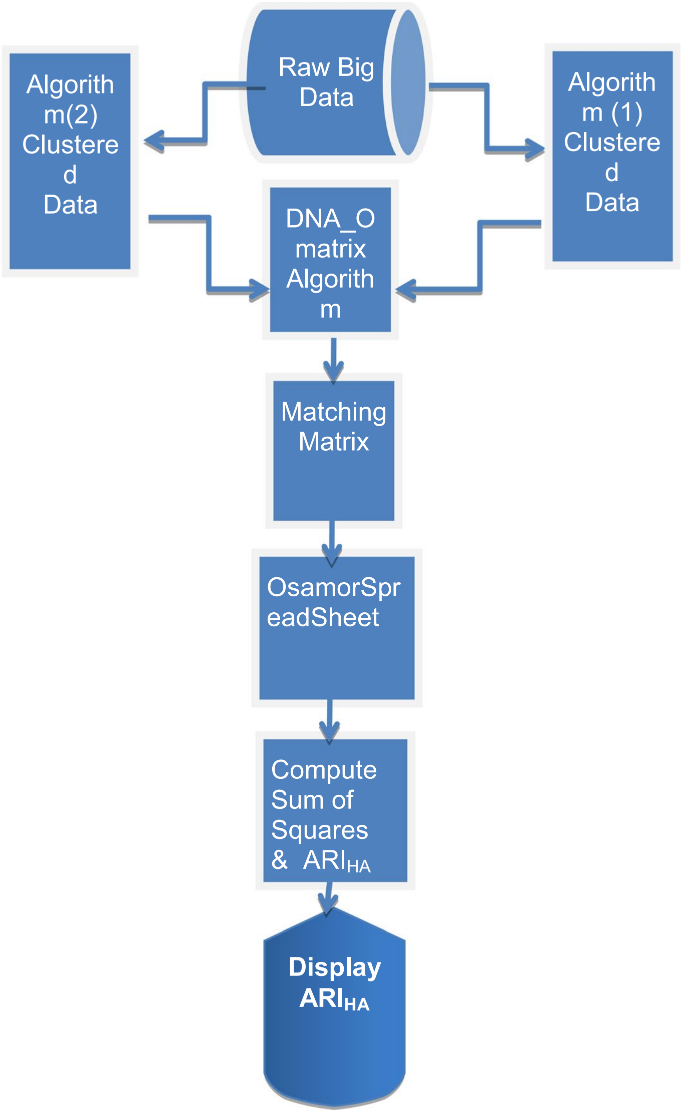
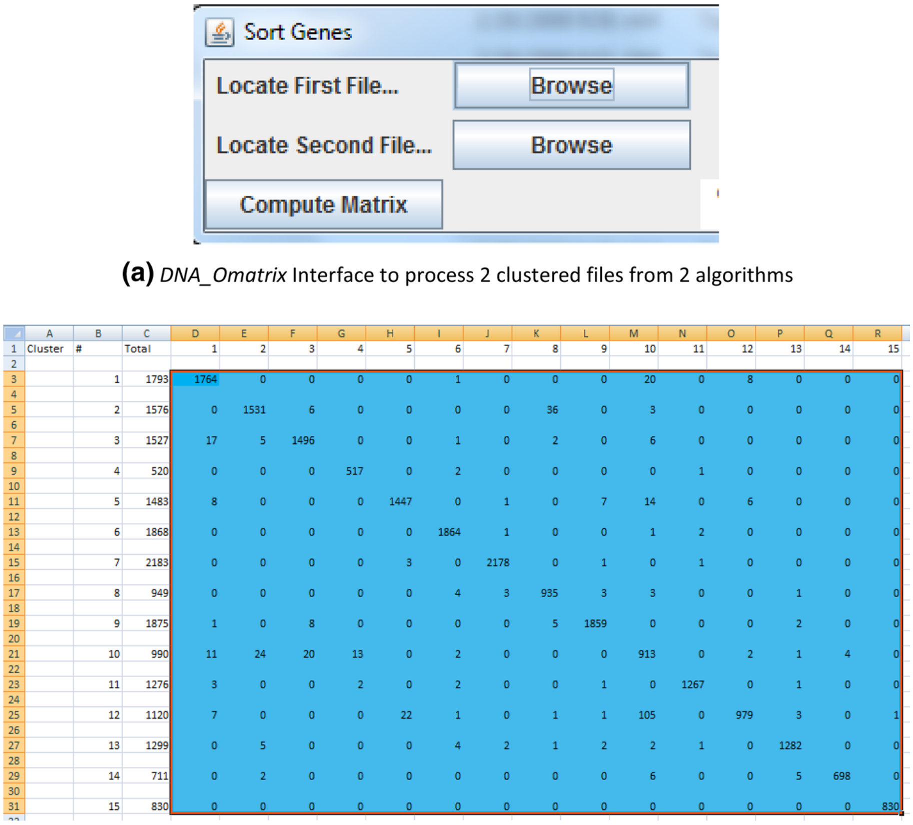

