Performance Validation of the Modified K-Means Clustering Algorithm Clusters Data

S. Govinda Rao
Associate Professor, Department of CSE, Gokaraju Rangaraju Institute of Engineering & Technology
govindsampathirao@gmail.com
Dr.A. Govardhan
Director, School of Information Technology, JNTU Hyderabad
govardhan_cse@yahoo.co.in

Abstract: In this paper, we present the Modified K-Means Clustering algorithm Analysis and performance, the clustering analysis can be used to partition the cluster data with number of choice clusters and perform each cluster if it can form properly or not and it can pertain by using the silhouette coefficient method. In this one the silhouette coefficient can apply on the group of author's H- and G-indices with same or different features [1]. The silhouette coefficient analysis can be used to separate the distance from each resulting clusters, the silhouette value measures and shows how each point in one cluster with other points in another cluster and also visually it provides how those cluster are formed with effectively the main functionality of the clustering analysis is to identify the quality assessment of the clustering results.

The silhouette index investigated and suggests that the use of the preprocessor improves the quality of clusters significantly for the h and g indices data sets. Furthermore, it is then shown that the modified *K*-means algorithm good quality, compact and well-separated clusters of the h and g indices data

Keywords: Modified K-Means Algorithm, Silhouette index, Cluster Performance.

1. INTRODUCTION

Cluster analysis is most important aspect in many areas like data mining, information systems etc.. Clustering results are obtained based clustering algorithms [2]. The importance of the cluster analysis is evaluation of the cluster results to find how those partitioned data can be clustered [3]. The job of any clustering algorithm is to provide clusters that are condensed and well-separated from one another. It pursue that a clustering job associate reducing the intra-cluster distance or the within-cluster diffusion and maximizing the inter-cluster distance or the between-cluster diffusion [4].

The K Means algorithm is most popular clustering algorithm and use this algorithm to partition the number of clustering for pre specified dataset. The number of clusters can be partitioned based on the trail and error process made more difficult to establish correct clusters. The performance of the K means algorithm can be calculated based on the K value, the selected values have to be significantly smaller than the number of objects in the data sets, which is the main motivation for performing data clustering [5]. The K means algorithm processes the h and g indices of the group of scientific authors, the h and g indices measure the scientific performance of the author. If the h and g value is high means that particular author performance is also high, else author performance is poor in case if h and g values are low. h-index evaluates the score generated from the papers published by the specific author as well as the number of papers published since the first publication[6].

The K Means algorithm suffers from some poor cluster performance issues, so that we have proposed new kind of clustering algorithm that is Modified K Means Clustering algorithm. modification of K-means algorithm that efficiently searches data to cluster points by compute the sum of squares within each cluster, which makes the program to select the most promising subset of classes for clustering. The h- and g- indices of few authors who have published scientific papers of excellence in the fields of computer science are segregated [7]

2. H- AND G- INDICES DATASET

The H,G indices are used to measure the performance of the scientific authors, these two are key indexing values to measure scientific performance of the authors if these values are high means that particular author performance is also high, otherwise if its values are low means that particular author performance is also low. A research author has h-index h if h of his n papers has at least h citations each and the other n- h papers have fewer than h + 1 citation each. the h index is also called as Hirch index.[8]. An scientific authors h index value is not greater than his /her number of publications The g-index is measured based on the allocation of citations received by a given authors publications, so that given a set of publications ranked in decreasing order of the number of citations that they received, the g-index is the unique largest number such that the top g articles received together at least g^2 citations[1]

3. METHODOLOGY

The K-Means Clustering algorithm is suffer from so many problems like bad computational time and efficiency ,so that we proposed modified K-Means algorithm can solve the above .the modified K-means algorithm forms the best clusters with good computational time and it proves the best efficiency of the algorithm[]. The main functionality of the modified K-Means Algorithm is to takes the set of H- and G-indices of the group of scientific authors and it can be clustered with k number of clusters. Modified k-means algorithm was presented where a metric was used to the sum of compute the sum of squares with in clusters to elect the best one. The sum of squares within the cluster indicates the sum of all distances between each data point and the centroid of its cluster and its value is smaller, more impact and it is best cluster. So that, for a given dataset, clusters with the smaller sum of squares within a cluster are regarded as generally better. The time required to perform both the algorithms are reported [8].

