International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue 6, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Performance Evaluation of Spectral Clustering Algorithm using Various Clustering Validity Indices

M. T. Somashekara
Department of Computer Science and Applications,
Bangalore University, Bangalore- 560056, Karnataka, India
Email: somashekar_mt@hotmail.com

D. Manjunatha
Department of Electronics Science,,
Tumkur University, Tumkur, Karnataka, India
Email: manjums08@gmail.com

***Abstract* – In spite of the popularity of spectral clustering algorithm, the evaluation procedures are still in developmental stage. In this article, we have taken benchmarking IRIS dataset for performing comparative study of twelve indices for evaluating spectral clustering algorithm. The results of the spectral clustering technique were also compared with k-mean algorithm. The validity of the indices was also verified with accuracy and (Normalized Mutual Information) NMI score. Spectral clustering algorithm showed better results when compared to k-mean algorithm. All indices showed consistent results with spectral clustering technique. Silhouette Index, Hartigan Index, Davies-Bouldin (DB) index and Krzanowski-Lai (KL) index failed to evaluate k-mean clustering. Surprisingly, all eleven indices showed acceptable results for spectral clustering algorithm. This article confirms the superiority of spectral clustering algorithm and also confirms that all 12 indices are suitable for evaluating spectral clustering.**

***Keywords* – Spectral Clustering, Validity Indices, NMI Score, K-Mean Algorithm.**

I. INTRODUCTION

Spectral clustering is the most popular clustering technique which has immense applications in the field of machine learning, exploratory data analysis, computer vision and speech processing. It contains eigen structure of a similarity matrix to partition points into disjoint clusters with points in the same cluster having high similarity and points in different clusters having low similarity [1]. There are wide varieties of spectral algorithms that use the eigenvectors in different ways. In this paper we have used the most popular Andrew's approach of spectral clustering [2].

Spectral clustering algorithm

Given a set of points $S = \{s_1, \ldots\ldots, s_n\}$ in R^l that we want to cluster into k subsets:

1. Form the affinity matrix $A \in R^{n \times n}$ defined by $A_{ij} = \exp(-\|s_i - s_j\|^2 / 2\sigma^2)$ if $i \neq j$, and $A_{ii} = 0$.
2. Define D to be the diagonal matrix whose (i,i)-element is the sum of A's i-th row, and construct the matrix $L = D^{-1/2} A D^{-1/2.l}$
3. Find $x_1, x_2 \ldots\ldots, x_k$, the k largest eigen vectors of L (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix $X = [x_1 x_2 \ldots\ldots x_k] \in R^{nxk}$ by stacking the eigenvectors in columns.
4. Form the matrix Y from X by renormalizing each of X's rows to have unit length (i.e. $Y_{ij} = X_{ij} / (\Sigma_j X_{ij}^2)^{1/2}$).
5. Treating each row of Y as a point in R^k, cluster them into k clusters via K-means or any other algorithm (that attempts to minimize distortion).
6. Finally, assign the original point s_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

The hardest problem in comparing different clustering algorithms is to find an algorithm independent measure to evaluate the quality of the clusters [3]. There are many clustering indices available for evaluating clustering algorithms but still there is no single index for evaluating all clustering algorithms.

II. CLUSTER VALIDITY INDICES

For validity evaluation of clustering solutions, we have used the following indices

Rand/Adjusted Rand: The Rand index or Rand measure in statistics, and in particular in data clustering, is a measure of the similarity between two data clustering. The adjusted-for-chance form of the Rand index is the adjusted Rand index [4,14].

Hubert index: It is a measure based on the comparison of object triples having the advantage of a probabilistic interpretation in addition to being corrected for chance (i.e., assuming a constant value under a reasonable null hypothesis) and bounded between ±1 [5].

Silhouette index: A composite index reflecting the compactness and separation of clusters; a larger average Silhouette index indicates a better overall quality of the clustering result, so the optimal NC is the one that gives the largest average Silhouette value [6,3]

Davies-Bouldin index: A measure of the average similarity between each cluster and its most similar one; small values correspond to clusters that are compact and have centers that are far away from each other; therefore, its minimum value determines the optimal NC [7,8].