Raw Big Data
Algorith m(2) Clustere d Data
Algorith m (1) Clustere d Data
DNA_O matrix Algorith m
Matching Matrix
OsamorSpr eadSheet
Compute Sum of Squares & ARI$_{HA}$
Display ARI$_{HA}$

Fig. 1 The proposed pipeline allows raw big data access into both Algorithm 1 and 2 whose cluster partitions is intended to be evaluated. The DNA_O Matrix Java program accepts (see Fig. 2a, b), these clustered partitions separately and processes it to generate a Matching matrix that serves as input into OsamorSpreadsheet for computation of model statistics and AHI$_{HA}$. Based on the computed values of the statistics, a decision is taken by the system to conclude one choice out of the 4 defined level of evaluation of similarity ranging from value of 0–1 as specified in Table 1

major attribute of a big data is that it is large and maybe structured or unstructured and imaging technology output and application usually constitute a big data problem as seen in this work. The spreadsheet design is capable of computing AHI$_{HA}$ for 20x20 clusters on a screen viewport (See Fig. 3a–c).

However, if higher dimension clusters are required, the worksheet is increased by simple insertion of additional rows and columns in the worksheet up to k = 16,384. An empty excel spreadsheet that serves as a single worksheet is made up of 1,048,576 rows by 16,384 columns, hence this Osamorspreadsheet application is able to process



(a) *DNA_Omatrix* Interface to process 2 clustered files from 2 algorithms

Cluster	#	Total	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	1793	1764	0	0	0	0	1	0	0	0	20	0	8	0	0	0
	2	1576	0	1531	6	0	0	0	0	36	0	3	0	0	0	0	0
	3	1527	17	5	1496	0	0	1	0	2	0	6	0	0	0	0	0
	4	520	0	0	0	517	0	2	0	0	0	0	1	0	0	0	0
	5	1483	8	0	0	0	1447	0	1	0	7	14	0	6	0	0	0
	6	1868	0	0	0	0	0	1864	1	0	0	1	2	0	0	0	0
	7	2183	0	0	0	0	3	0	2178	0	1	0	1	0	0	0	0
	8	949	0	0	0	0	0	4	3	935	3	3	0	0	1	0	0
	9	1875	1	0	8	0	0	0	0	5	1859	0	0	0	2	0	0
	10	990	11	24	20	13	0	2	0	0	0	913	0	2	1	4	0
	11	1276	3	0	0	2	0	2	0	0	1	0	1267	0	1	0	0
	12	1120	7	0	0	0	22	1	0	1	1	105	0	979	3	0	1
	13	1299	0	5	0	0	0	4	2	1	2	2	1	0	1282	0	0
	14	711	0	2	0	0	0	0	0	0	0	6	0	0	5	698	0
	15	830	0	0	0	0	0	0	0	0	0	0	0	0	0	0	830

