Indonesian Journal of Electrical Engineering and Computer Science
Vol. 5, No. 2, February 2017, pp. 410 ∼ 415
DOI: 10.11591/ijeecs.v5.i2.pp410-415

An Empirical Comparison of Latest Data Clustering Algorithms with State-of-the-Art

Xianjin Shi , Wanwan Wang , and Chongsheng Zhang*
Henan University
No.1 JinMing Street, 475001 KaiFeng, China, Tel: +86 13837150021
*Corresponding author, e-mail: chongsheng.zhang@yahoo.com

Abstract

Clustering technology has been applied in numerous applications. It can enhance the performance of information retrieval systems, it can also group Internet users to help improve the click-through rate of on-line advertising, etc. Over the past few decades, a great many data clustering algorithms have been developed, including K-Means, DBSCAN, Bi-Clustering and Spectral clustering, etc. In recent years, two new data clustering algorithms have been proposed, which are affinity propagation (AP, 2007) and density peak based clustering (DP, 2014). In this work, we empirically compare the performance of these two latest data clustering algorithms with state-of-the-art, using 6 external and 2 internal clustering validation metrics. Our experimental results on 16 public datasets show that, the two latest clustering algorithms, AP and DP, do not always outperform DBSCAN. Therefore, to find the best clustering algorithm for a specific dataset, all of AP, DP and DBSCAN should be considered. Moreover, we find that the comparison of different clustering algorithms is closely related to the clustering evaluation metrics adopted. For instance, when using the Silhouette clustering validation metric, the overall performance of K-Means is as good as AP and DP. This work has important reference values for researchers and engineers who need to select appropriate clustering algorithms for their specific applications.

Keywords: *Affinity Propagation, Density peak based clustering, Clustering Evaluation*

1. Introduction

Clustering or cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters) [1]. Clustering has been widely used in many applications, such as discovering customer groups based on their purchase behaviours to design targeted advertisements, identifying co-regulated genes to provide a genetic fingerprint for various diseases, differentiating between different types of tissue and blood in medical images, etc.

Over the past few decades, considerable research effort has been put into the development of new data clustering algorithms [2,3,4,5]. Among them, K-Means, DBSCAN [6], Bi-Clustering [7] and Spectral clustering [8] are the very well-known ones. K-Means is by far the most popular clustering tool used in scientific and industrial applications. Starting with random centroids, K-Means clustering iteratively re-assigns each data point to the nearest centroid, then computes a new centroid for each group of data points having the same centroid, then again, allocates each data point to the nearest centroid. DBSCAN [6] is a classic density-based clustering algorithm, it can discover clusters of arbitrary shape. Bi-Clustering and Co-Clustering [7] allow overlap between clusters, this class of algorithms have been widely used in bioinformatics. Spectral clustering [8] techniques perform dimensionality reduction before clustering, by utilizing the eigenvalues of the similarity matrix of the data.

In recent years, two novel and famous data clustering algorithms have been proposed. The first clustering algorithm is affinity propagation (hereafter referred to as AP) [9], which was published in *Science* in 2007. Its highlight is that it does not require users to specify the number of clusters. AP alternates two message passing steps: one is how well a data point is to serve

Received November 16, 2016; Revised January 6, 2017; Accepted January 19, 2017

as the cluster center for another data point; the other step takes into account other data points' preference for a data point to be a cluster center.

The second (and the latest data clustering) algorithm, published in *Science* in 2014, is the density peak based clustering algorithm (hereafter referred to as DP) [10]. It is based on the idea that the cluster centers are characterized by a higher density than their neighbours and by a relatively large distance from points with higher densities [10]. For each data point, DP computes its local density and its distance to points of higher density. Top-ranked data points in both metrics will be selected as cluster centers.

Overwhelmed with so many clustering algorithms, especially with the arrival of the two new clustering algorithms, a question which is naturally raised is: which is the best data clustering algorithm? Can the two latest data clustering algorithms, AP and DP, truly outperform state-of-the-art?

Problem Statement. In this work, we will empirically compare the performance of the two latest data clustering algorithms, which are DP and AP, with other well-established clustering algorithms, in particular K-means, DBSCAN, Bi-Clustering, Co-Clustering and Spectral clustering.

Contributions. This work reveals that, the two latest data clustering algorithms, AP and DP, do not always outperform the classic clustering algorithms, such as DBSCAN. Hence, to find the best clustering algorithm for a specific application, all of AP, DP and DBSCAN should be tested. Furthermore, we find the comparison of different clustering algorithms is closely related to the clustering validation metrics adopted. Thus, before selecting the best clustering algorithm for a given dataset/application, it is necessary to pick the clustering evaluation metric in advance.

