An Empirical Comparison of EM and K-means Algorithms for Binning Metagenomics Datasets

Comparación Empírica de los Algoritmos EM y K-medias para Binning de Conjuntos de Datos Metagenómicos

Patricio Tapia Reyes[1] Claudio Meneses Villegas[1]*

Recibido 27 de junio de 2018, aceptado 06 de agosto de 2018
Received: June 27, 2018 Accepted: August 06, 2018

ABSTRACT

Metagenomics is an area of microbiology that deals with the taxonomic classification of genomic samples taken directly from the environment. These samples are sequences of variable length and they may correspond to different species, some of which may be unknown or not previously stored in a genomic database. One of the main steps in metagenomics classification correspond to binning the sequence fragments into groups that may correspond to one species. Many approaches are used to perform binning, mainly machine learning algorithms to perform classification or clustering. This paper presents the results of an empirical evaluation of two well-known unsupervised algorithms to perform the metagenomics binning task: the EM versus the K-means algorithms. Both algorithms are tested on short and long reads of synthetic datasets, with different proportions and number of species. These empirical results show that K-means in general outperforms the EM algorithm, but EM results competitive in several of the short reads datasets used for evaluation.

Keywords: Binning, metagenomics analysis, unsupervised learning, clustering.

RESUMEN

La metagenómica es un área de la microbiología que trata con la clasificación taxonómica de muestras tomadas directamente del ambiente. Estas muestras son secuencias de largo variable que pueden pertenecer a distintas especies, algunas pueden ser desconocidas o no han sido almacenadas previamente en una base de datos genómica. Uno de los pasos principales en la clasificación metagenómica corresponde al proceso de binning de los fragmentos de secuencias en grupos que pueden corresponder a una especie. Se han usado varios acercamientos para realizar binning, principalmente algoritmos de machine learning para realizar la clasificación o agrupamiento. Este artículo presenta los resultados de una evaluación empírica de dos algoritmos no supervisados bien conocidos, para realizar la tarea de binning metagenómico: EM vs. K-medias. Ambos algoritmos son probados para secuencias largas y cortas de conjuntos de datos sintéticos, con diferentes proporciones y número de especies. Estos resultados empíricos muestran que K-medias en general tiene un mejor rendimiento que el algoritmo EM, pero los resultados de EM son competitivos cuando son probados con varios conjuntos de secuencias cortas.

Palabras clave: Binning, análisis metagenómico, aprendizaje no supervisado, agrupamiento.

[1] Departamento de Ingeniería en sistemas y computación. Universidad Católica del Norte. Av. Angamos 0610. Antofagasta, Chile. E-mail: patricio.tapia@alumnos.ucn.cl
* Corresponding Author: cmeneses@ucn.cl

INTRODUCTION

A recurrent problem in microbiology is to identify the microorganisms present in an environment, if we consider that only 1% of the microorganisms in the environment are cultivable in the laboratory [1], 99% of them should be studied by indirect methods. The main of these indirect methods is the sequencing, which consists of the ordered reading of each nucleotide molecule that makes up the DNA chain (Adenine, Thymine, Guanine, Cytosine) and its interpretation in a chain using a specific character for each type of Nucleotide (A, T, G, C). The sequencing of all the genetic material of an organism is called its Genome.

The metagenome is one of the methods of microbiology used to know the behavior of an environment, and consists of the sequencing of all the genetic material present in a sample. Among the so-called new sequencing techniques is Shotgun sequencing, which results in a collection of sequences between 70 and 200 characters long called reads. As the sample is environmental, it is unknown to which organism each reads belongs. The methods of binning are computational approaches for the grouping of these reads in genomes corresponding to their original species.

These methods may be presented as supervised or unsupervised. Supervised methods use previous information from databases to group reads into genomes, dragging the problem of classical microbiology from not knowing more than 1% of existing organisms in the environment; Unsupervised methods use extractable information from reads (e.g., character distribution, reads) to classify the sequences into genomes.

