RESEARCH ARTICLE OPEN ACCESS

Optimal Clustering Scheme For Repeated Bisection Partitional Algorithm

Kalyani Desikan *, G. Hannah Grace **
*(Department of Mathematics, VIT University, Chennai)
** (Department of Mathematics, VIT University, Chennai)

ABSTRACT
Text clustering divides a set of texts into clusters such that texts within each cluster are similar in content. It may be used to uncover the structure and content of unknown text sets as well as to give new perspectives on familiar ones. The focus of this paper is to experimentally evaluate the quality of clusters obtained using partitional clustering algorithms that employ different clustering schemes. The optimal clustering scheme that gives clusters of better quality is identified for three standard data sets. Also, the ideal clustering scheme that optimizes the *I2* criterion function is experimentally identified.

***Keywords* –** Cluster quality, Criterion Functions, Entropy and Purity

I. INTRODUCTION

The partitional clustering algorithms are well suited for clustering large document datasets due to their relatively low computational requirements according to study conducted by [1]. A report by [2] investigated the effect of the criterion functions to the problem of partitional clustering of documents and the results showed that different criterion functions lead to substantially different results. Another study reported by [3] examined the effect of the criterion functions on clustering document datasets using partitional and agglomerative clustering algorithms. Their results showed that partitional algorithms always led to better clustering results than agglomerative algorithms

The main focus of this paper is to perform experimental evaluation of various criterion functions in the context of the partitional approach, namely the repeated bisection clustering algorithm.

II. PRELIMINARIES

1.Cluster Quality

The quality of the clusters produced is measured using two external measures, namely, entropy [3][4][5] and purity . Entropy measures how various classes of documents are distributed within each cluster. The smaller the entropy values, better the clustering solution. Given a particular cluster S_r of size n_r, the entropy of this cluster is defined in [6] as

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}$$

where q is the number of classes in the dataset and n_r^i is the number of documents of the ith class that are assigned to the rth cluster. The entropy of the entire clustering is then the sum of the individual cluster entropies weighted according to the cluster size ,that is for k clusters, we have $Entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r)$. Smaller entropy values indicate better clustering solutions. Using the same Mathematical notation, the purity of a cluster is defined as in [7] $Pu(S_r) = \frac{1}{n_r}\max n_r^i$. The purity of the clustering solution is again the weighted sum of the individual cluster purities, $Purity = \sum_{r=1}^{k}\frac{n_r}{n}Pu(S_r)$. Larger purity value indicates better clustering solution.

2.Clustering Criterion Functions

Seven criterion conditions *I1*, *I2*, *E1*, *G1*,*G1p* ,*H1* and *H2* are mentioned in [2]. Theoretical analysis of the criterion functions by [8] shows that their relative performance depends on the (i) degree to which they can correctly operate when the clusters are of different tightness, and (ii) degree to which they can lead to reasonably balanced clusters. The main role of different clustering criterion functions is to determine which cluster to bisect next as discussed in [9].

In our experimental study, we have restricted our analysis to three criterion functions viz. *I2, E1* and *H2*. We did not consider *I1* since the only difference between *I1* and *I2* is that while calculating *I2* we take the square root of the similarity function. We also ignore *G1* and *G1p* since *G1* is similar to *E1* except that there is no square root in the denominator and *G1p* is similar to *E1* except that we have n_1^2 and that there is no square root in the denominator. Also, since *H1* is a hybrid function that maximizes *I1/E1,* we ignore it as we have not taken *I1* and *E1* into consideration.

I2 is an example of an internal criterion function that maximizes the similarity between each document and the centroid of the cluster to which it is assigned. *E1* is an external criterion function that focuses on optimizing a function that depends on the dissimilarity of the clusters. *E1* tries to minimize the similarity between the centroid vector of each cluster and the centroid vector of the entire collection. The contribution of each cluster is weighted based on the cluster size. Combinations of different clustering criterion functions provides a set of hybrid criterion functions that simultaneously optimize multiple individual criterion functions, for example, *H2* is obtained by combining *I2* with *E1.* Detailed descriptions of these criterion functions can be found in [2]. **Table1** gives the mathematical formulae for the criterion functions *I2, E1* and *H2.*

Table1

Criterion function	**Formula**
I2	**maximize** $\sum_{i=1}^{k}\sqrt{\sum_{u,v\in S_i} sim(u,v)}$
E1	**minimize** $\sum_{i=1}^{k} n_i \frac{\sum_{u\in S_i, v\in S} sim(u,v)}{\sqrt{\sum_{u,v\in S_i} sim(u,v)}}$
H2	**maximize** $\frac{I2}{E1}$

3. Clustering Schemes for Cluster Split

For Repeated bisection partitional clustering method, there are a number of clustering schemes to choose from. The clustering scheme determines the cluster to be bisected next as in [4]. The available schemes are “large”, which selects the largest cluster; “best” which selects the cluster that leads to the best cut; and “largess” which chooses the cluster that leads to the best reduction in subspace size. In this paper we investigate the three schemes experimentally and the results are shown. The detailed results of these experiments are omitted due to space limitation.

