International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.01 (Nov-2021)

http://dx.doi.org/10.12785/ijcds/1001102

OBKML-GO: Optimized Clustering Combination With Biological Knowledge For DNA Microarray Expression Data

Houda Fyad[1], Fatiha Barigou[1], Karim Bouamrane[1] and Baghdad Atmani[1]

[1] *Laboratoire d'Informatique d'Oran, Département d'Informatique, Université Oran1, BP 1524 El M'naouer Oran Algeria.*

Received 9 Jul.2020, Revised 4 April. 2021, Accepted 10 April. 2021, Published 25 Nov. 2021

Abstract: Several clustering techniques have been developed to help researchers analyze the large amount of information derived from genomic data. These techniques have led to the discovery of new expression patterns under different experimental conditions. One of the objectives of these methods is to cluster the profiles of co-expressed genes. However, the grouping of genes requires optimization and consistency with the reality of the biological data. This paper addresses these two aspects using the Bisecting KMeans (BKM) algorithm optimized with the WB validity index. For each cluster obtained at the end of the execution of the BKM algorithm, a profile representing this cluster that will be named leader is determined by the Leader Clustering algorithm. Then, the semantic computing of the Gene Ontology terms by the GOGO measurement is combined with the results of the optimized clustering. The proposed approach, called OBKML-GO (Optimized Bisecting KMeans Leader with Gene Ontology), is carried out on three benchmarks of model organisms: Yeast, Human and the plant *Arabidopsis thaliana*. The results show that this approach produces more relevant and coherent groups of co-expressed genes, reflecting at the same time the biological reality.

Keywords: Gene Expression, Bisecting KMeans, Optimized Clustering, Index ValidityWB, Gene Ontology.

1. INTRODUCTION

The increasing amount of massive data produced by DNA chips in the field of genes analysis, their transcripts, and their proteins, has led to the emergence of various machine learning techniques [1,2]. These techniques, in particular clustering methods, have made it possible to group or classify the profiles of co-expressed genes [3], to understand the behavior of the gene at different times and under different experimental conditions [4], and to assign the corresponding annotations to the genes [5]. However, these intensive data, provided by heterogeneous resources, are "noisy" and sometimes contain outliers that produce a number of "false positives" in biological inference [6,7]. These problems can influence the efficiency of computing approaches, i.e. the quality and consistency of the cluster results obtained.

One solution to overcome these problems may be to select relevant and optimal clusters according to an appropriate validity index or indexes [6,8,9]. These indexes, which use the concepts of separation and compactness, allow obtainingmore interesting clusters[6,8,9].

The integration of the biological knowledge provided by Gene Ontology (GO) in the previous proposal may be an alternative solution. In this case GO would be described by its different gene entities and associated annotations[10,11,12]. This can lead to homogeneous clusters and better coherence with the biological reality [10,11,12].

This paper presents a new OBKML-GO (Optimized Bisecting KMeans Leader with Gene Ontology) method for analyzing the expression of microarray data. This method combines two different aspects of gene expression: numerical and semantic. Using the numerical data, the Bisecting KMeans algorithm (BKM) is executed. The WB validity index is used to find the optimum clusters number. Then, using the Leader algorithm, each previously obtained cluster is provided with a representative gene profile whose distance from the elements of its group is minimal. In terms of semantics, the semantic similarity between genes is calculated using the GOGO measure. Clusters areformed using this distance. The Leader profile is used to find the

E-mail address: houdafyd82@gmail.com, fatbarigou@gmail.com, kbouamrane20@gmail.com,atmani.baghdad@gmail.com

semantic cluster corresponding to the numericalcluster.Fusion is performed, and the result is clustered again.

The remaining sections of the paper are organized in the following order: Section 2 displays an overview of the different methods using clustering and combination approaches for analyzing gene expression data. Section 3 details the proposed method process. Section 4 explains the experiments carried out, while Section 5 presents the findings obtained using the new approach. Then the results discussion isin the Section 6, and finally, Section 7 terminates the paper and provides suggestions for future research.

2. STATE OF ART

Several studies have highlighted the value of analyzing DNA chip data through the implementation of clustering techniques. These techniques have thus emerged as an effective means of identifying and grouping expression data of genes that behave similarly under various experimental conditions. They associate the expression data of genes with their biological functions and reveal the hidden patterns of these DNA chips. Various methods of clustering are reported in the literature: hierarchical methods (AGNES and DIANA) [13,14], partitioning methods(KMeans, PAM and CLARA) [15,14], Kohonen map-based methods (Self Organizing Map: SOM) [14], and fuzzy clustering (FCM) [15,14], graph-based methods (MST) [16,14], grid-based methods (STING, CLICK) [14], density-based methods (OPTICS, DBSCAN) [14], and Gaussian and spectral clustering methods [14].

Generally,for calculating clusters, each clustering approach has its own set of criterion. It is the nature of the datasets studied and the objectives targeted by the experimenter which predominate in the choice of the method to be used [17,18]. However, despite the usefulness of these methods, the main drawback is that with the high dimensionality of microarray data. The clustering of such data becomes a complex problem requiring optimization to improve the results quality. This optimization consists in some cases in applying validity measures. Examples of such indices are the Dunn index [8], the homogeneity [9] and separation indices [9], and the Davies & Bouldin index [9]. The aim is to use this type of index as a fitness function for better partitioning and convergence towards an optimal solution [9].

On the other hand,for improving coherence with biological reality, the integration of biological knowledge (GO) with the results of genetic clustering has became necessary. This has resulted creation and development of numerous controlled vocabularies and ontologies.They then have engendered the definition of the concepts of different biological terminologies and their products with the associated annotation [19]. The best known among biological ontologies is Gene Ontology (GO)which includes the description of three aspects: functions, processes, and components of genes for different species[19]. Measures of similarity have been used to capture the semantics of GO terms [20]. However, in the automatic comparison of genes, the semantic measurements are based solely on the analysis of biological aspects. This comparison is formalized by the hierarchical structure of the GO independently of particular genetic properties [21].

A new paradigm has been proposed for a more complete analysis taking into account the quality and consistency of the processed data. It consists to combine the cluster results obtained from expression profiles with GO as a source of biological information[22,23].This combination offers enrichment, improved annotation and helps researchers to discover more relevant models [23]. Other authors [24] have associated expression data with GO terms for the classification of expressed genes from plants under biotic and abiotic stress. Using a method called Gene Selection based on Biologically Relevant Clusters (GSBRC), these authors merged the numerical distance matrix of gene profile expression with the semantic matrix of GO terms [24]. The fusion matrix is subjected to the Affinity_Propagation (AP) algorithm which determines and groups genes by clustering. Then, the number of gene attributes is reduced by the Neighborhood Rough algorithm. Finally, these genes are classified by Support-Vector_Machine (SVM) [24]. The results of the integration of biological knowledge have led to an increase in the precision index [24]. In [11], the authors implemented the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) on the two combined matrices (numerical and semantic) of expression data. The clustering performed on this fusion is optimized by the Co-expression-Indicator (CI) andBiological-Homogeneity-Index (BHI) indexes. The results obtained indicate that the fusion and optimization provided homogeneous and biologically coherent clusters [11]. In the same perspective, the authors[12] also proposed an approach called Multi-Objective Clustering algorithm Guided by a-Priori Biological Knowledge which uses the three validity indexes (compactness, separation, and Xie-Beni index). These indexes were used as objective functions to determine the optimal clusters of gene expression data. The simultaneous incorporation of GO terms associated with gene functions during biological processes have been also done by these authors [12].