Unsupervised-type methods use clustering algorithms to determine which genome each read should correspond to. A widely used type of information is the repetition frequency of character sequences of a long l, in the literature this approach is found as a frequency of l-mers, n-grams, or k-mers. The use of l-mers is based on 2 properties [2]: i) the frequencies of l-mers of some read of a genome are linearly proportional to the abundance of such genome; ii) the frequency distribution of short l-mers is similar for similar genomes.

Related Work

A systematic search was performed with the keywords: "(metagenomic OR metagenome) AND binning AND (software OR method OR strategy OR proposal)", which resulted in the existence of at least 27 different binning methods from the 2012 through 2016 years, 23 of these methods use the l-mers frequency as classification information.

The most used value is l = 4, which is estimated to have a more stable frequency distribution for fragments of chromosomal DNA with a size of 500 to 10,000 base pairs [3], 15 of these 23 approaches are based exclusively on use of 4-mers. Depending on the author's interpretation, it may be a vector of 256 dimensions [4-9] or 136 dimensions, considering palindromic tetra-nucleotides as redundant information [10-15]. Table 1 summarizes the main features of the methods that consider 136-dimensional 4-mers frequency vectors.

This paper addresses the task of binning metagenomic data based on reads sequences from synthetic datasets. In particular, it is sought to empirically establish the performance of two widely known clustering algorithms used for the binning task of metagenomic data: the K-means algorithm and the EM algorithm. In this way, it is sought to establish a baseline with which to compare new methods to propose based on a semi-supervised approach, considering a research in development in which it is hypothesized that a semi-supervised approach may be superior in some cases and in others be competitive with respect to baseline methods.

MATERIALS AND METHODS

EM (Expectation-Maximization) Algorithm: This algorithm consists of two steps, the Expectation step (E-step), and the Maximization step (M-step). The E-step basically fills in the missing data (the class value in our case) based on the current estimation of the parameters. The M-step, which maximizes the likelihood, re-estimates the parameters. These steps are repeated until EM converges to a local minimum when the model parameters stabilize [16].

K-means Algorithm: The K-means algorithm is a clustering algorithm based on systematic partitions of data [17], and considers the following steps: i) Initially, K centroids are randomly generated; ii) The distance of each data of the dataset with respect to these centroids is measured; iii) Each data

Table 1. Articles of binning methods using 136-dimensional 4-mers frequency vectors.

Software Name	Clustering method	Autors	Year	Ref.
Metawatt	Interpolated Markov models	Marc Strous; Beate Kraft; Regina Bisdorf; Halina E. Tegetmeyer	2012	15
CONCONT	Gaussian mixture model	Johannes Alneberg; Brynjar Smari Njarnason; Ino de Brujin; Melanie Schirmer; Joshua Quick; Umer Z Ijaz; Leo Lathi; Nicholas J Loman; Anders F Anderson; Christopher Quince	2014	14
MaxBin	Expectation maximization	Yu-Wei Wu; Yung-Hsu Tang; Susannah G Tringe; Blake A Simmons; Steven W Singer	2014	13
MaxBin 2.0	Expectation maximization	Yu-Wei Wu; Blake A Simmons; Steven W Singer	2015	10
BiMeta	K-means	Le Van Vinh; Tran Van Lang; Le Thanh Binh; Tran Van Hoai	2015	11
No name	Support vector domain description models	Hou Tao; Liu Yun; Liu Fu; Wang Ke; Xie Jian	2015	12

is assigned to the group whose centroid is closest, thus forming K clusters; iv) The centroids of each group are recalculated. This process is repeated until the groups are stable, i.e., until all the groups from one iteration to the next do not change, or until a predefined number of iterations has been reached. Otherwise, the process is repeated from the step ii).

Datasets of Short and Long Reads: Twenty-three artificial datasets described by VanVinh, L. et al [11] were used, from which groups of reads were generated using the executable code of the BiMeta algorithm proposed in [11]. From these groups, 4-mers frequency vectors of 136 dimensions are generated to determine the species belonging to some group. It was considered that each group belonged to the species of the first read entered in the seed subgroup, the proportions of groups with respect to the genomes that they represent are different from the reads, since a group can contain from 1 to 50 different reads. The number of groups obtained from each artificial dataset and their distribution by genome is described in Table 2.