III. DOCUMENT DATASETS AND EXPERIMENTAL METHODOLODY

In this paper we have considered three datasets whose general characteristics are summarized in **Table 2**.

Table 2

Data Set	Number of Rows	Number of Columns	Number of Non - Zeros terms	Number of Classes
Classic	7094	41681	223839	4
Hitech	2301	126373	346881	6
mm	2521	126373	490062	2

The clustering tool we have used for our experiments is CLUTO. CLUTO is used for clustering low and high dimensional datasets and for analyzing the characteristics of the various clusters. CLUTO operates on very large set of documents as well as number of dimensions.

CLUTO can be used to cluster documents based on similarity/distance measures like cosine similarity, correlation coefficient, Euclidean distance and extended Jaccard coefficient. CLUTO provides cluster quality details such as Entropy and Purity.

We have experimentally analyzed cluster quality, based on entropy and purity, for the three data sets classic, hitech and mm for the three criterion conditions *I2*, *E1* and *H2*. We have applied all the three clustering schemes large, best and largess to all the three datasets to find out the scheme that gives the best cluster split.

1. Optimal Scheme for Cluster split

The first set of experiments was focused on evaluating the quality of clusters based on entropy and purity for the three schemes to ascertain the best cut cluster scheme.

2. Optimal Scheme for Optimizing *I2* Criterion Function

In the second set of experiments, our aim was to identify the clustering scheme that optimized the *I2* criterion function. We carried out our analysis for the three datasets by increasing the number of clusters from 2 to 100. We chose *I2* as the criterion function for our study because through an experimental study of Criterion functions, we have shown in our previous paper [10] that the *I2* function performs the best with respect to clustering time.

IV. RESULTS AND ANALYSIS

The results of entropy and purity obtained using repeated bisection method for various criterion functions are given for the three datasets in the figures below.