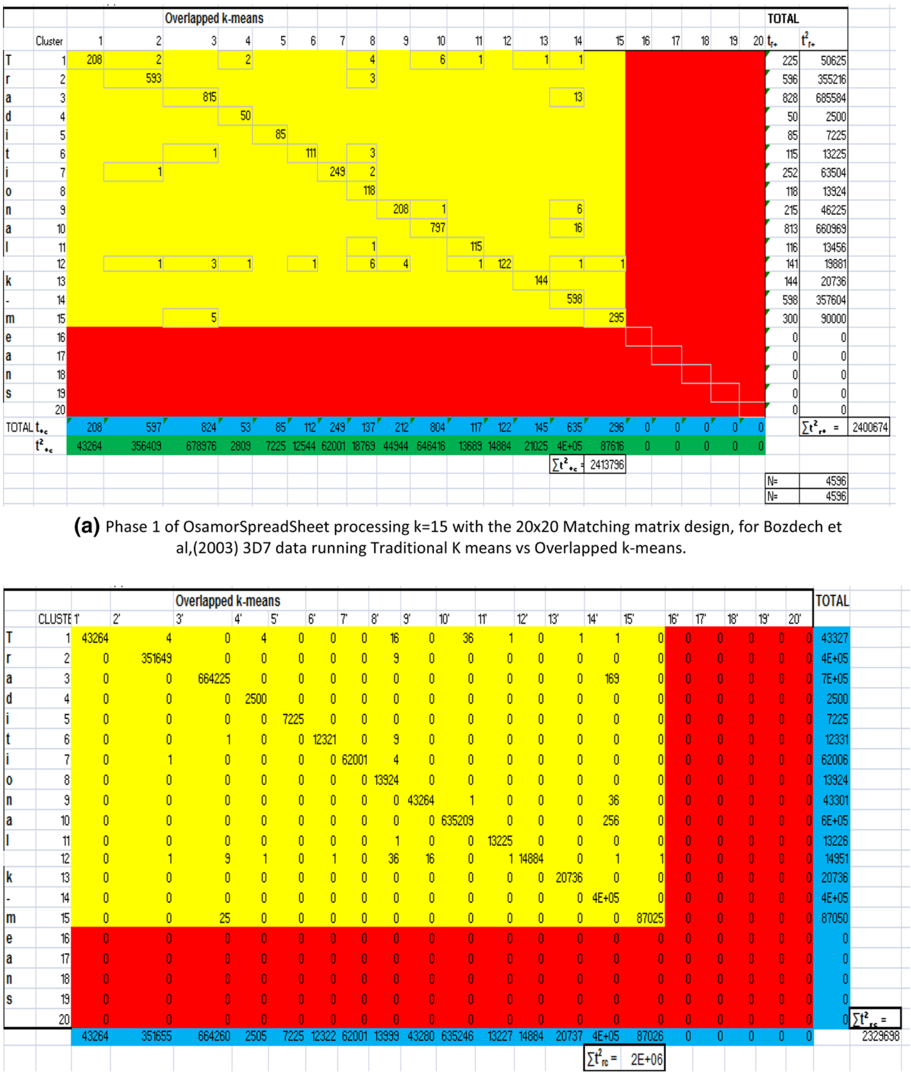
(b) Created matching matrix for k=15

Fig. 2 **a** Interface of *DNA_Omatrix* with (**b**) Matching matrix for Bozdech et al.(2003) 3D7 Data comparing two algorithms (Traditional k-means vs Overlapped k-means)

big data in the magnitude of specified by the capacity of the rows and column in the worksheet which transcends to a huge volume of data with its attendant simplicity in design and convenience to a wider user audience.