For the organization and format of the data, the software SPSS Statistics release 23.0.0.0 64-bit edition and arff Viewer of WEKA 3.8 [18] were used. Each dataset consists of a number of cases equal to the number of groups generated from the reads by the algorithm of BiMeta, retaining the name of the artificial dataset from which the reads come. Each case contains 138 attributes: the group number (from 0 to n, where n = number of groups), species of the group, and standardized frequencies of 4-mers for the whole group, labeled from f1 to f136 (see Table 3).

Evaluation Metrics: For the evaluation of the behavior of the algorithms for the classification of species, we used as metrics Accuracy, Recall, Precision and F-measure, calculated according to their standard definition [17]. In addition, the characterization of each cluster generated in terms of the measurement of Euclidean distances between the frequency vectors of each cluster was analyzed, assuming that the inter-cluster distance should be as large as possible and the lowest possible intra-cluster distance for 2 distinct clusters, generated from the same dataset by the same algorithm.

RESULTS

Table 4 summarizes the behavior of the two algorithms used to perform binning of the metagenomics data. The R1-R9 datasets contain long reads sequences, whereas the datasets L1-L6 and S1-S8 contain sequences of short reads.

According to these results it can be observed that both algorithms are competitive with each other for long reads datasets R3, R4, R5 and R8, and short reads datasets S6 and S7. For the case of the datasets R1, R2, R6 and R7 of long reads, and L1-L6, S1-S5 and S8 of short reads, the algorithm Simple K-Means (SKM) outperformed the EM algorithm. That is, in the case of datasets with long reads, in 50% of them SKM outperforms EM, while in the remaining 50% both algorithms present competitive results without

Table 2. Description of datasets used in this work, including datasets of groups of reads.

Dataset	Number of genomes	Number of reads	Genomer ratio	Number of groups	Genome Ratio
R1	2	82960	1:1	3171	1:1
R2	2	77293	1:1	1010	1:1
R3	2	93267	1:1	3895	1:1
R4	2	33457	1:1	1049	1:1
R5	2	40043	1:1	1337	1:1
R6	2	70550	1:1	2781	1:1
R7	3	290473	1:1:8	5842	1:1:1
R8	3	374830	1:1:8	8887	1:3:4
R9	6	588258	1:1:1:2:14	15225	1:1:1:3:4:4
L1	2	176688	1:1	3414	1:1
L2	2	259568	1:2	3362	1:1
L3	2	342448	1:3	3475	1:1
L4	2	425328	1:4	3555	1:1
L5	2	508209	1:5	3651	1:1
L6	2	591089	1:6	3665	1:1
S1	2	96367	1:1	1299	1:1
S2	2	195339	1:1	2481	1:2
S3	2	338725	1:1	3936	1:3
S4	2	375302	1:1	4784	1:5
S5	3	325400	1:1:1	4193	1:2:2
S6	3	713388	1:2:3	6546	1:2:5
S7	5	1653550	1:1:1:4:4	9852	1:2:2:3:3
S8	5	456224	3:5:7:9:11	12267	1:3:3:3:5

Table 3. Example of a dataset of group of reads, represented as frequency vectors.

Group	Specie	f1	f2	f3	f4	f5	f6	f7	f8	f9	…	f136
0	specie1	0.018	0.000	0.008	0.010	0.005	0.008	0.022	0.017	0.013	…	0.004
1	specie2	0.012	0.000	0.009	0.013	0.006	0.007	0.031	0.014	0.013	…	0.006

a statistically significant difference between them, in terms of the metrics used for evaluation. In general, the datasets with more than 3 species generated a results under 50% in F-measure for both algorithms, reaching the point where the amount of cluster generated by the algorithms was less than the amount of species existing in the dataset.

When viewing the results of Table 4 in the form of line graphs, we see a trend of SKM to present slightly higher results than those of EM for F-measure (Figure 1) and for Accuracy (Figure 2), mainly in datasets L1 to L6, where the vertical lines represent the standard deviation.