3. PROPOSED APPROACH

The combination of clustering-optimized numerical expression data processing using validity indices can confer facilitation and improvement of gene annotation results. However, the knowledge provided by the GO terms' semantic similarity must be added.From this perspective, this study displays a method that adds the semantic aspect with the numerical gene expression data,

called OBKML-GO.Figure 1 details each workflow step of the proposed approach.

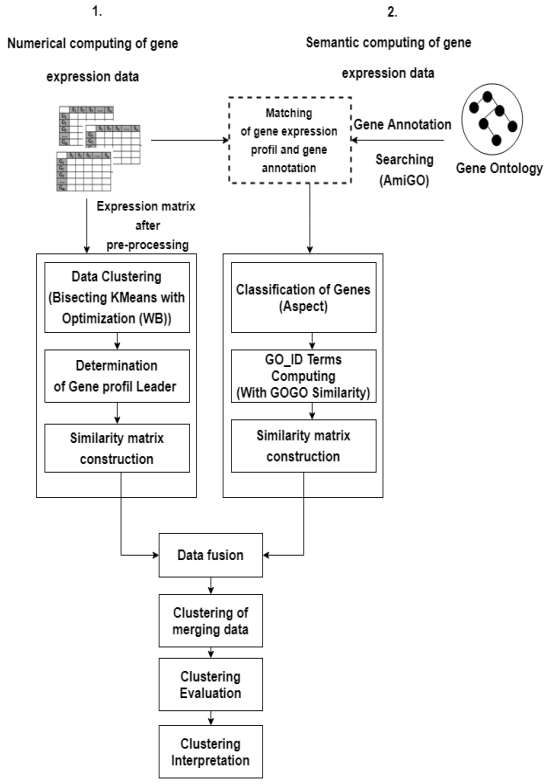

Figure 1. Workflow of the proposed approach.

A. Numerical computing of gene expression data

The numerical aspect of the presently exposed method is carried out in several steps. These steps involve the implementation of the Bisecting KMeans algorithm [25,26] enhanced by the validity indices (Sum Square Between(SSB), Sum Square Within(SSW) and WB)) for the gene expression numerical profiles. Then, a representative profile considered as the "Leader" profile at the level of each previously obtained optimal group is defined with the Leader algorithm [27]. This expression profile allows a minimum distance from the rest of the profiles of the same group.

a) Clustering of gene expression data: Bisecting KMeans Algorithm (BKM)

Bisecting KMeans (BKM) is a KMeans algorithm extension [25]. The basic idea of this algorithm is that from the data set D = {d_1, d_2...d_n}; two centroids are initially selected using the KMeans algorithm (K_{NC}=2). Then the Sum of the Squared Error (SSE) calculation is performed to determine which of the two previous clusters is to be divided in the next iteration. The one with the highest SSE value. As a result, until the final number K_{NC} of clusters is reached, the partitioning process is repeated [26].

As shown in Figure 1, the BKM algorithm is executed on the pre-processed numerical expression data of the three benchmarks [34, 35, 36]. The grouping of the expression profiles into homogeneous clusters resulting from this first step is performed using validity indices described in the following section.

b) Optimization of the BKM algorithm: Use of validity indices

Three validity measures are calculated during the BKM algorithm execution to obtain optimal expression profile clusters. The first two indices (SSB, SSW) have been widely used in the literature. This choice is guided by the contribution of these indices in terms of expected quality between clusters (SSB) and within clusters (SSW).The third index (WB) calculates the inter-and intra-cluster ratio weighted by the clustersnumber. This index improves considerably the clusters quality [28].

i. Between cluster variance (SSB)

$$SSB(k) = \sum_{i=1}^{k} n_i d^2(c_i, c)$$

k: indicates the set of clusters; where cluster *i*has n_i items and $c_i$is its centroid, c: is the centroid of the whole data set and d (,) : is a distance used in this algorithm.

ii. Within cluster variance (SSW)

$$SSW(k) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d^2(x_{ij}, c_i)$$

k: indicates the set of clusters where cluster *i*has n_i items and $c_i$is its centroid, x_{ij} is the j[th] data point in a cluster c_i and d (,) : is a distance used in this algorithm.

iii. Sum-of-Squares Based Cluster (WB)

$$WB = \frac{k * SSW}{SSB}$$

k: is the set of clusters, SSW (rep. SSB): represents the Within (resp. Between) cluster variance.

c) Determination of the Leader expression profile

The Leader algorithm is used to choose the Leader expression profile for each cluster. In the same cluster, the distance between the Leader vector with the other vectors must be lower or equal to a predefined threshold value [27]. Thus, each optimized cluster has a Leader profile.

The numerical computing on the expression data is summarized in the following pseudo-algorithm:

Input: Matrix of gene expression data (n*m) with: "n" genes profiles and "m" experimental conditions,
K_{NC} : number of clusters, α : the threshold value.
Output: C = {c_1, c_2…c_k} set of clusters. SSB, SS, WB Values, L = {l_1, l_2…l_k} a list of leader genes profiles vectors associated with their followers L_1= {f(l_1), f(l_2)… f(l_m)}.

//Optimized BKM algorithm steps:

1. Begin
2. Start with a single cluster GLOBAL(C) containing all genes profiles.
3. **Repeat**
4. **for**i := 1 to I a number of iteration **do**
5. Apply for GLOBAL (C) the KMeans algorithm (K_{NC}=2)
6. Calculate the SEE of each two clusters obtained and choose the largest SEE
7. Compute SSB, SSW and WB indices
8. Save the minimum value of WB index
9. **end for**
10. Add these two clusters in the bisecting cluster list with the most overall similarity.
11. **until** the number of clusters defined is found.