The remaining of the paper is organized as follows. In Section 2., we briefly review common clustering evaluation metrics. Next, in Section 3., we describe the setup for the experiments. We analyse the experimental results in Section 4. and conclude the paper in Section 5.

2. Clustering Evaluation Metrics

There are two types of measures to evaluate the results of the clustering algorithms (i.e., the quality of the clusters), which are the internal and external validation metrics [11]. The basic idea of the internal validation measures is to check whether the intra-cluster similarities (the similarities between the data points inside the same cluster) are high, while, at the same time, the inter-cluster similarities (the similarities between data points from different clusters) are low. For instance, an intuitive internal validation measure that easily comes into our mind is to simply divide the intra-cluster similarities by the inter-cluster similarities. In this work, we use two very well-known internal clustering validation metrics, which are Dunn and Silhouette [11,12].

The external validation metrics calculate for each cluster, the distribution of the true class labels for all the data points in the same cluster. Therefore, this type of clustering evaluation metrics require each data point to have a class label. If all or the majority of the data points in a cluster share the same class label, this implies that the clustering is very successful, then the corresponding score, in terms of an external clustering validation metric, will be high. In this paper, we utilize 6 external clustering validation metrics, which are Purity, Homogeneity, Completeness, V_measure, Adjusted_rand and Mutual_info_score (Mutual information) [11,12].

3. Experimental Setup

3.1. Clustering algorithms to be compared

In this work, we compare the performance of the two latest data clustering algorithms (i.e., AP and DP) with 5 classic clustering algorithms, which include K-Means, DBSCAN, Bi-Clustering, Co-clustering and Spectral clustering. We adopt the implementations from Scikit-learn[1] for AP and these 5 classic algorithms, while the implementation of DP was obtained from its official website [13].

[1] Scikit-learn is a well-known open source machine learning library. http://scikit-learn.org

Table 1. Summary of datasets

Dataset	Number of instances	Number of attributes	With class label?	Sources
Aggregation	788	7	Yes	[17]
Flame	240	2	Yes	[17]
Compound	399	6	Yes	[17]
Spiral	312	3	Yes	[17]
Pathbased	300	3	Yes	[17]
R15	600	15	Yes	[17]
D31	3100	31	Yes	[17]
Jain	373	2	Yes	[17]
Breast	699	2	No	[17]
Thyroid	215	5	No	[17]
Yeast	1484	8	No	[17]
Wine	178	13	No	[17]
Dim4	2701	4	No	[17]
Dim8	5401	8	No	[17]
Dim32	1009	32	No	[17]
Dim64	1024	64	No	[17]

3.2. Datasets

We use 16 datasets to validate the quality of different clustering algorithms. These datasets can be divided into two categories: 1) 8 datasets commonly used for clustering [14], including Aggregation, Flame, Compound, Spiral, Pathbased, R15, D31, and Jain. All of these datasets contain class labels (ground truth) for the data points. 2) 8 datasets that do not contain class labels, including Dim4, Dim8, Dim32, Dim64, Breast, Thyroid, Yeast, and Wine [14]. Table 1 is a summary of these datasets.

3.3. Parameter Settings

For K-Means clustering, we use the default value as the number of clusters, which is 8. We also try other cluster numbers such as 2, 3. DP needs to specify the initial cluster numbers (or an initial cluster) to automatically search for a good radius value. We try three different values, which are 2, 3, and 6.

DBSCAN has two parameters, eps, which is the maximum radius of the neighbourhood from a point, and minPts, which is the minimum number of data points within this distance. In our experiments, we try different eps values, such as 0.2, 0.4, 0.9, 1.0, 3.0, and different values for minPts, such as 13.

For all the clustering algorithms that need to tune the parameters, we manually choose the set of parameters that can achieve the best clustering quality.

4. Experimental Results

In Table 2, we depict the experimental results of the 7 clustering algorithms on the 8 datasets that contain class label information, using 6 external clustering validation metrics. In Table 3, we also present the clustering results on the other 8 datasets without class labels, evaluated in terms of 2 internal clustering validation metrics.

4.1. The Performance of Co-Clustering, Bi-Clustering and Spectral Clustering

We observe from Table 2 and Table 3 that, Co-Clustering algorithm has never shown outstanding performance, while Bi-Clustering only obtains leading clustering result on one dataset, which is *Thyroid*. Hence, both of them can be neglected when selecting the best clustering algorithm for a specific dataset.

Spectral clustering seldom achieves excellent clustering results, except on the *Aggregation* dataset using the Mutual_info_score metric, as well as the *Jain* and *Pathbased* datasets. However, it never shows best performance according to the two internal metrics, as can be observed in Table 4.

