Soft Clustering Criterion Functions for Partitional Document Clustering: A Summary of Results *

Ying Zhao
University of Minnesota, Department of Computer Science and Engineering
Minneapolis, MN 55455
yzhao@cs.umn.edu

George Karypis
University of Minnesota, Department of Computer Science and Engineering, Digital Technology Center, and Army HPC Research Center, Minneapolis, MN 55455
karypis@cs.umn.edu

ABSTRACT

Recently published studies have shown that partitional clustering algorithms that optimize certain criterion functions, which measure key aspects of inter- and intra-cluster similarity, are very effective in producing hard clustering solutions for document datasets and outperform traditional partitional and agglomerative algorithms. In this paper we study the extent to which these criterion functions can be modified to include soft membership functions and whether or not the resulting soft clustering algorithms can further improve the clustering solutions. Specifically, we focus on four of these hard criterion functions, derive their soft-clustering extensions, and present an experimental evaluation involving twelve different datasets. Our results show that introducing softness into the criterion functions tends to lead to better clustering results for most datasets.

Categories and Subject Descriptors: I.5.3 Clustering, Algorithms.

General Terms: Algorithms, Experimentation.

Keywords: Document clustering, Soft clustering.

1. INTRODUCTION

In recent years, soft clustering algorithms have been studied in document clustering and shown to be effective [4] in finding both overlapping and non-overlappying clusters. Studies have shown that "hardening" the results obtained by fuzzy C-means produces better hard clustering solutions than direct K-means [3], which suggests that including soft membership functions into other criterion functions may lead to better hard clustering solutions as well.

*This work was supported in part by NSF CCR-9972519, EIA-9986042, ACI-9982274, ACI-0133464, and ACI-0312828; the Digital Technology Center at the University of Minnesota; and by the Army High Performance Computing Research Center (AHPCRC) under the auspices of the Department of the Army, Army Research Laboratory (ARL) under Cooperative Agreement number DAAD19-01-2-0014. The content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. Access to research and computing facilities was provided by the Digital Technology Center and the Minnesota Supercomputing Institute.

CIKM'04, November 8–13,2004,Washington,DC,USA
ACM 1-58113-874-1/04/0011.

Recently, we studied seven different hard partitional clustering criterion functions in the context of document clustering, which optimize various aspects of intra-cluster similarity, inter-cluster dissimilarity, and their combinations [5]. The focus of this paper is to extend four of these hard criterion functions ($\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{E}_1$, $\mathcal{G}_1$ [5]) to allow soft membership functions, and to see whether or not introducing softness into these criterion functions leads to better clustering solutions. These criterion functions were selected because they include some of the best- and worst-performing schemes, and represent some of the most widely-used criterion functions for document clustering. In particular, the $\mathcal{I}_1$ criterion function maximizes the sum of the average pairwise similarities between the documents assigned to each cluster. $\mathcal{I}_2$ maximizes the similarity between each document and the centroid of the cluster that is assigned to. $\mathcal{E}_1$ minimizes the similarity between the centroid vector of each cluster and the centroid vector of the entire collection. $\mathcal{G}_1$ minimizes the edge-cut of each partition scaled by the internal edges.

We developed a hard-clustering based optimization algorithm that optimizes the various soft criterion functions. Although the experimental results show some dataset dependency, for most datasets the soft criterion functions tend to lead to better clustering results.

2. CRITERION FUNCTIONS

In our study we used the vector-space model and $tf - idf$ term weighting model to represent each document. Let n and k denote the number of documents and the number clusters, respectively. Let S denote the set of n documents that we want to cluster, $S_1, \ldots, S_k$ denote each one of the k clusters, $C_1, \ldots, C_k$ denote the centroids, and $n_1, \ldots, n_k$ denote the sizes of the corresponding clusters. If we use cosine as the similarity measure, then the various criterion functions can be written as in Table 1.

A natural and straight-forward way of deriving *soft* clustering solutions is to assign each document to multiple clusters. This is usually achieved by using membership functions [4, 2, 1] that for each document d_i and cluster S_j, they compute a non-negative weight, denoted by $\mu_{i,j}$, such that $\sum_j \mu_{i,j} = 1$, which indicates the extent to which document d_i belongs to cluster S_j. We define the size of the rth soft cluster $\overline{n}_r$ as $\sum_{i=1}^{n} \mu_{i,r}$, and the centroid of rth soft cluster $\overline{C}_r$ as $\sum_{i=1}^{n} \mu_{ir} d_i / \overline{n}_r$. Using the membership functions, centroids and sizes for soft clusters, we extended the various hard criterion functions and show the soft ones in Table 1 as well.

3. SOFT PARTITIONAL CLUSTERING ALGORITHM

We developed a soft partitional clustering algorithm that determines the values of the membership functions of the various documents following the induced fuzzy partitioning approach [1], and

Table 1: Clustering Criterion Functions.

	Hard Criterion Functions		Soft Criterion Functions
$\mathcal{I}_1$	maximize $\sum_{r=1}^{k} n_r \left(\frac{1}{n_r^2} \sum_{d_i,d_j \in S_r} \cos(d_i,d_j) \right)$	$\mathcal{SI}_1$	maximize $\sum_{r=1}^{k} \overline{n}_r \left(\frac{1}{\overline{n}_r^2} \sum_{i,j} \mu_{i,r}\mu_{j,r} \cos(d_i,d_j) \right)$
$\mathcal{I}_2$	maximize $\sum_{r=1}^{k} \sum_{d_i \in S_r} \cos(d_i, C_r)$	$\mathcal{SI}_2$	maximize $\sum_{r=1}^{k} \left(\sum_{i=1}^{N} \mu_{i,r} \cos(d_i, \overline{C}_r) \right)$
$\mathcal{E}_1$	minimize $\sum_{r=1}^{k} n_r \cos(C_r, C)$	$\mathcal{SE}_1$	minimize $\sum_{r=1}^{k} \overline{n}_r \cos(\overline{C}_r, C)$
$\mathcal{G}_1$	minimize $\sum_{r=1}^{k} \frac{\sum_{d_i \in S_r, d_j \notin S_r} \cos(d_i,d_j)}{\sum_{d_i,d_j \in S_r} \cos(d_i,d_j)}$	$\mathcal{SG}_1$	minimize $\sum_{r=1}^{k} \frac{\sum_{i,j} \mu_{i,r}(1-\mu_{j,r}) \cos(d_i,d_j)}{\sum_{i,j} \mu_{i,r}\mu_{j,r} \cos(d_i,d_j)}$

optimizes the soft criterion functions using a hard-clustering based optimization approach.

Given a hard k-way clustering solution $\{S_1, S_2, ..., S_k\}$, we define the membership of document d_i to cluster S_j to be

$$\mu_{i,j} = \frac{\cos(d_i, C_j)^m}{\sum_{r=1}^{k} \cos(d_i, C_r)^m}, \quad (1)$$

where C_r is the centroid of the hard cluster S_r.

The parameter m in Equation 1 is the *fuzzy factor* and controls the "softness" of the membership function and hence the "softness" of the clustering solution (the inclusion of the fuzzy factor was motivated by the formulation of the fuzzy C-means algorithm). In general, the softness of the clustering solution increases as the value of m decreases and vice versa.

Our proposed hard-clustering based optimization approach results in a pair of clustering solutions: a hard clustering solution and the induced soft clustering solution. In this paper, we focus on the hard clustering solution and used a clustering approach that determines the overall k-way clustering solution by performing a sequence of cluster bisections. During each step, we bisect the largest cluster available at that point of the clustering solution. Each of these bisections is performed in two steps. During the first step, an initial clustering solution is obtained by randomly assigning the documents to two clusters. During the second step, the initial clustering is repeatedly refined so that it optimizes the desired clustering criterion function.

The refinement strategy consists of a number of iterations. During each iteration, the documents are visited in a random order. For each document, d, we compute the change in the value of the soft criterion function obtained by moving d to the other cluster. This is done by deriving the membership values for the original and modified hard clustering solution and then calculate the change of the soft criterion function. If the change improves the criterion function, then d is moved to the cluster. The refinement phase ends, as soon as we perform an iteration in which no documents moved between clusters.

The time complexity of each iteration of the refinement of a 2-way clustering of a set of l documents is $O(l^2)$. If we assume that each successive bisection splits the documents into two roughly equal-size clusters then the overall amount of time required to compute all $k-1$ bisections is $O(n^2)$.

4. EXPERIMENTAL RESULTS

We experimentally evaluated the performance of the various soft criterion functions and compared them with the corresponding hard criterion functions using a number of different datasets.

In our experiments, we used a total of twelve different datasets. The smallest of these datasets contained 356 documents and the largest contained 1,170 documents. To ensure diversity in the datasets, we obtained them from different sources [5, 4]. For all datasets, we used a stop-list to remove common words, and the words were stemmed using Porter's suffix-stripping algorithm.

For each one of the different datasets we obtained a 10-way clustering solution that optimized the various hard and soft clustering criterion functions. Specifically, for each hard criterion function, we compared it with the corresponding soft criterion functions with the fuzzy factor m that achieves the best performance for each dataset. The quality of a clustering solution was evaluated using the ***entropy*** measure [5] that is based on how the various classes of documents are distributed within each cluster.

Table 2: Comparison of the Hard and Soft Criterion Functions.

	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$
Soft > Hard	8	10	7	9
Soft $\gg$ Hard	6	3	2	3

Table 2 shows the relative performance of the various soft criterion functions over the corresponding hard ones. The row labelled "Soft > Hard" shows the number of datasets on which soft criterion functions outperformed hard ones, whereas, the row labelled "Soft $\gg$ Hard" shows the number of datasets on which the improvement is more than 10%. As shown in Table 2, for most datasets, introducing softness improved the quality of the clustering solutions for most datasets. $\mathcal{SI}_2$ achieved the improvements more consistently, whereas, the improvements achieved by $\mathcal{SI}_1$ are most significant. The experiments on the effect of different fuzzy factor values are not shown in this paper due to the space limitation. The results show that the fuzzy factor values that achieved the best clustering solutions seemed to vary for different datasets, which suggests that the proper fuzzy factor values may relate to some characteristics of the datasets and their class conformations.

5. CONCLUSION

In this paper we extended four criterion functions that were studied in our previous work [5] to tackle the soft document clustering problem. We developed an approach similar to the induced fuzzy partition [1] to optimize various soft criterion functions. Our experimental results show that the soft criterion functions tend to lead to better clustering results for most datasets.

6. REFERENCES

[1] E. Backer. *Cluster Analysis by Optimal Decomposition of Induced Fuzzy Sets*. Delft University Press, Delft, The Netherlands, 1978.

[2] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[3] G. Hamerly and C. Elkan. Alternatives to the k-means algorithm that find better clusterings. In *Proc. of Int'l. Conf. on Information and Knowledge Management (CIKM-02)*, pages 600–607p, 2002.

[4] M. E. S. Mendes and L. Sacks. Evaluating fuzzy clustering for relevance-based information access. In *Proc. of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2003*, pages 648–653, May 2003.

[5] Ying Zhao and George Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, 2004.