//Leader-Followers algorithm steps

12. Consider each centroid in each cluster obtained before, as gene leader profile vector l_j
13. **for** all the rest of vectors v_i in the same cluster.
14. Calculate the distance between the gene leader profile vector and the rest of vectors
15. **if**$(Distance := argmin\|v_i - l_j\|^2 < \alpha)$**then**
16. Consider $l_j$as leader new profile vector
17. **else**
18. Determine $v_i$as leader new profile vector
19. **end if**
20. **End for**
21. **End.**

B. Semantic computing of gene expression data

Gene Ontology (GO) is required to define the biological similarity of genes, where terms representing the genes and gene products of different organisms are defined according to three angles: biological process (OntoBP), cellular component (OntoCC), and molecular function (OntoMF) [19,29]. Each term in GO is connected to one or more other terms in a Directed Acyclic Graph (DAG) [19,29]. A term in the GO has associated annotations that describe the function(s) of genes and gene products [19,29].
Each annotation contains the following main information that uniquely identifies it. It first contains the searched gene (or gene product), the term GO, the annotation reference. There is a code on how the annotation was made (manual, automatic, or other) by evidence codes. The date and the author of the annotation are indicated at the end.

a) Extraction of genes and their annotation

A web-based application called AmiGO [29] is available online to provide access to the information provided by the GO. It allows users to search for GO terms or gene symbols, and to browse and view the ontologies and annotations associated with those genes. Users can also refine the result of the query in the AmiGO search bar in different ways: they can specify the studied organism. They can choose one of the aspects of the GO (biological process (OntoBP), cell component (OntoCC), or molecular function (OntoMF). They can even search according to one of the particular evidence codes for example [29].

b) Correspondence with gene symbols for numerical computing

Extraction of the genes and annotations associated with these genes according to the three organisms in this study namely, Yeast, Human, and *Arabidopsis thaliana* plant was performed. The gene expression matrix, which defines a given gene symbol's expression profile under various experimental conditions, was then matched with the same gene symbol found in the AmiGO research.

c) Semantic similarity computing with the GOGO measure

To evaluate this biological similarity, several semantic measures have been suggested. The DAG topology of GO is used to define some of them that includes the nodes (or the edges) and the distance to the lowest common ancestor node (LCA) [20,30]. The informative content (IC) of the terms was used to develop other measures. Hybrid measures based on these two methods have been also proposed [20,30]. Among them the GOGO functional similarity, which has the advantage of taking into account the number of child nodes in the similarity calculation [31].

When T_U is the set of GO terms that encompasses U and its ancestors, E_U the set of relationships (edges) among nodes of T_U in DAG_U. In this case, a GO U term can be formally defined as the following triplet: DAG (U, T_U, E_U). In the first step, the weighting of the semantic contribution which is tributary of the relationship kind and the number of offspring, is calculated by using Equation (1) as follows:

$$\varpi_e = \frac{1}{(\varphi + n_{off}(t) + \eta)} \quad (1)$$

Where $n_{off}(t)$ denotes the total number of descendants for the GO term t, φ and η denote invariable parameters. The relationship type between a GO term and its parent nodes is defined by the parameterηwhich is similar to the d in Wang's semantic measure [32]. Then, using Equation (2), the S_value for the target term U is determined according to Wang's method. The Semantic_value of the GO term U is computed in this case, as reported by Equation (3).

$$\begin{cases} S_U(t)=1 & When \quad t=U \\ S_U(t)=\max\{\varpi_e * S_U(t') \mid t' \in children(t)\} & Other \end{cases} \quad (2)$$

$$SV(U)=\sum_{t\in T_U} S_U(t) \quad (3)$$

Hence, for the following pairwise U and W of GO terms, the semantic similarity between them is defined by Equation (4):

$$S_{GO}(U,W)=\frac{\sum_{t\in T_U \cap T_W}(S_U(t)+S_W(t))}{SV(U)+SV(W)} \quad (4)$$

In the last step, the functional similarity between g_1 with *m* terms of GO and g_2 with *n* terms of GO according to the Best-Match Average (BMA) strategy is calculated using Equation (5):

$$sim_{BMA}(g_1,g_2)=\frac{\sum_{i=1}^{m}\max_{i\le j\le n} sim(go_{1i},go_{2j})+\sum_{j=1}^{n}\max_{i\le i\le m} sim(go_{1i},go_{2j})}{n+m} \quad (5)$$

Given that i is a term from group g_1 (i varying from 1 to *m*) and j a term from group g_2 (j varying from 1 to *n*).

C. Fusion of numerical and semantic information of gene expression data

A numerical similarity matrix is created after performing the numerical computing of the expression data. It qualifies the correlation of the genes according to the value of the expression profiles. In this case, Pearson's correlation coefficient was used for measuring this correlation [33].Another semantic similarity matrix is established, reflecting the functional similarity of the genes from the biological annotation terms of Gene Ontology.

This semantic similarity is calculated using the GOGO measure explained above. The two matrices are merged into a single matrix (see Figure 2), with an associated weighting γ defined as follows [23]:

$$Similarity=\gamma Similarity_{numerical}+(1-\gamma)Similarity_{semantic} \quad (6)$$

$Similarity_{numerical}$ is computed on the numerical similarities calculation basis of the gene expression data. $Similarity_{semantic}$ is computed based on the semantic computation of the annotations of the terms of the GO. The coefficient γ is a parameter whose values are included in [0,1]. If the coefficient γ = 1, only the numerical distances of the expression data are used. However, if the coefficient γ = 0, only the semantic similarity is used. The parameter γ delimits the influence of biological knowledge on the quality and consistency of the outcomes [23].

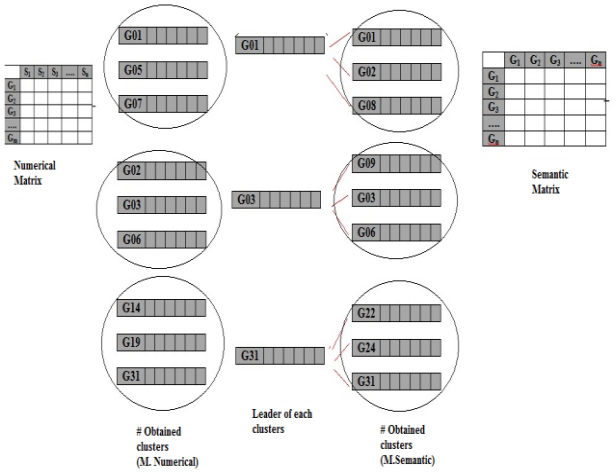

Figure 2. Numerical and semantic fusion of gene expression data.

4. EXPERIMENTATION

The effectiveness of the proposed approach applied to three benchmarks of real organisms: yeast, humans and the *Arabidopsis thaliana* plant, is analyzed and assessed using the various experiments described in this section.

A. Reference benchmarks used

Three reference expression benchmarks were utilized to assess the efficacy of the suggested approach: Yeast Sporulation [34], Human Fibroblasts Serum [35], and Arabidopsis thaliana [36]. Table I outlines the features of these three real-world benchmarks. The "Original genes" column concerns to the number of genes from the DNA micro-arrays experimentation. The "Genes treated" column corresponds to the number of genes obtained by pre-processing where duplicate genes and missing values of expression levels have been deleted from the initial data. The downsized genes number was utilized for the different experiments of our proposed approach.