Conclusion

The reproducibility of $\mathrm{ARI}_{\mathrm{HA}}$ results of Steinley et al. [30] and Osamor et al. [21] indicate that the design and automation of our method for the computation of $\mathrm{ARI}_{\mathrm{HA}}$ is empirically proven to be adequate. To the best of our knowledge, Kappa statistics' description set of 6, seems too many for describing a small range 0–1 for precise report desired for cluster agreement between two algorithms. From the result of our malaria microarray data analysis, our heuristics are more precise, compact and likely to serve as a fairly good enhancement for Kappa statistics' assigned descriptions of values, notwithstanding the differences in techniques. Our interesting contribution is the development of a software solution tailored as an off-the-shelf application package. Spreadsheet guarantees the likelihood of its wide usage and applicability due to ease-of-use. Surely, an application resident on a spreadsheet like *OsamorSpreadSheet* will command wider usage against tools that may require program compilation,



				Overlapped k-means																		TOTAL		
	Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	t_{r+}	t^2_{r+}	
T	1	208	2		2				4		6	1		1	1							225	50625	
r	2		593						3													596	355216	
a	3			815											13							828	685584	
d	4				50																	50	2500	
i	5					85																85	7225	
t	6			1			111		3													115	13225	
i	7		1					249	2													252	63504	
o	8								118													118	13924	
n	9									208	1				6							215	46225	
a	10										797				16							813	660969	
l	11								1			115										116	13456	
	12		1	3	1		1		6	4		1	122		1	1						141	19881	
k	13													144								144	20736	
-	14														598							598	357604	
m	15			5												295						300	90000	
e	16																					0	0	
a	17																					0	0	
n	18																					0	0	
s	19																					0	0	
	20																					0	0	
TOTAL	t_{+c}	208	597	824	53	85	112	249	137	212	804	117	122	145	635	296	0	0	0	0	0		$\sum t^2_{r+}$ =	2400674
	t^2_{+c}	43264	356409	678976	2809	7225	12544	62001	18769	44944	646416	13689	14884	21025	4E+05	87616	0	0	0	0	0			
															$\sum t^2_{+c}$ =	2413796								
																						N=	4596	
																						N=	4596	

(a) Phase 1 of OsamorSpreadSheet processing k=15 with the 20x20 Matching matrix design, for Bozdech et al,(2003) 3D7 data running Traditional K means vs Overlapped k-means.

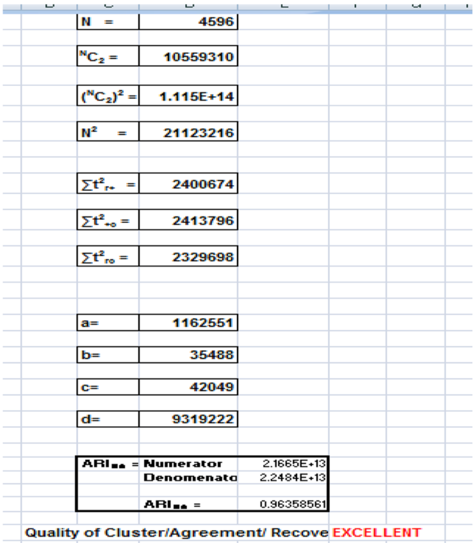
				Overlapped k-means																		TOTAL	
	CLUSTE	1'	2'	3'	4'	5'	6'	7'	8'	9'	10'	11'	12'	13'	14'	15'	16'	17'	18'	19'	20'		
T	1	43264	4	0	4	0	0	0	16	0	36	1	0	1	1	0	0	0	0	0	0	43327	
r	2	0	351649	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	4E+05	
a	3	0	0	664225	0	0	0	0	0	0	0	0	0	0	169	0	0	0	0	0	0	7E+05	
d	4	0	0	0	2500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2500	
i	5	0	0	0	0	7225	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7225	
t	6	0	0	1	0	0	12321	0	9	0	0	0	0	0	0	0	0	0	0	0	0	12331	
i	7	0	1	0	0	0	0	62001	4	0	0	0	0	0	0	0	0	0	0	0	0	62006	
o	8	0	0	0	0	0	0	0	13924	0	0	0	0	0	0	0	0	0	0	0	0	13924	
n	9	0	0	0	0	0	0	0	0	43264	1	0	0	0	36	0	0	0	0	0	0	43301	
a	10	0	0	0	0	0	0	0	0	0	635209	0	0	0	256	0	0	0	0	0	0	6E+05	
l	11	0	0	0	0	0	0	0	1	0	0	13225	0	0	0	0	0	0	0	0	0	13226	
	12	0	1	9	1	0	1	0	36	16	0	1	14884	0	1	1	0	0	0	0	0	14951	
k	13	0	0	0	0	0	0	0	0	0	0	0	0	20736	0	0	0	0	0	0	0	20736	
-	14	0	0	0	0	0	0	0	0	0	0	0	0	0	4E+05	0	0	0	0	0	0	4E+05	
m	15	0	0	25	0	0	0	0	0	0	0	0	0	0	0	87025	0	0	0	0	0	87050	
e	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
a	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
n	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
s	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\sum t^2_{rc}$ =
		43264	351655	664260	2505	7225	12322	62001	13999	43280	635246	13227	14884	20737	4E+05	87026	0	0	0	0	0		2329698
															$\sum t^2_{rc}$ =	2E+06							