Calinski-Harabasz index: The measures of between-cluster isolation and within-cluster coherence; its maximum value determines the optimal NC [9,10].

Hartigan index: Hargitan index (1975) [11], a statistical index to examine the relative change of fitness as number of clusters changes.

Weighted inter-intra index: A weighted inter-cluster edge ranking for clustered graphs that weighs edges (based on whether it is an inter-cluster or an intra-cluster edge) and nodes (based on the number of clusters it connects) [12].

Krzanowski-Lai index: It is based on the criteria for determining the number of groups in a data set using sum-of-squares clustering [6]

Homogenity and separation index: The goal is to partition the elements into subsets, which are called clusters, so that two criteria are satisfied: Homogeneity - elements in the same cluster are highly similar to each

other; and separation - elements from different clusters have low similarity to each other [13].

The results were also compared with accuracy and (Normalized Mutual Information) NMI scores. We have tested eleven indices on spectral clustering algorithm with benchmark IRIS dataset and also compared the results with accuracy and NMI score. We have compared the spectral clustering results with k-mean algorithm.

III. Materials and Methods

The results presented in this paper are obtained with the IRIS flower data set or Fisher's Iris data set. The dataset consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). The dataset can be freely downloaded from internet via (http://archive.ics.uci.edu/ml/machine-learning-databases /iris/iris.data). The data present in the dataset has to be divided into three clusters as per the reference. The indices which show three clusters can be considered as the best index. The K-means clustering and spectral clustering algorithms are applied to the IRIS flower dataset and the indices were evaluated at different clusters.

IV. Results and Discussion

The number of clusters provided by the two clustering algorithms in conjunction with the eleven validity indices for the IRIS flower data sets is provided in Table 1 and Table 2. For performance comparison between k-mean clustering and spectral clustering and for finding the best clustering validity index, a known standard reference IRIS flower data set was taken. The total number of clusters present in the IRIS reference flower dataset was three and if the optimal values predicted by the clustering validity index algorithms for IRIS reference dataset is at third cluster, then the corresponding validity index can be accepted [14]. The index is found to attain its maximum optimal value or minimum optimal value when the appropriate number of clusters is achieved. The clustering index is considered to be failed, if it is not following any pattern [15]. In this context, it is understood from the results (Table 2) that spectral clustering algorithm by Jordon approach showed consistent results for all eleven indices when compared with k-mean clustering algorithm. The optimal values of all clustering index algorithms were at the third cluster for Spectral Clustering which is equivalent to the reference cluster value from IRIS flower dataset.

Table 1: Clustering index from K-mean algorithm

K-mean Clustering	**Total Number of Clusters**									
	1	2	3	4	5	6	7	8	9	10
Rand Index	**0**	0.76367	**0.87973**	0.83919	0.84188	0.7711	0.84501	0.83132	0.82801	0.76787
Adjusted Rand Index	0	0.53992	**0.73024**	0.61625	0.61245	0.42156	0.6084	0.56666	0.5568	0.36802
Mirkin index	0	0.23633	**0.12027**	0.16081	0.15812	0.2289	0.15499	0.16868	0.17199	0.23213
Hubert Index	0	0.52734	**0.75946**	0.67839	0.68376	0.54219	0.69002	0.66264	0.65602	0.53575
Silhouette Index	0	0.68081	0.55259	0.49629	0.49101	0.3561	0.4627	0.43186	0.43355	0.29842
Hartigan Index	0	513.3	136.73	54.948	33.655	-3.2809	26.525	16.419	7.1932	16.135
Davies-Bouldin (DB) index	0	0.40483	0.58785	0.58207	0.66063	0.60977	0.81513	0.85038	0.80463	0.91836
Calinski-Harabasz (CH) index	0	513.3	**560.4**	529.02	493.92	382.82	379.98	363.16	332.53	329.1
Krzanowski-Lai (KL) index	0	5.9089	3.5765	2.0734	0.84328	1.2143	2.3363	4.7484	0.16581	0.16581
Weighted Inter-Intra (Wint)	0	0.55473	**0.69091**	0.63582	0.62916	0.60631	0.62871	0.63144	0.59968	0.57753
Homogeneity Separation Index	0	-2.6827	-2.3752	-2.2616	-2.1926	-2.0012	-2.0239	-1.9371	-1.8677	-1.7027
Accuracy	33.333	66.667	**89.333**	69.333	66.667	51.333	64	56.667	56	42
NMI Score	0	0.67932	**0.75821**	0.70796	0.70542	0.63404	0.70863	0.68183	0.67213	0.6441
Error Rate	0	3.57	51.99	100						