Table 2. Evaluation of algorithms using external validation metrics.

Dataset	Algorithm	Purity	Homegeneity	Completeness	V_measure	Rand_score	Mutual_info_score
Aggregation	AP	0.996	0.990	0.599	0.746	0.377	0.589
	DP	0.511	0.253	0.909	0.396	0.202	0.251
	DBSCAN	0.827	0.718	1.000	0.836	0.734	0.716
	Spectral	0.834	0.846	0.786	0.815	0.549	0.783
	Cocluster	0.346	0.010	0.011	0.010	0.001	0.001
	Bicluster	0.506	0.340	0.902	0.494	0.348	0.339
	K-Means	0.911	0.909	0.765	0.830	0.668	0.761
Compound	AP	0.910	0.855	0.563	0.679	0.389	0.548
	DP	0.627	0.416	1.000	0.588	0.437	0.414
	DBSCAN	0.935	0.725	0.920	0.811	0.784	0.721
	Spectral	0.875	0.857	0.667	0.750	0.481	0.659
	Cocluster	0.396	0.012	0.013	0.012	-0.004	-0.005
	Bicluster	0.627	0.416	1.000	0.588	0.437	0.414
	K-Means	0.875	0.769	0.596	0.671	0.453	0.586
Flame	AP	0.833	0.932	0.244	0.387	0.134	0.236
	DP	0.988	1.000	0.420	0.591	0.410	0.416
	DBSCAN	0.742	0.741	0.396	0.516	0.456	0.393
	Spectral	0.971	0.873	0.279	0.423	0.202	0.274
	Cocluster	0.654	0.024	0.010	0.015	0.004	0.005
	Bicluster	0.838	0.410	0.388	0.399	0.453	0.386
	K-Means	0.983	0.910	0.289	0.438	0.206	0.284
Jain	AP	1.000	1.000	0.238	0.385	0.120	0.233
	DP	1.000	1.000	0.812	0.897	0.954	0.812
	DBSCAN	0.997	0.247	0.375	0.298	0.337	0.243
	Spectral	1.000	1.000	0.285	0.444	0.185	0.282
	Cocluster	0.740	0.018	0.007	0.010	0.006	0.003
	Bicluster	0.786	0.407	0.337	0.369	0.324	0.336
	K-Means	0.987	0.924	0.261	0.406	0.166	0.257
Pathbased	AP	0.963	0.916	0.380	0.537	0.273	0.367
	DP	0.633	0.402	0.636	0.493	0.401	0.400
	DBSCAN	0.927	0.340	0.620	0.439	0.325	0.338
	Spectral	0.877	0.781	0.430	0.555	0.348	0.423
	Cocluster	0.387	0.006	0.004	0.005	-0.004	-0.005
	Bicluster	0.633	0.401	0.634	0.491	0.399	0.399
	K-Means	0.847	0.710	0.406	0.517	0.391	0.398
R15	AP	0.997	0.994	0.994	0.994	0.993	0.994
	DP	0.667	0.773	0.984	0.886	0.579	0.763
	DBSCAN	0.533	0.590	1.000	0.743	0.264	0.576
	Spectral	0.530	0.704	0.988	0.822	0.517	0.694
	Cocluster	0.108	0.021	0.039	0.027	0.001	0.003
	Bicluster	0.133	0.244	0.980	0.391	0.120	0.241
	K-Means	0.533	0.590	1.000	0.743	0.264	0.576
Spiral	AP	0.888	0.770	0.288	0.419	0.144	0.272
	DP	1.000	1.000	1.000	1.000	1.000	1.000
	DBSCAN	0.772	0.394	0.393	0.393	0.142	0.387
	Spectral	0.808	0.684	0.373	0.483	0.258	0.366
	Cocluster	0.397	0.013	0.009	0.010	0.001	0.001
	Bicluster	0.353	0.001	0.001	0.001	-0.003	-0.002
	K-Means	0.487	0.182	0.098	0.127	0.048	0.088
D31	AP	0.975	0.966	0.966	0.966	0.950	0.964
	DP	0.161	0.360	0.958	0.524	0.107	0.357
	DBSCAN	0.065	0.042	0.976	0.081	0.004	0.040
	Spectral	0.258	0.575	0.988	0.727	0.328	0.571
	Cocluster	0.045	0.006	0.0013	0.008	-0.000	-0.000
	Bicluster	0.065	0.188	0.933	0.313	0.060	0.187
	K-Means	0.258	0.566	0.944	0.708	0.336	0.562

4.2. The Performance of AP, DP, DBSCAN and K-Means

We can see from Table 2 and Table 4 that, on the external validation metrics, AP, DP and DBSCAN show very good clustering results. Moreover, DP and DBSCAN achieve the best clustering results on the Dunn internal metric.