Table 5 shows the intracluster and intercluster distance obtained for the different datasets used, for both algorithms. The intracluster distance is a quantitative measure of the degree of average closeness of the cases in each group. The clustering algorithms try to minimize this value. In contrast, the intercluster distance is a quantitative measure of the degree of average remoteness of the centroids of the different groups. Clustering algorithms attempt to maximize this value. When measuring the mean intracluster and intercluster Euclidean distance, for both EM and SimpleKMeans, two datasets showed statistically significant differences for the intracluster distance (R3 and S6) and a dataset presented a statistically significant difference in its intercluster distances (S5). This means that, in general, in terms of cohesion (intracluster distance) and coupling (intercluster distance) of clusters resulting from both algorithms, no statistically significant differences were found between the two algorithms. It should be noted that the calculation of the Euclidean intercluster distance in datasets with only two clusters is a simple summation of the distances of each of the frequencies that make up the vector of 4-mer frequencies of each cluster, therefore this does not present sufficient information to calculate a standard deviation

Table 4. Results of specie classification based on Simple K-means and EM algorithms, over 23 synthetic datasets. R1-R9 correspond to datasets of long reads; S1-S8 and L1-L6 correspond to datasets of short reads.

Data-set	Groups	Species	SimpleKMeans algorithm				Expectation Maximization algorithm			
			Accuracy	Mean precision ± standard deviation	Mean Recall ± standard deviation	Mean F-measure ± standard deviation	Accuracy	Mean precision ± standard deviation	Mean Recall ± standard deviation	Mean F-measure ± standard deviation
R1	3171	2	0.928	0.933 ± 0.07	0.929 ± 0.08	0.928 ± 0.06	0.771	0.839 ± 0.22	0.774 ± 0.32	0.757 ± 0.23
R2	1010	2	0.831	0.830 ± 0.01	0.830 ± 0.03	0.830 ± 0.02	0.534	0.526 ± 0.03	0.522 ± 0.28	0.487 ± 0.16
R3	3895	2	0.823	0.824 ± 0.05	0.824 ± 0.05	0.823 ± 0.04	0.808	0.808 ± 0.04	0.809 ± 0.03	0.808 ± 0.03
R4	1049	2	0.994	0.994 ± 0.00	0.994 ± 0.00	0.994 ± 0.00	0.995	0.995 ± 0.01	0.995 ± 0.01	0.995 ± 0.01
R5	1337	2	0.975	0.975 ± 0.01	0.974 ± 0.01	0.974 ± 0.01	0.980	0.980 ± 0.00	0.980 ± 0.00	0.980 ± 0.00
R6	2781	2	0.968	0.964 ± 0.04	0.972 ± 0.03	0.967 ± 0.03	0.818	0.847 ± 0.21	0.844 ± 0.22	0.818 ± 0.18
R7	5842	3	0.737	0.744 ± 0.22	0.753 ± 0.18	0.710 ± 0.18	0.612	0.608 ± 0.14	0.691 ± 0.28	0.605 ± 0.20
R8	8887	3	0.920	0.910 ± 0.03	0.889 ± 0.10	0.895 ± 0.07	0.934	0.907 ± 0.09	0.948 ± 0.08	0.922 ± 0.08
R9	15225	6	0.823	0.755 ± 0.38	0.716 ± 0.39	0.011 ± 0.37	0.754	0.723 ± 0.41	0.664 ± 0.38	0.693 ± 0.38
L1	3414	2	0.973	0.973 ± 0.07	0.972 ± 0.01	0.973 ± 0.08	0.524	0.512 ± 0.05	0.510 ± 0.28	0.469 ± 0.16
L2	3362	2	0.977	0.976 ± 0.02	0.977 ± 0.01	0.976 ± 0.01	0.902	0.921 ± 0.10	0.893 ± 0.14	0.898 ± 0.10
L3	3475	2	0.979	0.979 ± 0.02	0.980 ± 0.02	0.979 ± 0.01	0.899	0.916 ± 0.10	0.895 ± 0.14	0.897 ± 0.10
L4	3555	2	0.976	0.976 ± 0.02	0.977 ± 0.02	0.976 ± 0.02	0.899	0.915 ± 0.11	0.896 ± 0.14	0.897 ± 0.10
L5	3651	2	0.975	0.975 ± 0.02	0.976 ± 0.02	0.975 ± 0.02	0.899	0.913 ± 0.10	0.896 ± 0.13	0.897 ± 0.09
L6	3665	2	0.973	0.973 ± 0.01	0.973 ± 0.01	0.973 ± 0.01	0.897	0.910 ± 0.11	0.896 ± 0.13	0.896 ± 0.10
S1	1299	2	0.978	0.978 ± 0.01	0.979 ± 0.01	0.978 ± 0.01	0.850	0.880 ± 0.17	0.857 ± 0.20	0.848 ± 0.15
S2	2481	2	0.839	0.831 ± 0.06	0.835 ± 0.03	0.832 ± 0.04	0.576	0.535 ± 0.12	0.529 ± 0.32	0.472 ± 0.20
S3	3936	2	0.965	0.945 ± 0.07	0.971 ± 0.02	0.957 ± 0.04	0.818	0.797 ± 0.29	0.875 ± 0.17	0.798 ± 0.20
S4	4784	2	0.986	0.963 ± 0.05	0.986 ± 0.00	0.974 ± 0.03	0.991	0.980 ± 0.02	0.988 ± 0.01	0.984 ± 0.02
S5	4193	3	0.815	0.815 ± 0.05	0.808 ± 0.15	0.802 ± 0.10	0.608	0.568 ± 0.26	0.602 ± 0.35	0.459 ± 0.28
S6	6546	3	0.986	0.987 ± 0.00	0.975 ± 0.03	0.981 ± 0.02	0.989	0.987 ± 0.01	0.984 ± 0.01	0.985 ± 0.01
S7	9852	5	0.559	0.506 ± 0.26	0.582 ± 0.38	0.197 ± 0.31	0.527	0.469 ± 0.25	0.540 ± 0.39	0.306 ± 0.31
S8	12267	5	0.579	0.618 ± 0.20	0.598 ± 0.28	0.496 ± 0.23	0.536	0.546 ± 0.15	0.545 ± 0.29	0.443 ± 0.22