TABLE I. MICROARRAY BENCHMARKS DESCRIPTION

Benchmarks	#Original Genes	#Treated Genes	#Samples
Yeast Sporulation	6118	474	7
Human fibroblasts serum	8613	517	13
Arabidopsis thaliana	138	133	8

a) Benchmark 1: Yeast Sporulation

The benchmark corresponding to Yeast Sporulation was downloadedfrom http://cmgm.stanford.edu/pbrown/sporulation. It shows the levels of gene expression in 6118 genes at seven different intervals (t0, t0.5, t2, t5, t7, t9 and t11.5). Of these genes, only 474 that were considered to differ significantly in their expression levels were retained during this sporulation process [37].

b) Benchmark 2: Human fibroblast serum

The Information on expression levels of 8613 genes is provided in the Human Fibroblast Serum Benchmark available athttp://www.sciencemag.org/feature/data/984559.shl.At different sampling times, from 0 to 96 quarters of an hour as follows: (0,1, 2, 4, 8 16, 24, 32, 48, 64, 80 and 96), as well as at an asynchronous time point (each sample is taken at 13 time intervals). For gene expression analysis, only a set of 517 genes was found to be significant to be chosen[37].

c) Benchmark 3: Arabidopsis thaliana

The benchmark *Arabidopsis thaliana* was downloaded fromhttp://anirbanmukhopadhyay.50webs.com/data.html. It contains the expression values of 138 genes for 8 periods (15 min, 30 min, 60 min, 90 min, 180 min, 360 min, 540 min and 1440 min). A subset of 133 genes was submitted for analysis [37].

B. Algorithms used

The performance of the OBKML-GO algorithm in numerical and semantic computing was compared to the performance of two algorithms, KMeans Basic and KMeans Optimized by the WB index.

a) KMeans Algorithm

The first of the two comparison algorithms used is the KMeans. This algorithm is described below:

Input : $D = d_i\{i = 1, \ldots . n\}$. set of data points and K_{NC} a clusters number.

Output: K_{NC} set of clusters.

1. Choose arbitrary K points from D as the initial centroids
2. **Repeat**
3. (Re) assign each point to the clusters which has the closestmeans.
4. Update the cluster means calculation between each point and the new cluster.
5. **Until** no change.

b) Optimized KMeans Algorithm

The second one is the Optimized KMeans. This algorithm is described below:

Input : $D = d_i\{i = 1, \ldots . n\}$ set of data points and K_{NC}a clusters number.

Output: K_{NC} set of clusters. WB index value

1. Choose arbitrary K points from D as the initial centroids
2. **repeat**
3. (Re) assign each point to the clusters which has the closestmeans.
4. Compute WB index
5. Save the minimum value of WB index
6. Update the cluster means calculation between each point and the new cluster.
7. **until** no change.

C. Validation indices

In the literature, various validation indices have been used to describe the quality clustering system. Overall, these indices have been subdivided into two main categories: internal measures and external measures [38]. Internal measures refer to the inter- and intra-cluster quality, without a priori information [38]. The external measures compare a result of the obtained partition with another partition known in advance.

In this study, Calinski-Harabasz (I_{CH}) [17] and Hartigan (I_H) [39] were used as internal evaluation indices. The purity measure (I_P) [40] was used for the external evaluation.

a) Calinski-Harabasz (I_{CH}) index

The Calinski-Harabasz (ICH) index is defined as the fraction of the global variance of the sum of inter-cluster squares (SSB) (compactness criterion) on the global variance of the sum of intra-cluster squares (SSW) (separation criterion) [17]. This fraction is maximized when the clusters have been well separated and compacted [17]. The I_{CH}measure is calculated using Equation (7):

$$I_{CH} = \frac{(N - K_{NC})}{(K_{NC} - 1)} \frac{SSB}{SSW} \quad (7)$$

Where, SSB is a variance between clusters; SSW is a variance within clusters; K_{NC} is the clusters number; N is the entire observations number (the total data points number).

b) Hartigan (I_H) index

The Hartigan index (I_H) is defined as the logarithm of the global variance of the sum of inter-cluster squares (SSB), i.e. the compactness criterion on the global variance of the sum of intra-cluster squares (SSW), i.e. the separation criterion [39]. The $I_H$measure is calculated using Equation (8):

$$I_H = \log_2 \frac{SSB}{SSW} \quad (8)$$

Where, SSB is a variance between clusters and SSW is a variance within clusters.

c) Purity (I_P) index

The purity index is expressed by a rate of the total number of points that have been correctly classified; these points are considered "correctly classified" if for each C_i cluster, a group of points is identified as belonging to the same class or reference cluster [40]. The purity of a $C_i$cluster is defined as:

$$Purity_i = \frac{1}{n_i} \max_{j=1}^{k_{NC}} \{n_{ij}\} \quad (9)$$

Where, $n_i$is the size of the obtained cluster C_i,k_{NC} is the number of clusters actually assigned, and n_{ij} is the number of common points between cluster C_i and the actually assigned partition R_j. The overall clustering solution purity can be given by Equation (10):

$$Purity = \frac{1}{n} \sum_{i=1}^{r} \max_{j=1}^{k_{NC}} \{n_{ij}\} \quad (10)$$

Where,r is the clusters number obtained using certain clustering algorithms and n is the total points number in a d-dimensional data set D. $D = x_i\{i = 1, \ldots . n\}$.

5. RESULTS

The clustering performance results obtained by the OBKML-GO algorithm and those of the other two algorithms are compared in this section. These calculations were performed on the numerical, semantic and combined gene expression data profiles for the three benchmarks used.

B. Case γ = 1 numerical similarity

Table II presents the numerical results (γ = 1). The OBKML-GO algorithm, except for the *Arabidopsis thaliana* benchmark, provides the best values for the I_{CH} index regardless of the benchmark used or the number of clusters.In this case, the OBKML-GO algorithm is outperformed by the optimized KMeans algorithm if the 9 number of clusters is reached. The two other indices$I_H$andI_P, show the same pattern.The values of these indices for the OBKML-GO algorithm are higher than those of the two KMeans and Optimized KMeans algorithms. In this instance the benchmarks used and the clusters number are not taken into consideration.

Overall, the values of the indices (I_{CH},$I_H$andI_P) are relatively close for a number of clusters K_{NC}={7,9 and 10} compared to the clusters number K_{NC}=5. This situation is similar for the Yeast, Human, and *Arabidopsis thaliana* benchmarks, regardless of the algorithm used. For example, for the OBKML-GO algorithm with a clusters number K_{NC}={7,9 and 10}, the value of I_{CH}varies from 37.634 to 38.500, the value of $I_H$varies from 0.662 to 0.826 and $I_P$varies from 0.768 to 0.794 for the Yeast benchmark.

This can also be noticed for the two other benchmarks Human, and *Arabidopsis thaliana* have I_{CH}index has values ranging from 36.896 to 40.151 and from 30.565 to 35.237 respectively. For $I_H$index, the first one has values from 0.613 to 0.780, and the second one has values from 0.690 to 0.778. Finally, for the $I_P$index, Human benchmark has values ranging from 0.726 to 0.751, and for *Arabidopsis thaliana* benchmark has values varying from 0.786 to 0.719.