(b) Phase 2 of OsamorSpreadSheet with Automatic Computation of Sum of Squares

Fig. 3 **a** Phase 1 of OsamorSpreadSheet processing k = 15 with the 20x20 Matching matrix design, for Bozdech et al. [50] 3D7 data running Traditional K means vs Overlapped k-means. **b** Phase 2 of OsamorSpreadSheet with Automatic Computation of Sum of Squares. **c** Phase 3 of OsamorSpreadSheet completes the computation of other statistics required to display the ARI_{HA} and the appropriate description of cluster validity according to our proposed design

system troubleshooting and/or complications often associated with character user interface applications.

Furthermore, another contribution may be the use of this tool to give an idea on the suitable number of clusters for an optimum initial k cluster value selection thereby, adding to the solution of finding the value of initial k cluster value problem. The method gives clue to the suitable number of k clusters to be applied from various checks of ARI_{HA} and ultimately make biological results easier to interpret. It is also important to accommodate the core idea and intuition that good clustering produces stable clustering result [53, 54] but since this does not occur in most cases, *Osamor-Soft* tool becomes very handy measure to ensure robustness against the randomness



N =	4596
$^{N}C_2$ =	10559310
$(^{N}C_2)^2$ =	1.115E+14
N^2 =	21123216
$\sum t^2_{r+}$ =	2400674
$\sum t^2_{+o}$ =	2413796
$\sum t^2_{ro}$ =	2329698
a=	1162551
b=	35488
c=	42049
d=	9319222

ARI_{HA} =	Numerator	2.1665E+13
	Denomenato	2.2484E+13
	ARI_{HA} =	0.96358561

Quality of Cluster/Agreement/ Recove EXCELLENT

(c) Phase 3 of OsamorSpreadSheet completes the computation of other statistics required to display the ARI_{HA} and the appropriate description of cluster validity according to our proposed design

Fig. 3 continued

of clustering results. Further work is desirable in evolving a framework to automate the integration of several other cluster validity schemes for easy comparative analysis and also testing the solution in HDFS or Spark or test the Java application on a distributed system.

Abbreviations

ARI_{HA}: Hubert-Arabie Adjusted Rand Index; ARI_{MA}: Morey and Agresti Adjusted Rand Index; CST: Correlation Search Technique; eQTL: Expression Quantitative Trait Loci; GEO: Gene Expression Omnibus; MMK-means: Metric Matrices K-means; QT: Quality Threshold; SOM: Self Organizing Map.

Acknowledgements

I wish to thank Covenant University for their support towards the publication expenses of this manuscript. Also, I gladly appreciate Akinnusi Opeyemi for his support during the preparation of this manuscript.