4. CLUSTER PERFORMANCE

The cluster performance can be calculated based on objects of within the clusters are more close to its centroids. In this paper we shown the performance of the modified K-Means Clustering Algorithm and this algorithm takes the h and g indices dataset and all the data in the data set is normalized before we applied modified k means clustering algorithm[9]. In the Modified K-Means algorithm all objects in each cluster are closer to its centroids, but in convention k means algorithm some of the objects are closer to two or more cluster centroids[10]. So that compared to conventional k means the modified k means cluster performance and quality of the cluster is good.

4.1 Silhouette Index

Finding the right number of clusters is a challenging issue in cluster analysis literature, for which no unique solution exists. Therefore, different approaches have been proposed. One of the most popular methods to select the right value of K is by means of the silhouette coefficients. For a given point i in a cluster A, the silhouette of i, s(i) is defined as follows

$$\boxed{S_i = (b_i - a_i) / \max \{a_i, b_i\}} \qquad (1)$$

Where, a(i) is the average dissimilarity between point I and all other points in A (the cluster to which i belongs) and b(i) is the average dissimilarity between point i and the points in the closest cluster to A, which is B in this case. The average of all silhouettes in the data set S'is called the average silhouettes width for all points in the data set. The value S' will be denoted by S'(K), which is used for the selection of the right value of the number of clusters, K, by choosing that k for which S'(K) is as high as possible[11].