Figure 3 illustrates the performance of our OBKML-GO algorithm for the three validity indices (I_{CH}, I_H, andI_P) with respect to the other two (KMeans and Optimized KMeans) on the three benchmarks for a number of clusters K_{NC} = 10 in the numerical aspect.

TABLE II. RESULTS OF THE THREE VALIDITY INDICES FOR CLUSTERING ALGORITHMS ON THE THREE BENCHMARKS (NUMERICAL PART)

		K_{NC}=5			K_{NC}=7			K_{NC}=9			K_{NC}=10		
		Validity indices											
Yeast Sporulation		I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P
	KMeans	30.939	0.491	0.527	34.081	0.592	0.629	35.563	0.679	0.737	**37.334**	**0.647**	**0.737**
	Optimized KMeans	32.762	0.683	0.643	34.197	0.612	0.738	36.824	0.726	0.769	**34.444**	**0.770**	**0.779**
	OBKML-GO	37.498	0.721	0.778	37.634	0.662	0.768	38.499	0.733	0.784	**38.500**	**0.826**	**0.794**
		K_{NC}=5			K_{NC}=7			K_{NC}=9			K_{NC}=10		
Human Serum Fibroblast		I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P
	KMeans	33.602	0.401	0.686	36.744	0.502	0.688	38.226	0.589	0.696	**38.499**	**0.525**	**0.754**
	Optimized KMeans	34.419	0.433	0.736	36.248	0.560	0.712	38.475	0.476	0.723	**40.080**	**0.563**	**0.748**
	OBKML-GO	34.860	0.671	0.764	36.896	0.613	0.726	38.861	0.683	0.745	**40.151**	**0.780**	**0.751**
		K_{NC}=5			K_{NC}=7			K_{NC}=9			K_{NC}=10		
Arabidopsis thaliana		I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P
	KMeans	28.368	0.375	0.613	31.510	0.476	0.615	32.992	0.563	0.623	**31.185**	**0.537**	**0.678**
	Optimized KMeans	29.505	0.682	0.644	31.334	0.527	0.664	33.561	0.725	0.675	**34.766**	**0.774**	**0.629**
	OBKML-GO	30.565	0.690	0.786	32.601	0.577	0.702	31.566	0.702	0.711	**35.237**	**0.778**	**0.719**

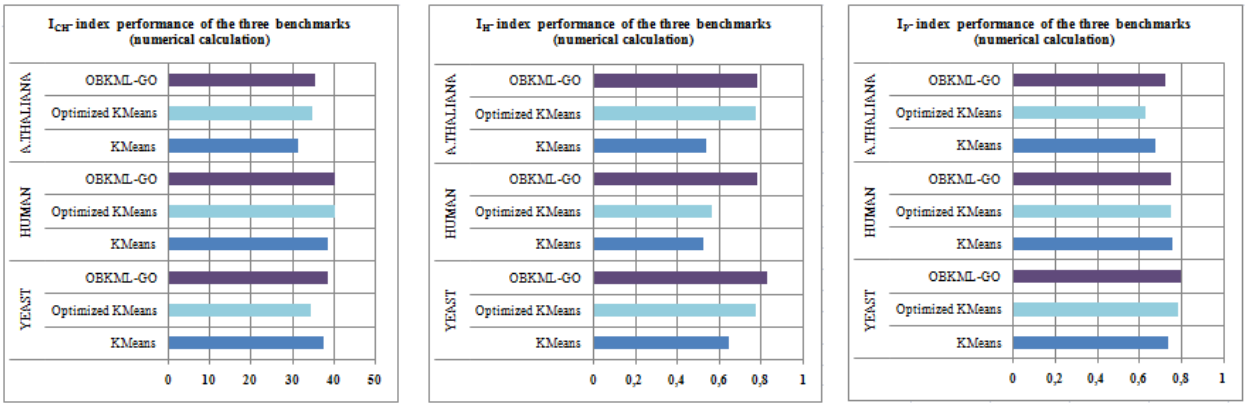

Figure 3. I_{CH}- index (left), I_H- index (center) and I_P- index (right) performance of the three benchmarks (numerical calculation).

C. *Case γ = 0 semantic similarity*

Table III shows the semantic results (γ = 0). It can be noticed that the best values obtained for the I_{CH}index are given by the OBKML-GO algorithm regardless of the benchmark used, and regardless of the number of clusters, except for the *Arabidopsis thaliana* benchmark. In this case, the OBKML-GO algorithm is outperformed by the optimized KMeans algorithm if the 9 number of clusters is reached. As an identical result can be reported for the two other indices $I_H$andI_P, i.e. the values of these indices for the OBKML-GO algorithm are higher than those of the two KMeans and Optimized KMeans algorithms independently of the benchmark used and independently of the clustersnumber. The indicesvalues(I_{CH},$I_H$andI_P) are generally relatively close for a number of clusters K_{NC}={7,9 and 10} compared to the clusters number K_{NC}=5. This is true for the Yeast, Human, and *Arabidopsis thaliana* benchmarks, regardless of the algorithm used. For instance, for the OBKML-GO algorithm with a number of clusters K_{NC}={7,9 and 10}, the value of I_{CH}ranging from 37.109 to 37.801, the value of $I_H$ranging from 0.685 to 0.797 and $I_P$ranging from 0.751 to 0.770 for the Yeast benchmark. The same result applies for the two other datasets Human and the plant *Arabidopsis thaliana* have I_{CH}index has

values range from 34.726 to 38.220 and from 33.102 to 36.395 respectively. For $I_H$index the first one has values from 0.598 to 0.798, and the second one has values from 0.686 to 0.851. Finally, for the $I_P$index, Human benchmark has values range from 0.671 to 0.753, and for *Arabidopsis thaliana* benchmark has values vary from 0.749 and 0.753.

Figure 4 illustrates the performance of our OBKML-GO algorithm for the three validity indices (I_{CH}, I_H, andI_P) with respect to the other two (KMeans and Optimized KMeans) on the three benchmarks for a number of clusters K_{NC} = 10 in the semantic aspect.