Authors' contributions

VCO initiated the idea of the work while IPO and VCO did the experiment and wrote the manuscript. All authors read and approved the final manuscript.

Funding

Covenant University is funding the cost of the publication.

Availability of data and materials

Not Available.

Competing interests

The authors do not have any competing interest.

Author details

[1] Department of Accounting, Faculty of Management Sciences, Lagos State University, Ojo Campus, Lagos, Nigeria. [2] Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria.

Received: 2 March 2019 Accepted: 2 July 2020
Published online: 09 July 2020

References

1. MacQueen J. Some methods for classification and analysis of multi-variate observations, in Proc. of the Fifth Berkeley Symp. on Math., LeCam, L.M., and Neyman, J., (eds.) Statistics and Probability, 1967.
2. Gower JC, Legendre P. Metric and Euclidean properties of dissimilarity coefficients. J Classif. 1986;3(1):5–48.
3. Batagelj V, Bren M. Comparing resemblance measures. J Classif. 1995;12(1):73–90.
4. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. A local search approximation algorithm for k-means clustering. Comput Geom. 2004;28(2–3):89–112.
5. Albatineh AN, Niewiadomska-Bugaj M, Mihalko D. On Similarity indices and correction for chance agreement. J Classif. 2006;23(2):301–13.
6. Milligan GW, Cooper MC. A Study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behav Res. 1986;21(4):441–58.
7. Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. 1999;9(11):1106–15.
8. Tamayo P, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci. 1999;96(6):2907–12.
9. Tseng VS, Kao CP. Efficiently mining gene expression data via a novel parameterless clustering method. IEEE/ACM Trans Comput Biol Bioinform. 2005;2(4):355–65.
10. Friedler SA, Mount DM. Approximation algorithm for the kinetic robust K-center problem. Comput Geom. 2010;43(6–7):572–86.
11. Fahim AM, Salem AM, Torkey FA, Ramadan MA. An efficient enhanced k-means clustering algorithm. J Zhejiang Univ Sci A. 2006;7(10):1626–33.
12. Gerso A, Gray RM. Vector quantization and signal compression. 1992;159.
13. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3):37.
14. Scott AJ, Symons MJ. Clustering methods based on likelihood ratio criteria. Biometrics. 1971;27(2):387–97.
15. Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. Pattern Anal Mach Intell IEEE Trans. 1997;19(2):153–8.
16. Marriott FHC. Practical problems in a method of cluster analysis. Biometrics. 1971;27(3):501–14.
17. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA. 1998;95(25):14863–8.
18. Cho RJ, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell. 1998;2(1):65–73.
19. Chu S, et al. The transcriptional program of sporulation in budding yeast. Science. 1998;282(5389):699–705.
20. Wen X, et al. Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci USA. 1998;95(1):334–9.
21. Osamor VC, Adebiyi EF, Oyelade JO, Doumbia S. Reducing the time requirement of k-means algorithm". PLoS ONE. 2012;7:12.
22. D'Argenio V. The high-throughput analyses era: are we ready for the data struggle? High Throughput. 2018;7:1. https://doi.org/10.3390/ht7010008.
23. Krieger AM, Green PE. A generalized rand-index method for consensus clustering of separate partitions of the same data base. J Classif. 1999;16(1):63–89.
24. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF, et al. Clustering algorithms: a comparative approach. PLoS ONE. 2019;14:1. https://doi.org/10.1371/journal.pone.0210236.
25. Hämäläinen J, Jauhiainen S, Kärkkäinen T. Comparison of internal clustering validation indices for prototype-based clustering. Algorithms. 2017;10:3. https://doi.org/10.3390/a10030105.
26. Pirim H, Ekşioğlu B, Perkins A, Yüceer C. Clustering of high throughput gene expression data. Comput Oper Res. 2012;39(12):3046–61. https://doi.org/10.1016/j.cor.2012.03.008.
27. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66(336):846.
28. Morey LC, Blashfield RK, Skinner HA. A comparison of cluster analysis techniques withing a sequential validation framework. Multivariate Behav Res. 1983;18(3):309–29.
29. Morey LC, Agresti A. The measurement of classification agreement: an adjustment to the rand statistic for chance agreement. Educ Psychol Meas. 1984;44(1):33–7.
30. Steinley D. Properties of the hubert-arabie adjusted rand index. Psychol Methods. 2004;9(3):386–96.
31. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(1):193–218.
32. Warrens MJ. On the equivalence of cohen's kappa and the hubert-arabie adjusted rand index. J Classif. 2008;25(2):177–83.
33. Llet R, Ortiz MC, Sarabia LA, Sánchez MS. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. Anal Chim Acta. 2004;515(1):87–100.
34. Milligan GW. A monte carlo study of thirty internal criterion measures for cluster analysis. Psychometrika. 1981;46(2):187–99.
35. Dunn JC. Well-separated clusters and optimal fuzzy partitions. J Cybern. 1974;4(1):95–104.
36. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
37. McClain JO, Rao VR. Clustisz: a program to test for the quality of clustering of a set of objects. J Mark Res. 1975;12(4):456–60.
38. Saltstone R, Stange K. A computer program to calculate Hubert and Arabie's adjusted rand index. J Classif. 1996;13(1):169–72.
39. Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. J Am Stat Assoc. 1983;78(383):553–69.
40. Yeung KY, Ruzzo WL. Details of the adjusted Rand index and clustering algorithms, supplement to the paper 'An empirical study on principal component analysis for clustering gene expression data. Bioinformatics. 2001;17(9):763–74.