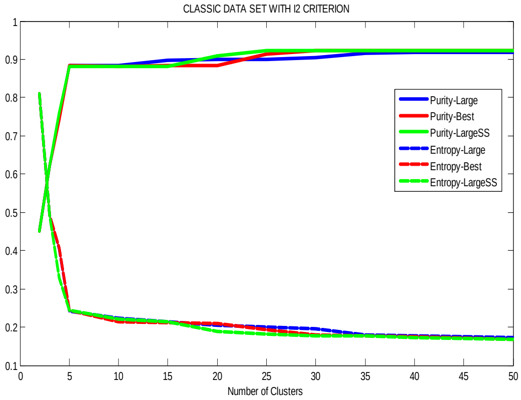

Fig.1

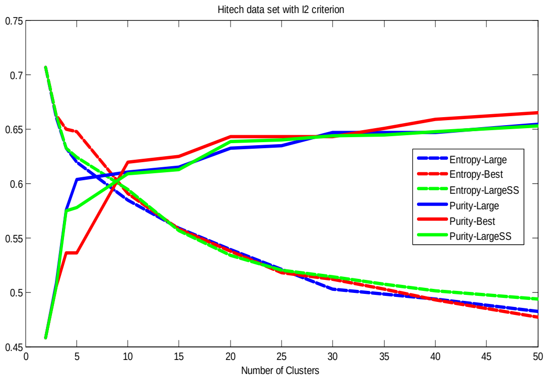

Fig.2

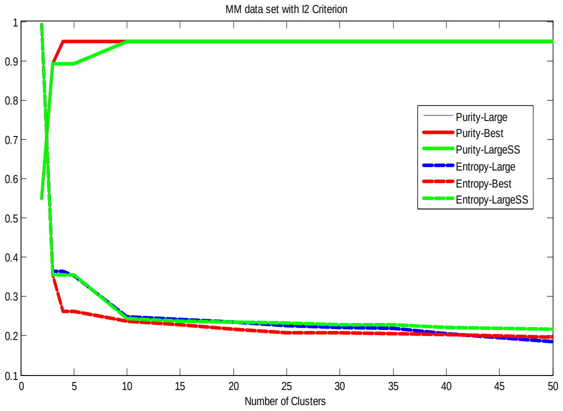

Fig.3

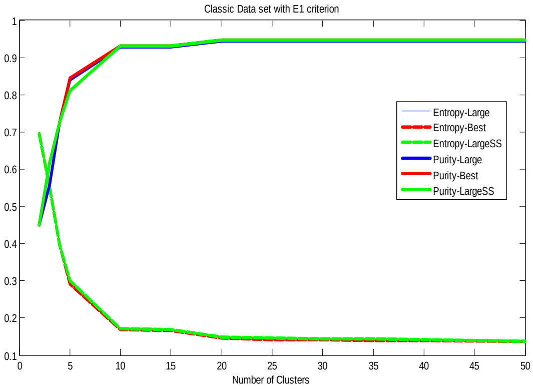

Fig.4

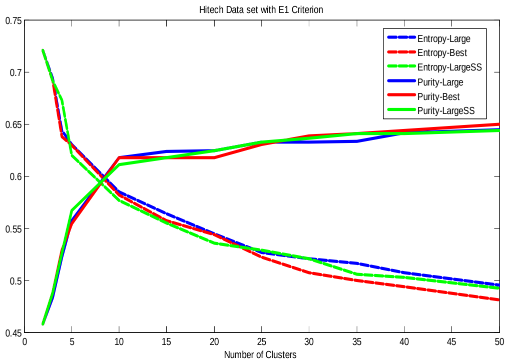

Fig.5

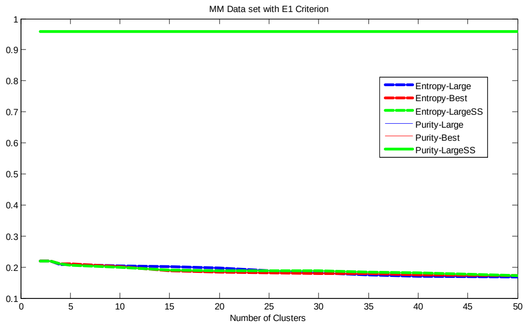

Fig.6

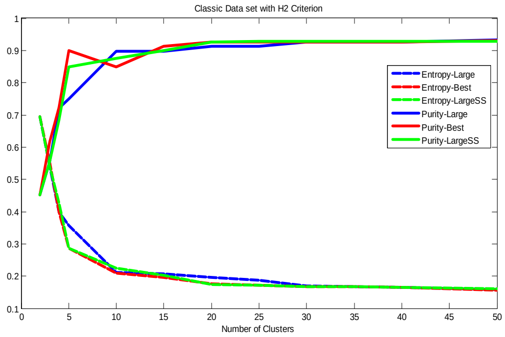

Fig.7

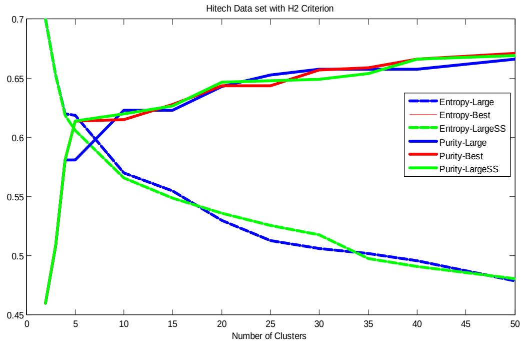

Fig.8

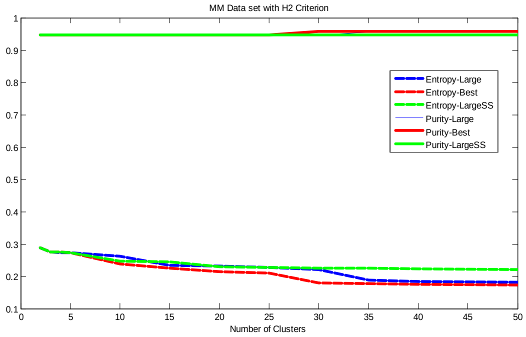

Fig.9

Fig.1 through **Fig.9** show the behavior of entropy and purity as we increase the number of clusters for the three data sets classic, hitech and mm. We have studied the behavior of entropy and purity by varying the clustering scheme used for the cluster split. We have analyzed for the three clustering schemes: large, best and largess.

Our results show that in most of the cases there is no significant difference between the three clustering schemes in terms of quality of the clusters obtained. Nevertheless, the "best" scheme is seen to be marginally better than the other two schemes.

Fig.10 to Fig.12 show the behaviour of the *I2* criterion function for the three schemes: best, large and largess as we increase the number of clusters, for the three datasets, classic, hitech and mm respectively. The Number of clusters is taken along x-axis and criterion condition *I2* is taken along y-axis.