Table 2: Clustering index from Spectral Clustering Algorithm

Spectral Clustering [2]	**Total Number of Clusters**									
	1	2	3	4	5	6	7	8	9	10
Rand Index	**0**	0.61915	**0.80286**	0.79973	0.79517	0.74416	0.72	0.72519	0.72116	0.72089
Adjusted Rand Index	0	0.24695	**0.55406**	0.51965	0.49271	0.33775	0.28136	0.27613	0.23907	0.23533
Mirkin index	0	0.38085	**0.19714**	0.20027	0.20483	0.25584	0.28	0.27481	0.27884	0.27911
Hubert Index	0	0.2383	**0.60573**	0.59946	0.59034	0.48832	0.44	0.45038	0.44233	0.44179
Silhouette Index	0	0.19724	**0.4787**	0.36093	0.32937	0.18298	0.14547	0.10655	0.1378	0.13072
Hartigan Index	0	15.789	739.86	**24.023**	**16.236**	**11.585**	**-11.674**	**14.345**	**32.694**	**2.689**
Davies-Bouldin (DB)index	0	2.4087	**0.68391**	0.83133	1.1979	1.1684	1.1364	1.252	1.3714	1.4246
Calinski-Harabasz (CH) index	0	15.789	**417.24**	329.73	277.17	240.23	180.63	171.32	187.45	168.92
Krzanowski-Lai (KL) index	0	0.27285	**464.46**	1.5642	0.39051	0.077721	5.1566	0.21521	4.716	4.716
Weighted Inter-Intra (Wint)	0	0.37043	**0.65684**	0.63696	0.63532	0.60767	0.59474	0.60492	0.57895	0.56854
Homogeneity Separation Index	0	-0.39921	**-2.2206**	-2.0759	-2.007	-1.8268	-1.6542	-1.6157	-1.6486	-1.5869
Accuracy	33.333	58.667	**79.333**	69.333	62.667	50	46.667	43.333	34.667	34.667
NMI Score	0	0.27953	**0.60011**	0.55028	0.52196	0.48877	0.47739	0.46895	0.46411	0.45763
Error Rate	0	24.24	**51.85**	100						

Rand, Adjusted Rand, Hubert, Silhouette, Hartigan, CH, KL, Wint and NMI score showed optima maxima and the remaining indices showed optima minima values at third cluster for spectral clustering algorithm. Compared to the k-mean clustering, it is found to be more consistent and reliable in predicting the correct number of clusters. The K-mean clustering failed to represent Silhouette Index, Hartigan Index, Davies-Bouldin (DB) index, Krzanowski-Lai (KL) index, Homogeneity Separation Index as the results were not following any pattern. It is also suggested not to use the above clustering indices for validating the results from k-mean clustering algorithm.

The total number of clusters can be finalized based on the results from the validity indices. Higher the rand and adjusted rand index, more the accuracy. Lowest Mirkin [16], Davies-Bouldin (DB) index, Homogeneity Separation Index will give good results. The clusters can be selected based on the results from the indices. From the above tables (Table 1 and Table 2), it is clearly observed that all indices are either decreasing or increasing after third cluster and from the indices, we can confirm the third cluster is best when compared with other clusters. The optimal values were bolded in Table 1 and Table 2 for illustration purpose. To evaluate the quality of document clusters, it is customary to use the Normalized Mutual Information (NMI) [17] and accuracy, which is a standard way to measure the cluster quality. The highest NMI score with highest accuracy percentage is considered to be best cluster and in both k-mean and spectral clustering, the values at third cluster satisfy the criteria.