41. Santos JM, Embrechts M. On the use of the adjusted rand index as a metric for evaluating supervised classification. Berlin: Springer; 2009.
42. Alonso-Betanzos A, Bolón-Canedo V, Morán-Fernández L, Sánchez-Maroño N. A review of microarray datasets: where to find them and specific characteristics. Methods Mol Biol. 2019;1986:65–85. https://doi.org/10.1007/978-1-4939-9442-7_4.
43. Rogers LRK, de los Campos G, Mias GI. Microarray gene expression dataset re-analysis reveals variability in influenza infection and vaccination. Front Immunol. 2019;10:2616. https://doi.org/10.3389/fimmu.2019.02616.
44. Osamor V, Adebiyi E, Doumbia S. Comparative functional classification of *Plasmodium falciparum* genes using k-means clustering, in computer science and information technology-spring conference, 2009. IACSITSC'09. International Association of. 2009; 491–495.
45. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
46. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360–3.
47. Karmakar B, Das S, Bhattacharya S, et al. Tight clustering for large datasets with an application to gene expression data. Sci Rep. 2019;9:3053. https://doi.org/10.1038/s41598-019-39459-w.
48. Shirkhorshidi AS, Aghabozorgi S, Wah TY. A comparison study on similarity and dissimilarity measures in clustering continuous data. PLoS ONE. 2015;10(12):e0144059. https://doi.org/10.1371/journal.pone.0144059.
49. Zhang Z, Fang H. Multiple-vs non-or single-imputation based fuzzy clustering for incomplete longitudinal behavioral intervention data. In 2016 IEEE first international conference on connected health: applications, systems and engineering technologies (CHASE). 2016; 219–228.
50. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol. 2003;1(1):5.
51. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL. Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol. 2003;4(2):R9.
52. Roch KG, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science. 2003;301(5639):1503–8.
53. Xu Q, Zhang Q, Liu J, Luo B. Efficient synthetical clustering validity indexes for hierarchical clustering. Expert Syst Appl. 2020;151:113367.
54. Wang H, Mahmud MS, Fang H, Wang C. Wireless Health, SpringerBriefs in Computer Science. 2016; 30

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.