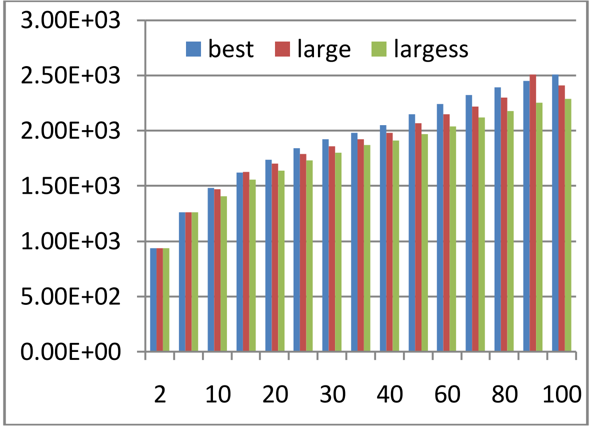

Fig.10

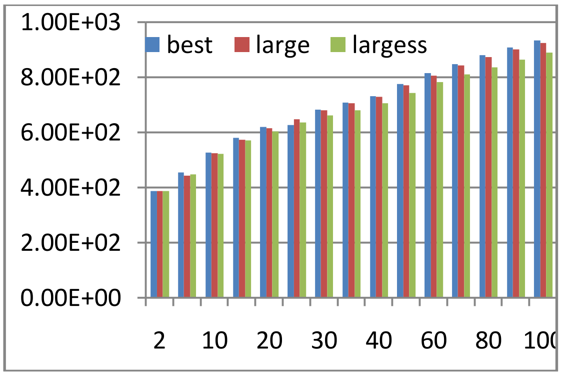

Fig.11

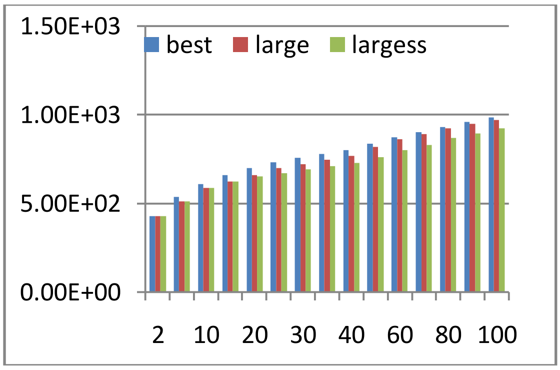

Fig.12

It can be seen that the *I2* criterion functional value continues to increase as we increase the number of clusters. Also, across all the three datasets the *I2* criterion function attains a maximum value only when we use the'best' clustering scheme for the cluster split.

V. CONCLUSION

We have experimentally shown that we can get better quality clusters, evaluated in terms of entropy and purity, by applying the 'best' clustering scheme. But from the graphs it can also be seen that the cluster quality obtained by using this scheme is only marginally better than that obtained by applying the other two schemes.

We have also shown that the *I2* criterion function is maximized, irrespective of the number of clusters, when the 'best' clustering scheme is used.

REFERENCES

[1] C. Charu, C. Stephen, S. Philip, On the Merits of Building Categorization Systems by Supervised Clustering. *In Proceeding of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 352-356, 1999.

[2] Y. Zhao, G. Karypis, "*Criterion functions for document clustering: Experiments and analysis". Technical Report TR #01–40,* Department of Computer Science, University of Minnesota, Minneapolis, MN, Feb 2002.

[3] Y. Zhao, G. Karypis, "Comparison of Agglomerative and Partitional Document Clustering Algorithms*". The SIAM workshop on Clustering High-dimensional Data and Its Applications, Washington, DC,* April 2002.

[4] George Karypis. *CLUTO: A Clustering Toolkit. Technical Report: #02-017.* University of Minnesota, Department of Computer Science. November 28, 2003

[5] Michael Steinbach, George Karypis, Vipin Kumar. *A Comparison of Document Clustering Techniques. Technical Report #00-034.* Department of Computer Science and Engineering. University of Minnesota. USA.

[6] Anna Huang, university of Waikato, Newzealand, Similarity Measures for Text Document Clustering. *In proceedings of the New Zealand Computer Science Research Student Conference 2008*

[7] Van de Cruys, Tim "*Mining for meaning: the extraction of lexico-semantic knowledge from text*" *Dissertation*, Evaluation of cluster quality, chapter 6 , University of Groningen , 2010 .

[8] Y. Zhao, G. Karypis. "*Soft Clustering Criterion Functions for Partitional Document Clustering". Technical Report #04-022,* Department of Computer Science, University of Minnesota, Minneapolis, MN, 2002 http://www.cs.umn.edu/˜karypis

[9] Fathi H.Saad, B.de la Iglesia and G.D.Bell. A Comparison of Two Document Clustering Approaches for Clustering Medical Documents. *Conference on Data Mining (DMIN 2006)*

[10] G.Hannah Grace, Kalyani Desikan, Estimation of the number of clusters based on cluster quality- *Submitted to Journal* .April 2013