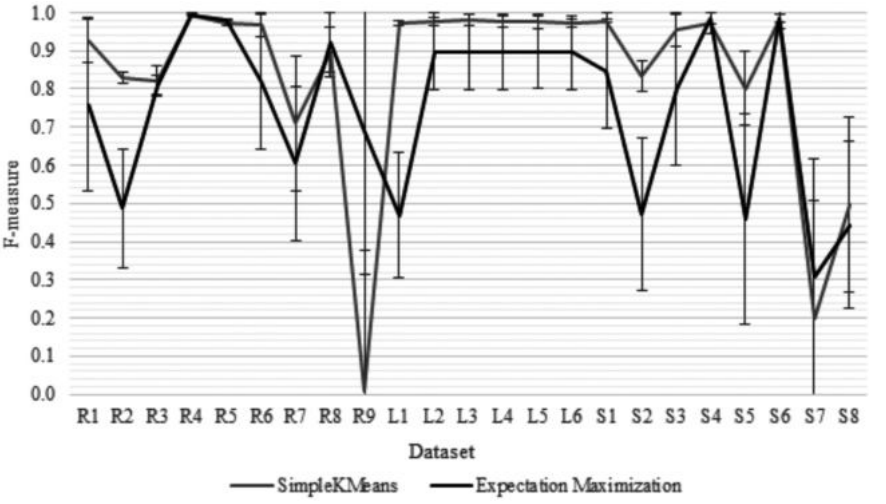

Figure 1. F-measure measurement for each dataset with the SimpleKMeans and Expectation Maximization algorithms.