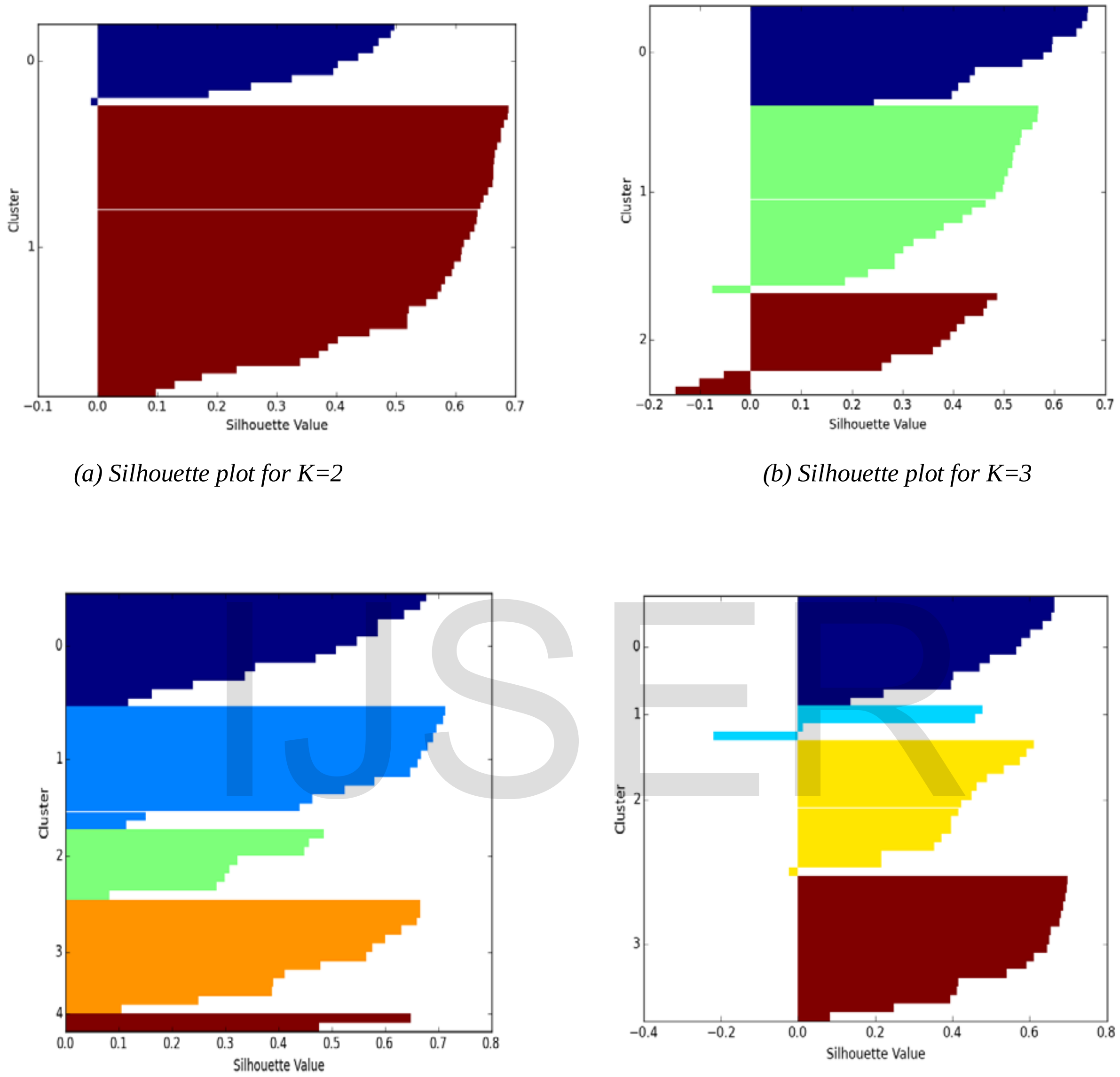
5. EXPERIMENTAL RESULTS AND DISCUSSIONS

The Modified K-Means Algorithm can be applied on experimental dataset H and G Indices of the set of scientific authors. The algorithm was applied more than once on a data set to see the effect of *K* in the clustering process [12]. Results of this investigation are presented in Table 1. In the given table.1 both the algorithms can run in different runs on h and g indices dataset and IRIS dataset. In this example algorithms run time can be changed every time and based on K value it can generates that particular number of clusters, the cluster objects can be close to their centroids at k=5 cluster but in remaining clusters objects are not formed properly in the clusters..We applied the modified k means on datasets IRIS with 150 records and also the h and g indices of 150 author's dataset

Table1. it shows the clustering results for the datasets

Dataset	Number of Records	Number of Clusters(K)	Conventional K Means Algorithm runs in sec	Modified K-Means Algorithm runs in sec
IRIS	150	K=5	13.19	0.56
H-G Indices	150	K=5	12.14	0.47

The silhouette coefficient is interpretation and validation of the within a clusters of data and it provides the successful graphical representation and it shows how well each object is lies within its clusters



(a) Silhouette plot for K=2

(b) Silhouette plot for K=3

(d) silhouette plot for K=5

(c) Silhouette plot for K=4

Fig.1 Silhouette plots for the clusters high dimensional data with different K values

6. CONCLUSION

The modified K-Means Algorithm is the best suitable clustering algorithm; in this paper we validating the Modified K -mean algorithm Cluster data. The cluster validation can be performed by using the silhouette coefficient index. This index can validated the cluster data with different k values. Based on the k valued the algorithm can partition that particular number of clusters. In this paper Modified K means is new but silhouette coefficient index is not new ,in future the silhouette index can upgraded to validate cluster data and the modified K means algorithm can be applied to complex datasets.

REFERENCES

[1]. S Govinda Rao, Dr A Govardhan. "Assessing h- and g- indices of scientific papers using k-means clustering" International Journal of Computer Applications(0975-8887), Vol.100-No.11,August 2014.

[2]. Halkidi, M.; Batistakis, Y.; and Vazirgiannis, M. 2001. On Clustering Validation Techniques. Journal of Intelligent Information Systems.

[3]. [Kaijun Wang, Baijie Wang , and Liuqing Peng "CVAP: VALIDATION FOR CLUSTER ANALYSES" Data Science Journal, Volume 8, 20 May 2009.

[4]. Bariléé Barisi Baridam " More Work on K -Means Clustering Algorithm: The Dimensionality Problem" International Journal of Computer Applications (0975 – 8887) Volume 44– No.2, April 2012

[5]. D T Pham🐦, S S Dimov, and C D Nguyen "Selection of K in K-means clustering" Manufacturing Engineering Centre, Cardiff University, Cardiff, UK September 2004

[6]. Jacso, P. (2008b). The pros and cons of computing the h-index using Google Scholar. Online Information Review, 32(3), 437–452

[7]. S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics 3: 273–289

[8]. T. Anderson, R. Hankin, and P. Killworth. Beyond the Durfee square: Enhancing the h-index to score total publication output. Scientometrics, 2008. In press

[9]. D T Pham🐦, S S Dimov, and C D Nguyen "Selection of K in K-means clustering" Proc. IMechE Vol. 219 Part C: J. Mechanical Engineering Science, September 2004

[10]. Nidhi Singh,.Divakar Singh "Performance Evaluation of K-Means and Heirarichal Clustering in Terms of Accuracy and Running Time" International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012,4119-4121

[11]. Moh'd Belal Al- Zoubi and Mohammad al Rawi "An Efficient Approach for Computing Silhouette Coefficients" Journal of Computer Science 4 (3): 252-255, 2008 ISSN 1549-3636

[12]. Campello, R.; Hruschka, E. 2006. A Fuzzy Extension of the Silhouette Width Criterion for Cluster Analysis. Fuzzy Sets and Systems, 157: 2858-2875.