TABLE III. RESULTS OF THE THREE VALIDITY INDICES FOR CLUSTERING ALGORITHMS ON THE THREE BENCHMARKS (SEMANTICAL PART)

		K_{NC}=5			K_{NC}=7			K_{NC}=9			K_{NC}=10		
		Validity indices											
Yeast Sporulation		I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P
	KMeans	30.678	0.457	0.689	33.026	0.537	0.741	33.625	0.637	0.719	**36.388**	**0.637**	**0.721**
	Optimized KMeans	33.285	0.627	0.721	35.186	0.717	0.755	37.281	0.650	0.673	**34.854**	**0.751**	**0.726**
	OBKML-GO	36.985	0.679	0.706	37.109	0.685	0.751	38.165	0.702	0.777	**37.801**	**0.797**	**0.770**
		K_{NC}=5			K_{NC}=7			K_{NC}=9			K_{NC}=10		
Human Serum Fibroblast		I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P
	KMeans	32.146	0.390	0.689	34.494	0.473	0.675	35.088	0.670	0.678	**35.172**	**0.579**	**0.678**
	Optimized KMeans	33.604	0.537	0.745	35.654	0.659	0.677	35.597	0.760	0.682	**36.395**	**0.657**	**0.695**
	OBKML-GO	34.602	0.629	0.755	34.726	0.598	0.671	37.780	0.772	0.743	**38.220**	**0.798**	**0.753**
		K_{NC}=5			K_{NC}=7			K_{NC}=9			K_{NC}=10		
Arabidopsis thaliana		I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P	I_{CH}	I_H	I_P
	KMeans	29.418	0.574	0.614	31.766	0.654	0.723	32.360	0.754	0.735	**32.705**	**0.755**	**0.736**
	Optimized KMeans	31.555	0.625	0.736	32.926	0.834	0.734	**35.550**	0.648	0.745	**33.125**	**0.763**	**0.745**
	OBKML-GO	32.978	0.731	0.752	33.102	0.6865	0.749	34.156	0.754	0.753	**36.395**	**0.851**	**0.753**

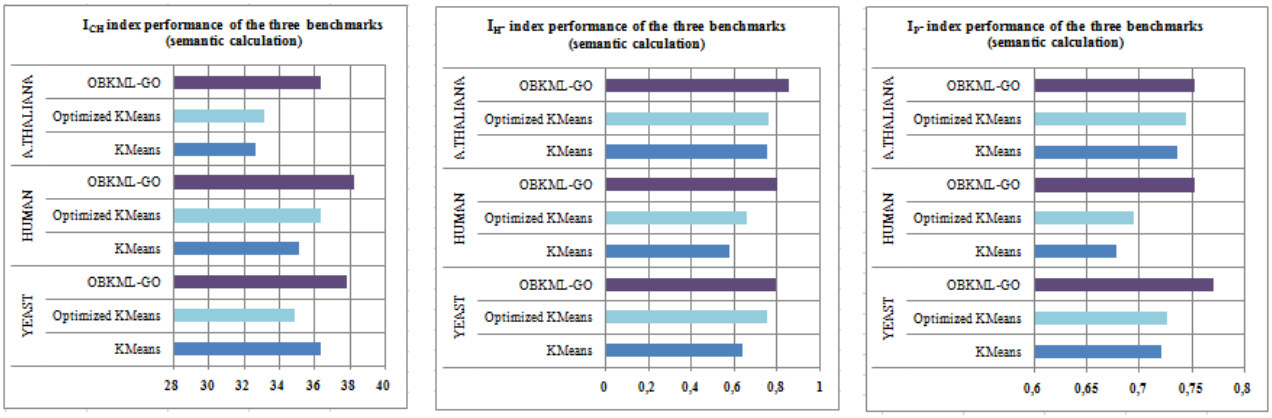

Figure 4. I_{CH}- index (left), I_H- index (center) and I_P- index (right) performance of the three benchmarks (semantical calculation).

D. *Cases $\gamma \neq 0$ and $\gamma \neq 1$: combined similarities*

The influence of the parameter γ was tested with several values ranging from 0 to 1 with a step of 0.1. Table V shows the best combination results obtained by the three validity indices with the corresponding γ value carried out on the three benchmarks.

These results show that for the Yeast benchmark, the best index values are I_{CH}= 33.087, I_H= 0.653, and I_P= 0.746 with γ = 0.1 and K_{NC}= 10. For the Human benchmark, the best index values are I_{CH}= 34.953, I_H= 0.713, and I_P= 0.768 with γ = 0.7 and K_{NC}= 9. As for the *Arabidopsis thaliana* benchmark, the best index values are I_{CH}= 32.359, I_H= 0.648, and I_P= 0.735 with γ = 0.2 and K_{NC}= 8.

TABLE IV. THE COMBINED SIMILARITIES RESULTS

Benchmarks	γ value	# Cluster	Index values		
			I_{CH}	I_H	I_P
Yeast Sporulation	0.1	10	33.087	0.653	0.746
Human fibroblasts serum	0.7	9	34.953	0.713	0.768
Arabidopsis thaliana	0.2	8	32.359	0.648	0.735

6. DISCUSSION

The results obtained by clustering the expression data of the three benchmarks for both the numerical and semantic aspects seem to indicate that in general, the OBKML-GO algorithm provides the best results compared to the two algorithms KMeans and Optimized-KMeans. This is valid for the three validity indices (I_{CH},I_H andI_P) and when the number of clusters has not been taken into account.

This result was to be expected due to the optimization which generally leads to a better clustering compared to classical clustering methods [41,42]. Moreover, this is also because of the use of the BKM algorithm. This algorithm is thought to be more interesting than the KMeans algorithm. The obtained clusters are more homogeneous and not empty, as opposed to the KMeans algorithm, which can generate them [26].

In addition, the use of validity indices contributes to this optimization. Indeed, the WB index, which supports inter- and intra-cluster variation as well as the number K_{NC} of clusters, can contribute to the determination of these optimal clusters [28].

The Leader has the role of representative for each cluster. It allows having a minimum distance from its own group's elements and a maximum distance from the elements of the other groups.Biological knowledge also provides semantic information and therefore an interesting global result in terms of homogeneity and coherence of clusters in adequacy with the biological reality. Approaches reported in the literature that involve the semantic part, underlining the positive aspect of the fusion of the two types of data, those of numerical origin and those of semantic origin [11,12,24]. Regardless of the benchmark, and regardless of the algorithm used, the number of clusters for which the index values are the best is K_{NC}={7, 9 and 10}.

However, this does not mean that the different algorithms result in the same cluster quality since the index values are higher in the OBKML-GO algorithm compared to the other two [26].

The key points to be retained from OBKML-GO approach are the easiness of and the relative low time complexity. Indeed, as reported in [43], the use of the Bisecting KMeans algorithm during its partitioning process (in K_{NC}=2 at each iteration) has a more interesting temporal complexity $o(k-1)nI$ (where k indicates the set of clusters obtained, n the number of data points, the number of iterations) than that of KMeans which is of $o(knI)$ and also an optimized computing using the WB index.

The limitations of the proposed approach could include (i) the size of the benchmarks, which should be larger and of similar size, (ii) the analysis of a larger number of benchmarks, (iii) the evaluation of the quality of clustering by more validity indices, and (iv) the integration of the GO terms associated with the other two aspects (i.e., the cellular component (OntoCC) and molecular function (OntoMF) aspects).

7. CONCLUSION

This paper proposed an approach called OBKML-GO (Optimized Bisecting KMeans Leader with Gene Ontology) that combines data from gene expression profiles and biological knowledge derived from terms and annotations related to Gene Ontology (GO).