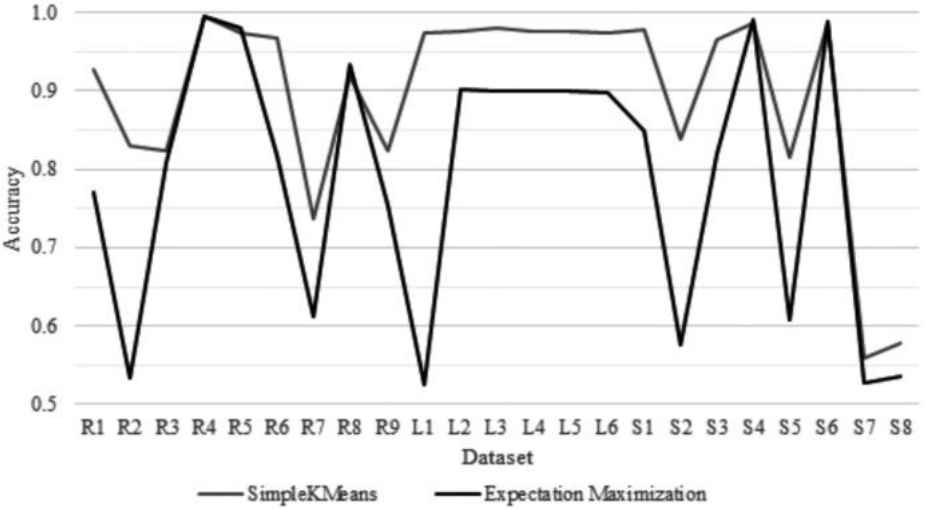

Figure 2. Accuracy measurement for each dataset with the SimpleKMeans and Expectation Maximization algorithms.

Table 5. Comparison of resulting clusters generated by K-means and EM algorithm, based on the intracluster and intercluster average distance.

Dataset	Clusters	SimpleKMeans algorithm		Expectation Maximization algorithm	
		Mean IntraCluster Distance ± standard deviation	Mean InterCluster Distance ± standard deviation	Mean IntraCluster Distance ± standard deviation	Mean InterCluster Distance ± standard deviation
R1	2	0.0252541 ± 0.0085317	0.0156655	0.0270378 ± 0.0007638	0.0206100
R2	2	0.0259683 ± 0.0122143	0.0101579	0.0205995 ± 0.0005631	0.0213644
R3	2	0.0874562 ± 0.0086373	0.0280106	0.0287656 ± 0.0001663	0.0279549
R4	2	0.0272987 ± 0.0002738	0.0346150	0.0272654 ± 0.0007450	0.0347295
R5	2	0.0995836 ± 0.0212610	0.0351438	0.0997911 ± 0.0214979	0.0352843
R6	2	0.0294728 ± 0.0102472	0.0346440	0.0294728 ± 0.0102472	0.0346440
R7	3	0.0323306 ± 0.0085582	0.0346425 ± 0.0125191	0.0336210 ± 0.0026424	0.0351133 ± 0.0096335
R8	3	0.0299328 ± 0.0116063	0.0771026 ± 0.0314250	0.0317694 ± 0.0065680	0.0738762 ± 0.0300574
R9	6	0.0336484 ± 0.0045021	0.0572017 ± 0.0176787	0.0323857 ± 0.0062913	0.0598808 ± 0.0204103
L1	2	0.1571469 ± 0.0242263	0.0031724	0.1526578 ± 0.0530349	0.0406151
L2	2	0.1543704 ± 0.0370218	0.0348746	0.1494912 ± 0.0466247	0.0407553
L3	2	0.1545883 ± 0.0351402	0.0345795	0.1495891 ± 0.0464405	0.0409151
L4	2	0.1546097 ± 0.0361639	0.0346760	0.1491224 ± 0.0487726	0.0406383
L5	2	0.1526745 ± 0.0379499	0.0341652	0.1470982 ± 0.0479603	0.0405508
L6	2	0.1541340 ± 0.0356566	0.0343172	0.1490432 ± 0.0440333	0.0407737
S1	2	0.0665941 ± 0.0174995	0.0379279	0.0635155 ± 0.0029405	0.0474583
S2	2	0.0666234 ± 0.0305250	0.0058861	0.0607143 ± 0.0015441	0.0237358
S3	2	0.0740767 ± 0.0217884	0.0467122	0.0605464 ± 0.0028861	0.0691053
S4	2	0.0587966 ± 0.0011660	0.1016099	0.0585097 ± 0.0027094	0.1000344
S5	3	0.0460514 ± 0.0287384	0.0288425 ± 0.0126795	0.0373041 ± 0.0019136	0.0959204 ± 0.0358429
S6	3	0.0401389 ± 0.0007670	0.0386924 ± 0.0132037	0.0379667 ± 0.0005475	0.0944913 ± 0.0347768
S7	5	0.0361684 ± 0.0297462	0.0254207 ± 0.0135238	0.0298624 ± 0.0169002	0.0370670 ± 0.0142258
S8	5	0.0361684 ± 0.0297462	0.0254207 ± 0.0135238	0.0298624 ± 0.0169002	0.0370670 ± 0.0142258

When visualizing the results of table 5 in the form of line graphs, we do not observe distinctly different results between the intracluster distance for both algorithms (Figure 3), except for datasets R3 and S4, in which SKM is shown to be significantly higher, presenting a smaller distance between the data present in each cluster. On the other hand, the EM algorithm generated clusters with an intercluster distance slightly superior to the clusters generated with SKM (Figure 4). However, these results are not decisive because they do not have enough information to calculate if they are significantly different from each other

CONCLUSIONS AND FUTURE WORK

This article presents empirical results to establish a baseline regarding the application of unsupervised learning methods to perform the binning task in the classification of metagenomic data. The data used to evaluate the K-means and EM algorithms correspond to 9 datasets of long reads sequences and 14 datasets of short reads sequences, both types of data were generated in a synthetic way. The preprocessing considered the calculation of repeating frequency vectors of sequences of characters of a long l.

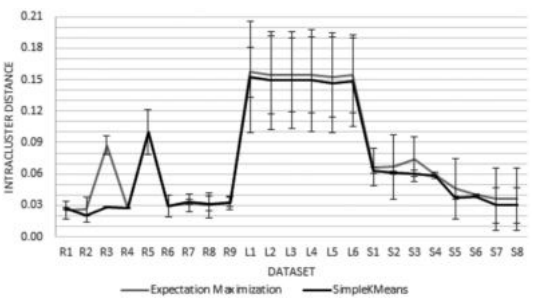

Figure 3. Measurement of the intracluster distance for each dataset with the SimpleKMeans and Expectation Maximization algorithms, vertical lines represent the standard deviation.

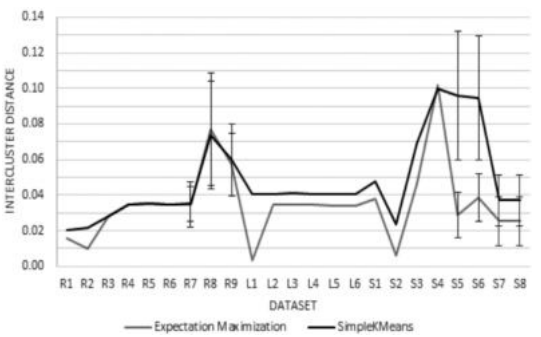

Figure 4. Measuring the intercluster distance for each dataset with the SimpleKMeans and Expectation Maximization algorithms, the vertical lines represent the standard deviation.Maximization algorithms, vertical lines represent the standard deviation.

The application of the datasets to both clustering algorithms resulted in a partition of the frequency vectors representing a group of reads that was evaluated in terms of the cohesion and coupling of the clusters generated by each algorithm and in terms of accuracy, Precision, recall, and F-measure. Regarding the evaluation in terms of cohesion and coupling, although there are differences between the average values generated by both methods, these differences are not statistically significant, indicating that in terms of these two properties (cohesion and coupling) of the Clusters generated both algorithms are competitive with each other. Regarding the classification of the reads groups in their corresponding species, the K-means algorithm presented a superior behavior with respect to the EM algorithm, in particular for sequences of short reads. For long reads sequences both algorithms were generally competitive with each other.

One of the limitations of these algorithms, when evaluating metrics such as Recall, Precision, Accuracy and F-measure, is the dependence on the correct prior classification of the data. In this case, it is possible to improve the quality of the evaluations by considering the number of reads that are contained in each group generated by the BiMeta algorithm [11], used to generate the frequency vectors associated with each group.

This work is part of the previous analysis of the behavior of non-supervised clustering algorithms for the grouping of readings coming from metagenomes, with the subsequent objective of being compared empirically with semi-supervised methods as part of a research in development. Thus, along with extending the empirical evaluation to other unsupervised algorithms, a semi-supervised approach will be designed which is expected to be better in this domain and competitive in others with respect to the results obtained with the algorithms included in the baseline.

REFERENCES

[1] J.A. Eisen. "Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes". PLOS Biology. Vol. 5 Issue 3, p. e82. 2007.

[2] S. Karlin, I. Ladunga and B.E. Blaisdell. "Heterogeneity of genomes: measures and values". Proceedings of the National Academy of Sciences. Vol. 91, Issue 26, pp. 12837-12841. 1994.

[3] Zhou F., V. Olman and Y. Xu, "Barcodes for genomes and applications". BMC Bioinformatics. Vol. 9, pp. 546-546. 2008.

[4] T.K. Moon. "The expectation-maximization algorithm". IEEE Signal Processing Magazine. Vol. 13, Issue 6, pp. 47-60. 1996.

[5] Liu Y., F. Liu, T. Hou and K. Wang. Unsupervised binning of metagenomic datasets using cluster size insensitive fuzzy c-means method. in 35th Chinese Control Conference, CCC 2016. 2016. IEEE Computer Society.

[6] Hou T., Y. Liu, J. Xue, M. Li and F. Liu. Taxonomic classification DNA fragment of metagenome with a novel model. in 35th Chinese Control Conference, CCC 2016. 2016. IEEE Computer Society.

[7] Zhang R.C., Z.Z. Cheng, J.H. Guan and S.G. Zhou, "Exploiting topic modeling to boost metagenomic reads binning". Bmc Bioinformatics. Vol. 16, p. 10. 2015.

[8] Liao R., R. Zhang, J. Guan and S. Zhou, "A new unsupervised binning approach for metagenomic sequences based on N-grams and automatic feature weighting". IEEE/ACM Trans. Comput. Biol. Bioinformatics. Vol. 11, Issue 1, p. 42-54. 2014.

[9] Reddy R.M., M.H. Mohammed and S.S. Mande, "TWARIT: An extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences". Gene. Vol. 505, Issue 2, pp. 259-265. 2012.

[10] Wu Y.W., B.A. Simmons and S.W. Singer, "MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets". Bioinformatics. Vol. 32, Issue 4, pp. 605-607. 2015.

[11] VanVinh, L., T. Van Lang, L.T. Binh and T. Van Hoai, "A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads". Algorithms for Molecular Biology. Vol. 10, Issue 1. 2015.

[12] Tao H., L. Yun, L. Fu, W. Ke and X. Jian. Binning DNA fragment of metagenome using a novel model. in 27th Chinese Control and Decision Conference, CCDC 2015. 2015. Institute of Electrical and Electronics Engineers Inc.

[13] Wu Y.W., Y.H. Tang, S.G. Tringe, B.A. Simmons and S.W. Singer, "MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm". Microbiome. Vol. 2, p. 18. 2014.

[14] Alneberg J., B.S. Bjarnason, I. De Bruijn, M. Schirmer, J. Quick, U.Z. Ijaz, L. Lahti, N.J. Loman, A.F. Andersson and C. Quince, "Binning metagenomic contigs by coverage and composition". Nature Methods. Vol. 11, Issue 11, pp. 1144-1146. 2014.

[15] Strous M., B. Kraft, R. Bisdorf and H.E. Tegetmeyer, "The binning of metagenomic contigs for microbial physiology of mixed cultures". Frontiers in Microbiology, 2012. 3(DEC).

[16] Dempster A., N. Laird and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological). Vol. 39, Issue 1, pp. 1-38. 1977.

[17] Liu B., Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. 2011, Berlin, Heidelberg: Springer Berlin Heidelberg. Pp. 136-143.

[18] Hall M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA data mining software: an update". SIGKDD Explor. Newsl. Vol. 11, Issue 1, pp. 10-18. 2009.