Surprisingly, on the Silhouette internal metric, the overall performance of K-Means is as good as DP and AP.

We now compare the two density-based clustering algorithms, i.e., DP vs DBSCAN. We see that, on two datasets, *Aggregation* and *Compound*, DBSCAN outperforms DP, while on the rest 6 datasets DP outperforms DBSCAN in almost all the metrics (with few exceptions), as can be seen in Table 4.

4.3. Efficiency Comparison of Different Clustering Algorithms

We have also checked the efficiency of different clustering algorithms. We find that, AP is very time-consuming, especially when the number of data points is large, say, more than 3000. DP is faster than AP, but slower than Co-Clustering, Bi-Clustering, and K-Means in general. In Figure 1, the running time of these algorithms on *Aggregation*, *Yeast*, and *Dim8* datasets is depicted. It should be mentioned that in *Yeast* and *Dim8* datasets, AP was also found to be the slowest algorithm, at least 3 times slower than DP, so we removed it from the plots for clarity reasons.

Table 3. Evaluation of algorithms using internal validation metrics.

Metrics \ Algorithm / Data sets		AP	DP	DBSCAN	Spectral	Cocluster	Bicluster	K-Means
Breast	Dunn	0.00	0.408	0.093	0.000	0.000	0.041	0.051
	Silhouette	0.182	-0.298	0.631	-0.057	0.067	0.122	0.722
Thyroid	Dunn	0.067	0.019	0.050	0.017	0.008	0.091	0.047
	Silhouette	0.230	-0.022	0.651	0.194	0.354	0.577	0.462
Yeast	Dunn	0.054	0.087	0.019	0.000	0.000	0.025	0.026
	Silhouette	0.139	-0.042	-0.282	0.204	0.323	0.191	0.356
Wine	Dunn	0.178	0.256	0.265	0.190	0.109	0.256	0.147
	Silhouette	0.115	0.279	0.310	0.118	0.304	0.280	0.473
Dim4	Dunn	0.561	0.561	4.904	0.003	0.001	0.463	0.601
	Silhouette	0.951	0.951	0.950	0.480	0.871	0.511	0.912
Dim8	Dunn	0.068	0.453	4.171	0.292	0.003	0.745	5.085
	Silhouette	0.837	0.932	0.995	0.583	0.604	0.311	0.851
Dim32	Dunn	4.035	4.035	0.016	0.003	0.000	0.645	0.771
	Silhouette	0.945	0.945	0.050	-0.212	0.389	0.191	0.523
Dim64	Dunn	5.820	5.820	0.012	0.004	0.001	0.777	0.764
	Silhouette	0.966	0.966	0.077	-0.194	0.355	0.154	0.511

Table 4. Total number of datasets where a classifier ranked first.

Total number \ Alg / Metrics	AP	DP	DBSCAN	K-Means	Spectral
Dunn	2	4	2	1	0
Silhouette	3	3	2	3	0
Purity	5	2	1	0	1
Homegeneity	5	3	0	0	2
Completeness	0	5	2	0	1
V_measure	2	3	2	0	1
Rand_score	2	3	3	0	0
Mutual_info_score	2	3	1	0	1

4.4. Summary of the Results

In summary, we draw the following summary from the above analyses:

1. Although AP and DP are the two latest (and very popular) clustering algorithms, they do not always outperform DBSCAN.
2. AP, DP, and DBSCAN, when put together as a group, show very outstanding performance than the other clustering algorithms. Therefore, AP, DP and DBSCAN should be the major candidates for the clustering tasks in real-world applications.
3. Co-Clustering, K-Means, and Bi-Clustering are generally the most efficient clustering algorithms. AP is the slowest one, whereas DP is more than 3 times faster than AP.
4. The comparison of different clustering algorithms depends on the evaluation metrics.

5. Conclusions

In this work, we empirically compare the performance of the two latest data clustering algorithms, which are affinity propagation (AP) and density peak based clustering (DP), with state-of-the-art clustering algorithms, using 6 external and 2 internal clustering validation metrics. Our experimental results demonstrate that, the two latest clustering algorithms AP and DP do not always outperform DBSCAN. This means that, even with the two latest clustering algorithms, there is no single clustering algorithm that is the best for all the datasets. Therefore, to select the best clustering algorithm for a specific dataset, all of AP, DP and DBSCAN should be considered/tested. In addition, we find the comparison of different clustering algorithms is also closely related to the clustering evaluation metrics adopted.