The performance of this approach was assessed on three real-life gene expression datasets against two KMeans and Optimized-KMeans clustering algorithms. In contrast to the other two algorithms, OBKMLGO generated the best results for the three validity indices (I_{CH},I_H andI_P) regardless of the clusters number. This is because the Bisecting algorithm is more effective than the KMeans algorithm. This is mainly due to its cluster partitioning process at the level of each iteration, by an SSE value computing. The BKM algorithm provides at the end coherent and non-empty clusters, unlike the KMeans which can generate them. The WB index calculates the inter-cluster and intra-cluster variation and the number of clusters that can contribute to obtaining these optimal clusters. Moreover, using the Leader (one representative for each cluster) allows having a minimum distance from its own group's elements and a maximum distance from the other groups' elements.Furthermore, the addition of biological knowledge that considers the information content of the child nodes improves the relevance of the semantics. As a result, the overall result is interesting in terms of homogeneity and coherence of clusters in adequacy with the biological context. Thus, the combination results for the Yeast benchmark for the best index values are I_{CH}= 33.087, I_H= 0.653, and I_P= 0.746 with γ = 0.1 and K_{NC} = 10. For the Human benchmark the best index values are I_{CH}= 34.953, I_H= 0.713, and I_P= 0.768 with γ = 0.7 and K_{NC} = 9. According to the *Arabidopsis thaliana* benchmark, the best index values are I_{CH}= 32.359, I_H= 0.648, and I_P= 0.735 with γ = 0.2 and K_{NC} = 8.

A deeper insight into OBKML-GO approach would consist to utilize other bigger and comprehensive benchmarks, to check the quality of clustering through more validity indices. The incorporation of the GO terms related to the other two facets (i.e. the cell component aspect (OntoCC) and the molecular function aspect (OntoMF)) can also be added and the semantic contribution of these aspects to gene clustering can be compared. Finally, other biological annotation resources can also be associated with the numerical expression profiles to enhance the semantic aspect.

ACKNOWLEDGMENT

The authors would like to thank the Directorate General of Science Research and Technological Development (DGRSDT), Ministry of Higher Education and Scientific Research of Algeria for their support in this work.

REFERENCES

[1] M. W. Libbrecht, and W. S. Noble, "Machine learning applications in genetics and genomics," Nature Reviews Genetics, vol. 16, no 6, pp. 321-332, 2015.

DOI: https://doi.org/10.1038/nrg3920.

[2] E. Lin, and H.-Y. Lane, "Machine learning and systems genomics approaches for multi-omics data," Biomarker research, vol. 5, no. 1, pp. 2, 2017.

DOI: https://doi.org/10.1186/s40364-017-0082-y.

[3] S. Saha, A. K. Alok, and A. Ekbal. "Use of semisupervised clustering and feature-selection techniques for identification of co-expressed genes," IEEE journal of biomedical and health informatics, vol. 20, no. 4, pp. 1171-1177, 2015.

[4] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–a decade review" Information Systems, vol. 53, pp. 16-38, 2015.

DOI: https://doi.org/10.1016/j.is.2015.04.007.

[5] G. Leale, D. H. Milone, A. E. Bayá, P. M. Granitto, and G. Stegmayer, "A novel clustering approach for biological data using a new distance based on gene ontology," in XIV Argentine Symposium on Artificial Intelligence (ASAI)-JAIIO 42 (2013), 2013.

[6] Z. M. Hira, and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," Advances in bioinformatics, vol. 2015, p.13, 2015.

DOI: https://doi.org/10.1155/2015/198363.

[7] A.G. Spampinato, and S. Cavallaro, "Meta-analysis of genomic data: Between strengths, weaknesses and new perspective," Int. J. Biomed. Data Min, vol. 5, pp. 117, 2016.

DOI: https://doi.org/10.4172/2090-4924.1000117.

[8] Alok. A. K, Saha, S. and Ekbal, A. "Semi-supervised clustering for gene-expression data in multiobjective optimization framework," International Journal of Machine Learning and Cybernetics, vol. 8, no. 2, pp. 421-439, 2017.

DOI: https://doi.org/10.1109/JBHI.2015.2451735.

[9] Z. Zareizadeh, M. S. Helfroush, A. Rahideh, and K. Kazemi, "A robust gene clustering algorithm based on clonal selection in multiobjective optimization framework," Expert Systems with Applications, vol. 113, pp. 301-314, 2018.

DOI: https://doi.org/10.1016/j.eswa.2018.06.047.

[10] G. Leale, A. E. Bayá, D. H. Milone, P. M. Granitto, and G. Stegmayer, "Inferring unknown biological function by integration of GO annotations and gene expression data," IEEE/ACM transactions on computational biology and bioinformatics, vol. 15, no. 1, pp. 168-180, 2016.

[11] J. Parraga-Alava, and M. Inostroza-Ponta, "A bi-objective clustering algorithm for gene expression data," CLEI Electronic Journal, vol 20, no. 2, pp. 1-17, 2017.

DOI: https://doi.org/10.19153/cleiej.20.2.4

[12] J. Parraga-Alava, M. Dorn, and M. Inostroza-Ponta, "A multi-objective gene clustering algorithm guided by apriori biological knowledge with intensification and diversification strategies". BioData mining, vol. 11, no. 1, pp. 16, 2018.

[13] M. B. Hossen, H. A. Siraj-Ud-Doulah, and A. Hoque, "Methods for evaluating agglomerative hierarchical clustering for gene expression data: a comparative study," Computational Biology and Bioinformatics, vol. 3, no.6, pp.88-94, 2015.

DOI: https://doi.org/10.11648/j.cbb.20150306.12.

[14] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O.P. Patel, A. Tiwari,……and C.T. Lin, "A review of clustering techniques and developments," Neurocomputing, vol. 267, pp. 664-681, 2017.

DOI: https://doi.org/10.1016/j.neucom.2017.06.053.

[15] S. Anusuya, B. DNU and Kasthuri, E. "yeast gene expression analysis using k means and FCM," International Journal of Pharma and Bio Sciences, vol. 6, no. 3, pp. B395-B400, 2015.

[16] D. Elsayad, M. E. Khalifa, A. Khalifa, and E. S. El-Horbaty, "An improved parallel minimum spanning tree based clustering algorithm for microarrays data analysis," in 2012 8th International Conference on Informatics and Systems (INFOS), pp. DE-66-DE-72. IEEE, 2012.

[17] A. Bihari, S. Tripathi, and A. Deepak, "Gene Expression Analysis Using Clustering Techniques and Evaluation Indices," Available at SSRN 3350332, 2019.

[18] H. Fyad, F. Barigou, and K. Bouamrane, "An Experimental Study on Microarray Expression Data from Plants under Salt Stress by using Clustering Methods,". International Journal of Interactive Multimedia and Artificial Intelligence, vol.6, no.2, 2020.

DOI: 10.9781/ijimai.2020.05.004 .

[19] S. Carbon, H. Dietze, S.E. Lewis, C.J. Mungall, M.C. Munoz-Torres, S. Basu,…. and H. Mi, "Expansion of the gene ontology knowledgebase and resources, " Nucleic Acids Research, vol. 45, no. D1, 2017.