V. Conclusion

Cluster validation is an important and necessary step in cluster analysis. A spectral clustering algorithm is proposed to facilitate cluster validation and cluster analysis. The algorithm provides the necessary methods and tools as well as an analysis environment for clustering and cluster validation and can help a user accomplish his clustering task faster and better. All indices were successfully applied to the spectral clustering technique when compared with the K-mean algorithm.

References

[1] Bach, Francis R., and Michael I. Jordan, "Learning graphical models with Mercer kernels", In *Advances in Neural Information Processing Systems*, pp. 1009-1016. 2002.

[2] Ng, Andrew Y., Michael I. Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm", *Advances in neural information processing systems*, Volume 2, 2002, Pages 849-856.

[3] Chen, Gengxin, Saied A. Jaradat, Nila Banerjee, Tetsuya S. Tanaka, Minoru SH Ko, and Michael Q. Zhang, "Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data", *Statistica Sinica*, Volume 12, Issue 1, 2002, pages 241-262.

[4] Rand WM, "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association (American Statistical Association)*, Volume 66, Issue 336, 1971, Pages 846–850.

[5] Hubert. L. and Arabie. P, "Comparing Partitions". *Journal of Classification*, Volume 2, 1985, Pages 193-218

[6] W. J. Krzanowski and Y. T. Lai, "A criterion for determing the number of groups in a data set using sum-of-squares clustering", *Biometrics*, Volume 44, 1985, Pages 23–34.

[7] Dimitriadou, E., Dolnicar, S., & Weingessel, A, "An examination of indexes for determining the Number of Cluster in binary data sets". *Psychometrika*, Volume 67, Issue 1, 2002, Pages 137-160.

[8] Bolshakova, N. & Azuaje, F, "Cluster validation techniques for genome expression data". *SignalProcessing*, Volume 83, Issue 4, 2003, Pages 825-833.

[9] Dudoit, S. & Fridlyand, J, "A prediction-based resampling method for estimating the number of clusters in a dataset". *Genome Biology*, Volume 3, Issue 7, 2002, Pages: 0036.1-21.

[10] Shu, G., Zeng, B., Chen, Y. P., & Smith, O. H, "Performance assessment of kernel density clustering for gene expression profile data", *Comparative and Functional Genomics*, Volume 4, Issue 3, 2003, Pages 287-299.

[11] Hartigan JA "Clustering algorithms". Wiley series in Probability and Mathematical Statistics, John Wiley & Sons, Inc. 1975, USA

[12] Padmanabhan, Divya, Prasanna Desikan, Jaideep Srivastava, and Kashif Riaz. "WICER: a weighted inter-cluster edge ranking for clustered graphs." In *Web Intelligence*, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on, pp. 522-528. IEEE, 2005.

[13] Sharan, R., Maron-Katz, A., and Shamir, R, "CLICK and EXPANDER: A System for Clustering and Visualizing Gene Expression Data". *Bioinformatics*, Volume 19, 2003, Pages 1787-1799.

[14] Wang Kaijun, Baijie Wang, and Liuqing Peng, "CVAP: Validation for cluster analyses", *Data Science Journal*, Volume 8, 2009, Pages 88-93.

[15] Maulik, Ujjwal, and Sanghamitra Bandyopadhyay. "Performance evaluation of some clustering algorithms and validity indices." Pattern Analysis and Machine Intelligence, *IEEE Transactions* on 24, no. Volume 12, 2002, Pages 1650-1654.

[16] Mirkin, BG, Cherny LB, "On a distance measure between partitions of a finite set". *Automation and remote control*, Volume 31, Issue 5, 1970, Pages 91-98

[17] Strehl, Alexander, and Joydeep Ghosh. "Cluster ensembles---a knowledge reuse framework for combining multiple partitions", *The Journal of Machine Learning Research*, Volume 3, 2003, Pages 583-617.

Author's Profile

M. T. Somashekara
is currently working as Assistant Professor in the Dept of Computer Science and Applications, Bangalore University. His areas of research include Bioinformatics, Pattern Recognition and Image Processing.
Email: somashekara_mt@hotmail.com

D. Manjunatha
is currently working as Assistant professor in the Dept of Electronics Science, Tumkur University, Tumkur. He has many research articles to his credit. His areas of research include Wireless Sensor Networks and Algorithms.
Email: manjums08@gmail.com