In terms of running time efficiency, our experiments show that AP is the least efficient clustering algorithm, it consumes at least 3 times more running time than the others, while Co-Clustering, K-Means, and Bi-Clustering are usually the top-3 most efficient algorithms.

This work provides valuable empirical reference for researchers and engineering who need to select the best clustering algorithm for their specific applications.

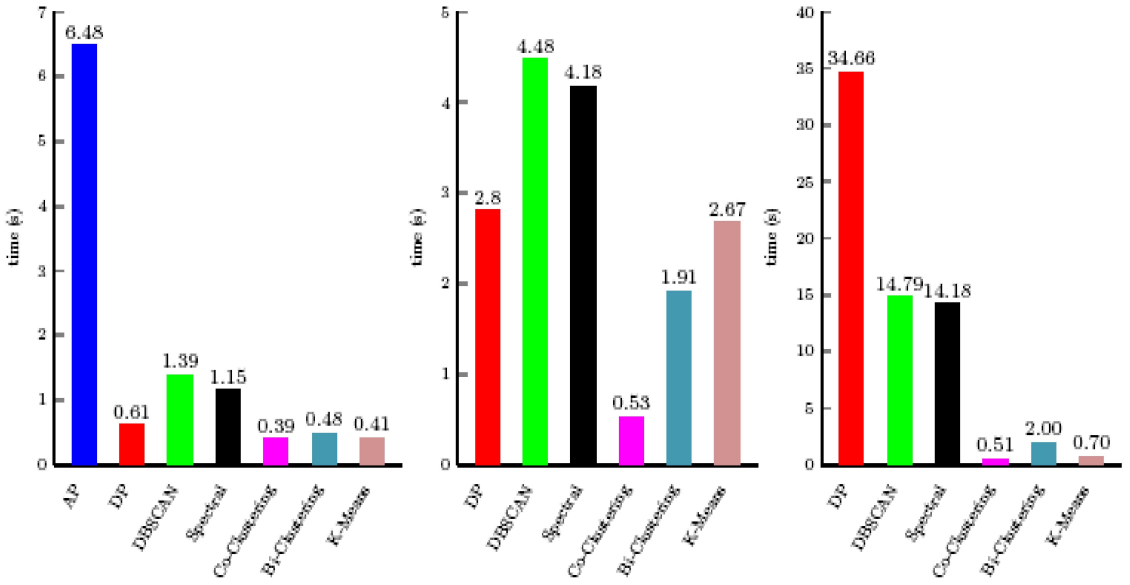

Figure 1. Running time of different algorithms on Aggregation (left), Yeast (center), and Dim8 (right) datasets.

References

[1] Cluster analysis. Wikipedia. https://en.wikipedia.org/wiki/Cluster_analysis.
[2] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. ACM Comput. Surv. 31, 3, 1999, pp 264-323.
[3] S. Kotsiantis, P. Pintelas, Recent Advances in Clustering: A Brief Survey, WSEAS Transactions on Information Science and Applications, Vol 1, No 1, 2004, pp 73-81.
[4] Rui Xu and D. Wunsch. Survey of clustering algorithms. IEEE Transactions on Neural Networks, 2005, pp 645-678.
[5] P. Berkhin, A survey of clustering data mining techniques, in: J. Kogan, C. Nicholas, M. Teboulle (Eds.), Grouping Multidimensional Data, Springer, Berlin, Heidelberg, 2006, pp. 2571.
[6] Martin Ester, Hans-Perer Kriegel, Jorg Sander, Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.1996.
[7] Mirkin, Boris. Mathematical Classification and Clustering. Kluwer Academic Publishers. 1996.
[8] Sepandar Kamvar, Dan Klein, Christopher Manning. Spectral learning. In IJCAI03, pp 561–566, 2003.
[9] Brendan J.Frey, Delbert Dueck. Clustering by Passing Messages Between Data Points. Science. 2007.
[10] Alex Rodriguez ,Alessandro Laio. Clustering by fast search and find of density peaks. 2014.
[11] Erendira Rendon,Itzel Abundez,Alejandra Arizmendi,Elvia M.Quiroz. Internal versus External cluster validation indexes. International Journal of Computers and Communications. Issue 1, Volume 5, 2011.
[12] Clustering. Scikit-learn-0.17.1-documentation. http://scikit-learn.org/stable/modules/clustering.html.
[13] http://people.sissa.it/ laio/Research/Res_clustering.php.
[14] Clustering datasets. http://cs.joensuu.fi/sipu/datasets/.
[15] http://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets.