Evaluation of Clustering Algorithms: A Case Study

Guillaume Artignan, Mountaz Hascoët

RESEARCH REPORT

EVALUATION OF CLUSTERING ALGORITHMS: A CASE STUDY

Guillaume Artignan

Mountaz Hascoët

Guillaume Artignan and Mountaz Hascoët. 2010. Evaluation of clustering algorithms: a Case Study. Research Report.

Evaluation of clustering algorithms: a Case Study

Mountaz Hascoët Guillaume Artignan
LIRMM, UMR5506, Univ. Montpellier II
161, rue Ada, 34090 Montpellier France
{mountaz,artignan}@lirmm.fr

ABSTRACT

In many situations, the choice of the most appropriate algorithm for clustering can turn into a real dilemma. Numerical criteria have been proposed to evaluate the quality of the results of clustering algorithms. However, so many different criteria have been proposed that the dilemma is even worsen. Most quality indices reveal different aspects of the quality of the results and hide others. The aim of this paper is to help with the understanding of this domain and to facilitate the comparison and the choice of clustering algorithm. Our proposal consists in studying both evaluation criteria and clustering algorithms. We start by discussing a selected set of representative criteria, and further conduct a case study on a large set of real data, measuring not only the quality of different representative clustering algorithms but also the impact of each criterion on the ranking of the algorithms. By providing both analytical and empirical results, we hope to clarify the field and facilitate designers choices.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Clustering
General Terms: Experimentation, Data Clustering, Analysis, Comparison.
Keywords: Experimentation, Data Clustering, Analysis, Comparison.

1. INTRODUCTION

In information retrieval, clustering can be used at various stages. Either as a post treatment for a search engine to cluster the results instead of displaying ranked lists, or as a tool to automatically extract thesauri from a set of documents. It can also be used to automatically organize a collection of documents into a catalog or a directory. In all cases, the designer eager to perform clustering has to choose amongst thousands of algorithms. "There is no best clustering algorithm" said Jain in his recent review on that subject [5]. The choice of the appropriate algorithm for one purpose is multi-factorial in nature. Amongst all possible important factors that would impact the choice, the quality or nature of the results is probably the most obscure and difficult to evaluate. Very few efforts have been made to help non clustering experts to understand what is at stake and how to compare the quality of various clustering algorithms.

Our aim is to provide such elements. Our approach is based on graph representation which help simplifying and unifying notations. Most current data sets used in information retrieval can be conceptually represented by graphs.

2. QUALITY CRITERIA

Notations. A graph G is composed of a set of nodes and a set of edges that represent links between nodes. We note E (resp. N) the number of edges (resp. Nodes). Applying clustering to G consists in getting K clusters $\{C_1...C_K\}$ that are K subsets of the set of nodes of G. For all clusters C_i and C_j, we note (a) E_i, (b) E'_i, (c) E_{ij} the number of edges (a) in C_i, (b) outgoing of C_i, (c) between C_i and C_j. We note N_i the number of nodes in a cluster C_i. We note $E^{\infty}_{N_i}$ (resp. E^{∞}_{N}) the maximal number of possible edges in C_i (resp. in G). For an undirected graph we have : $E^{\infty}_{N_i} = N_i(N_i-1)/2$. For a directed graph we have : $E^{\infty}_{N_i} = N_i(N_i-1)$.

The rest of this section introduces six different evaluation criteria.

Cut [3] is computed as the sum for each cluster of a ratio of the number of extra-edges over the number of intra edges.

$$Cut(G) = \sum_{i=1}^{K} \frac{E'_i}{E_i} = \sum_{i=1}^{K} \frac{A}{B} \quad (1)$$

Perf [3] takes into account (A) the number of links between clusters, as well as (B) the number of couples grouped in the same cluster but without a link between them. Perf further computes a ratio of bad links (undesired links (A) plus missing links (B)) to the total number of possible links (C).

$$Perf(G) = 1 - \frac{A+B}{C} = 1 - \frac{\sum_{i<j}^{K} E_{ij} + \sum_{i=1}^{K}(E^{\infty}_{N_i} - E_i)}{E^{\infty}_{N}} \quad (2)$$