DOI: https://doi.org/10.1093/nar/gkw1108.

[20] G. K. Mazandu, E. R. Chimusa, and N. J. Mulder, "Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery," Briefings in bioinformatics, vol. 18, no. 5, pp.886-901, 2017.

[21] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic similarity from natural language and ontology analysis," Synthesis Lectures on Human Language Technologies, vol. 8, no. 1, pp. 1-254, 2015.

DOI: http://doi.org/10.2200/S00639ED1V01Y201504HLT027.

[22] R. Kustra, and A. Zagdanski, "Incorporating gene ontology in clustering gene expression data," in 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06). IEEE, pp. 555-563, 2006.

[23] B. Y, Kang, S. Ko, and D.W. Kim, "SICAGO: Semi-supervised cluster analysis using semantic distance between gene pairs in Gene Ontology," Bioinformatics, vol. 26, no. 10, pp. 1384-1385, 2010.

DOI: https://doi.org/10.1093/bioinformatics/btq133.

[24] J. Meng, R. Li, and Y. Luan, "Classification by integrating plant stress response gene expression data with biological knowledge," Mathematical Biosciences, vol. 266, pp. 65-72, 2015.

[25] K. Abirami, and P. Mayilvahanan, "Performance Analysis of K-Means and Bisecting K-Means Algorithms in Weblog Data," International Journal of Emerging Technologies in Engineering Research (IJETER), vol. 4, no. 8, 2016.

[26] M.R. Mahmud, M. A. Mamun, M. A. Hossain, and M. P. Uddin, "Comparative Analysis of K-Means and Bisecting K-Means Algorithms for Brain Tumor Detection," in 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), pp 1-4, IEEE, 2018.

[27] J. Anuradha, and B. Tripathy, "Hierarchical clustering algorithm based on attribute dependency for attention deficit hyperactive disorder," International Journal of Intelligent Systems and Applications, vol. 6, no. 6, pp. 37, 2014.

[28] Q. Zhao, and P. Fränti, "WB-index: A sum-of-squares based index for cluster validity," Data & Knowledge Engineering, vol. 92, pp. 77-89, 2014.

[29] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, ... and Web Presence Working Group, "AmiGO: online access to ontology and annotation data," Bioinformatics, vol. 25, no. 2, pp. 288-289, 2009.

DOI: https://doi.org/10.1093/bioinformatics/btn615.

[30] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," PLoS computational biology, vol. 5, no. 7, 2009.

[31] C. Zhao, and Z. Wang, "GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms," Scientific reports, vol. 8, no. 1, pp. 1-10, 2018.

[32] J. Z. Wang, Z. Du, R. Payattakool, P.S. Yu, and C.F. Chen, "A new method to measure the semantic similarity of GO terms," Bioinformatics, vol. 23, no. 10, pp. 1274-1281, 2007.

DOI: https://doi.org/10.1093/bioinformatics/btm087.

[33] J. C. de Winter, S. D. Gosling, and J. Potter, "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data," Psychological methods, vol. 21, no. 3, pp. 273, 2016.

[34] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz, "The transcriptional program of sporulation in budding yeast," Science, vol. 282,no. 5389, pp. 699-705, 1998.

[35] V. R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J. Lee,….and D. Lashkari, "The transcriptional program in the response of human fibroblasts to serum," Science, vol. 283, no. 5398, pp. 83-87, 1999.

[36] P. Reymond, H. Weber, M. Damond, and E. E. Farmer, "Differential gene expression in response to mechanical wounding and insect feeding in Arabidopsis," The Plant Cell, vol. 12, no. 5, pp. 707-719, 2000.

[37] U. Maulik, A. Mukhopadhyay, and S. Bandyopadhyay, "Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes," BMC bioinformatics, vol. 10, no. 1, pp. 27, 2009.

DOI: https://doi.org/10.1186/1471-2105-10-27.

[38] S. Zerabi, S. Meshoul, A. Merniz, and R. Melal, "Towards clustering validation in big data context," in Proceedings of the 2nd international Conference on Big Data, Cloud and Applications, pp. 1-6, 2017.

[39] R. Dash, and B. B. Misra, "Performance analysis of clustering techniques over microarray data: A case study," Physica A: Statistical Mechanics and its Applications, vol. 493, pp. 162-176, 2018.

[40] K. Swapna, and M. P. Babu, "Critical Evaluation of Predictive Analytics Techniques for the Design of Knowledge Base," in International Conference on E-Business and Telecommunications. Springer, Cham. pp. 385-392, 2019

[41] D. K. Roy, and L. K. Sharma, "Genetic k-means clustering algorithm for mixed numeric and categorical data sets," International Journal of Artificial Intelligence and Applications, vol. 1, no. 2, pp. 23-28, 2010.

[42] Y. K. Lam, P. W. M. Tsang, and C. S. Leung, "PSO-based K-Means clustering with enhanced cluster matching for gene expression data," Neural Computing and Applications, vol. 22, no7-8, pp. 1349-1355, 2013.

[43] H. Gaudani, K. Lakhani, and R. Chhatrala, "Survey of Document Clustering," International Journal of Computer Science and Mobile Computing, vol. 3, no. 5, pp. 871-874, 2014.

HoudaFyad is an Assistant Professor in computer science at University of Oran 2, Algeria. She received her engineering degree in computer science department (2006) from University of Oran1. She also received her magister degree (2011) with specialization Informatique&Automatique from the same university. Currently she is preparing her PhD thesis within the Computer Science Department in the University of Oran1 with specialization Diagnostic and Decision-Making Assistance and Human Interaction Machine. In research field, she works on Machine Learning, Data mining, Bioinformatics and Ontologies.

Fatiha Barigou is a university lecturer at Computer Science Department at Université Oran 1. She is a research member of the AIR team in the LIO laboratory. She does research in Text Data Mining, Big data and Artificial Intelligence. Her current projects are Sentiment Analysis, AI, Fog and Cloud Computing in healthcare.

Karim Bouamranereceived the PhD Degree in computer science from the Oran University in 2006. He is Professor of computer Science at the same university. He is the head of computer science laboratory (LIO) and Decision and piloting system team. His current research interests deal with decision support system in maritime transportation, urban transportation system, production system, and application of bio-inspired based optimization metaheuristic. He participates in several scientific committees' international/national conferences in Algeria and others countries in the same domain and collaborates in Algerian-French scientific projects. He is co-author of more than 40 scientific publications.

Baghdad Atmani is a Professor of Computer Science at the University of Oran 1 Ahmed Benbella. His field of interests is Data Mining and Machine Learning Tools. His research is based on Knowledge Representation, Knowledge-based Systems and CBR, Data and Information Integration and Modelling, Data Mining Algorithms, Expert Systems and Decision Support Systems. His research is guided and evaluated through various applications in the field of control systems, scheduling, production, maintenance information retrieval, simulation, data integration and spatial data mining.