Cond [3] criterion is an average over the conductance of each pair of clusters. The conductance of a pair of cluster Ci and Cj is the proportion of edges between Ci to Cj divided by the minimum number of edges within Ci and Cj.

$$Cond(G) = \frac{\sum_{i<j}^{K} \frac{E_{ij}}{min(E_i,E_j)}}{K(K-1)/2} = \frac{\sum_{i<j}^{K} A}{B} \quad (3)$$

Cov [3] is the ratio of (A) edges in clusters to (B) the total edges in the graph.

$$Cov(G) = \frac{\sum_{i=1}^{K} E_i}{E} = \frac{\sum_{i=1}^{K} A}{B} \quad (4)$$

MQ [3] MQ is a difference between (A) the average intra-cluster edge density and (B) the inter-cluster edge density. Therefore it varies between -1 and +1, and highest values correspond to best clustering results.

$$MQ(G) = \frac{\sum_{i=1}^{K}(E_i/E^{\infty}_{N_i})}{K} - \frac{\sum_{i<j}^{K}(E_{ij}/N_iN_j)}{(K(K-1))/2} = A - B \quad (5)$$

Mod [4] can be considered as a measure of the density of intra-edges corrected by the density of extra-edges. Therefore, the highest values for Mod correspond to best clustering results.

$$Mod(G) = \sum_{i=1}^{K} \frac{Ei}{E} - \left(\frac{Ei + (Ei'/2)}{E} \right)^2 = \sum_{i=1}^{K} A - B^2 \quad (6)$$

For most criteria, high values indicate best clustering quality, except for both cut and cond. In the case of cut and cond, low values indicate best clustering results.

3. CASE STUDY

The dataset chosen for our case study is a lexical networks composed of 111 701 nodes and 620 043 edges [6].

We have chosen six clustering algorithms amongst the most representative approaches: CNM [4] is an agglomerative algorithm, BGLL [2] is a multilevel algorithm, CMJA [1] is a divisive algorithm, InfoMap [8] is a random-walk algorithm, LinLog [7] is a layout algorithm assigning a position and a cluster for each node and K-Means [9] a *partitional* algorithm (cf. rows Tab.1).

Tab 1 summarizes the results of our analysis. The five first lines correspond to the five first algorithms; the five next lines correspond to variations of K-means for different values of K. The five last lines show the results of a random clustering and make a sort of control test for the number of clusters resulting from tested algorithms.

Table 1: Compared algorithms using quality indices

Algo.	K	Mod	Cut	Perf.	Cond	Cov.	MQ
CNM	595	0.50795	681	0.8064	0.0038	0.7836	0.4684970
BGLL	34*	0.56433	43	0.9504	0.0579	0.6245	0.0100815
CMJA	1804*	0.00021	1914*2	0.0323	0.0011*2	0.9958	0.9963701*2
InfoMap	4678	0.47631	9621*2	0.9976	0.0008*2	0.4790	0.0000825*2
LinLog	31	0.59714	32	0.9398	0.0526	0.6835	0.0511209
595-M	595	0.32770	2831	0.9977	0.0108	0.3302	0.0126973
34-M	34	0.16104	400	0.9692	0.5489	0.2048	0.0003049
1804-M	1802	0.27958	15145*2	0.9991	0.0072*2	0.2805	0.0740601*2
4678-M	4661	0.23371	65839*2	0.9995	0.0050*2	0.2341	0.0894950*2
31-M	31	0.49033	57	0.9652	0.0747	0.5289	0.0009109
595-Rand	595	0.00012	785760*2	0.9982	2.8708*2	0.0019	0.0000118
34-Rand	34	-0.00031	2277	0.9705	2.1369	0.0291	-0.0000004
1804-Rand	1804	-0.00010	1194451*2	0.9993	0.3798*2	0.00005	-0.0000008*2
4678-Rand	4678	-0.00007	1232754*2	0.9997	0.0566*2	0.00022	0.0000040*2
31-Rand	31	-0.00011	1875	0.9677	2.1251	0.03219	0.0000002

* Clustering depending of users' parameters

*2 Normally these result are undetermined processing a division by zero

The first important result is the important variation in terms of number of clusters. Algorithms fall in at least 4 categories: (1) LinLog that results in low number of clusters, (2) InfoMap that has a high number of clusters, (3) CNM that has a relatively average number of clusters and (4) CMJA, BGLL and K-Means in which the number of clusters is a parameter of the algorithm.

The second important result is the instability of the criteria reported in table 1. Some algorithms that rank best according to a given criterion can be worse according to another. It is the case for CMJA, for example, that is best in terms of MQ and worse in terms of Perf. Figure 1 shows how the criteria relate. Each criterion is depicted as a node and the number close to the link between two criteria indicates the proportion of similarly ranked algorithms between the two criteria. It shows, for example, that with Perf and Cut, only 13% of the items ranks are consistent with the two criteria.

The third important result lies in the limits of the criteria that were illustrated by this case study. We have identified at least five important problems shared at various degrees amongst criteria: (1) the number of resulting clusters wrongly impacts the value of the criterion (2) the density of the original dataset impacts the discriminating power of the criterion (3) some criteria do not compute for extreme situations like, for instance, singleton clusters, (4) some criteria are redundant while other are contradictory, and (5) the identification of undesirable clusters such as very small clusters, very big clusters, of very sparse clusters s not fully accounted for.

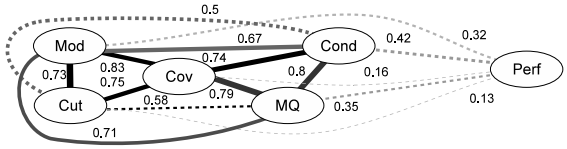

Figure 1: The similarity graph of measures

Cut falls in the first category of limits and strongly favor algorithms with low numbers of clusters. It even ranks two random algorithms (34-Rand and 31-rand) better than one of the published algorithm (InfoMap) mainly because of the difference in terms of cluster numbers. Cond, conceptually close, to cut successfully accounts for this and is not too much influenced by the number of clusters. However, Cond falls in the third category of limits. Perf can be very confusing and seems to fail at capturing the quality of a clustering for some real datasets and even ranked first the random clustering.

Cov is the most consensual of all criteria. Mod, Cut, MQ and Cond share more than 70% of the ranking of Cov, which indicates that Cov captures most of what is captured by the other four criteria. Cov is also very simple to interpret so it might be ideal for non expert users. However cov is affected by first and fifth limits above. Mod is more accurate than cov and tend to better avoid the pitfalls mentionned above.

4. CONCLUSION

The analysis of different published criteria and clustering algorithms both analytically and empirically shows that simplification is possible. Our conclusion is that using Cov or Mod criteria would capture most of what other criteria would capture of the quality of a clustering and that other important aspects such as the number of clusters or the distribution of nodes in clusters and the number of undesirable clusters are factors that can be used in conjonction with Cov or Mod in the analysis of clustering.

5. REFERENCES

[1] D. Auber et al. Multiscale visualization of small world networks. In *IV*, 2003.

[2] V. Blondel et al. Fast unfolding of communities in large networks. *JSM*, 2008.

[3] F. Boutin et al. Cluster validity indices for graph partitioning. In *IV*, 2004.

[4] A. Clauset et al. Finding community structure in very large networks. *Phys. Rev.*, 2004.

[5] A. Jain. Data clustering: 50 years beyond K-means. *PRL*, 2010.

[6] M. Lafourcade. Making people play for lexical acquisition with the jeuxdemots prototype. 2007.

[7] A. Noack. An energy model for visual graph clustering. In *GD*, 2004.

[8] M. Rosvall et al. Maps of random walks on complex networks reveal community structure. *PNAS*, 2008.

[9] H. Steinhaus. Sur la division des corp materiels en parties. *BAPS*, 1956.