Partitional Clustering Experiments on Document Datasets

December 31, 2009

Abstract

The purpose of this study is evaluation and comparison of some criterion functions used for document clustering. Each function is evaluated by using different clustering methods and different datasets. Detailed experiments show that some clustering criterion functions perform better than rest. Results of experiments are also consistent with previous works which compares same criterion functions.

1 Introduction

Document clustering is the unsupervised organization of documents. Documents are grouped such that related documents are assigned to the same group. It is a technique that can be used for automatic document organization, fast information retrieval.

In this work, the success of a clustering solution is measured by some criterion function. The clustering operation is treated as an optimization of the selected criterion function. Selected clustering method tries to optimize selected function to make the clustering solution better iteratively. In this work, some such criterion functions are evaluated and compared. Different clustering methods and different datasets are used. Results are consistent with previous related works.

The organization of this report is as follows. Section 2 gives information about previous works related to this study. Section 3 gives the relevant definitions and notations. It also gives the definitions of clustering criterion functions which are compared in this study and it explains different clustering methods used in this work. Section 4 is about experiments done. It contains information about datasets used, experiment setup. Detailed experiment results and analysis of them are given in this section. Section 5 concludes the report.

2 Related Work

In [1], different types of clustering techniques, namely partitional and hierarchical clustering techniques are compared with experiments. Related works on which this study based are [2] and [3]. They compare several clustering criterion functions, some of which are also evaluated in this study. These papers give results of experiments in detail as well as some theoretical analysis on these clustering criterion functions.

3 Document Clustering

Some definitions and notations used throughout this report are as follows.

Document representation In this study, each document is represented by using *term frequency-inverse document frequency* model. Each document d is represented as a vector d_{tf-idf}.

$$d_{tf-idf} = \begin{bmatrix} tf_1 log(n/df_1) & tf_2 log(n/df_2) & \dots & tf_m log(n/df_m) \end{bmatrix} \tag{1}$$

where tf_i is the frequency of the ith term in d, n is the total number of documents, df_i is the number of documents that contain ith term.

Similarity measure To measure the similarity between documents, *cosine similarity* measure is used. Cosine similarity between two documents d_i and d_j is defined as

$$cos(d_i, d_j) = \frac{{d_i}^T}{\|d_i\|\|d_j\|} \tag{2}$$

Symbols Some symbols that are used throughout this report are given below.

n : the number of documents
m : the number of terms
k : the number of clusters
S : the set of n documents to be clustered
S_i : ith cluster
n_i : number of documents in S_i.

By using these definitions and symbols, the document clustering problem can be defined as assigning each document d in S of size n to one of the subsets $\{S_1, S_2, \dots, S_k\}$, such that similar documents should be assigned to the same subset S_i.

3.1 Clustering Criterion Functions

3.1.1 Internal Criterion Functions

This type of criterion functions measure only the *internal* similarity of each clusters. They only quantify the similarity of documents which belong to the same cluster.

The first internal criterion function maximizes the average pairwise similarities between documents in the same cluster. It can be written as:

$$\mathcal{I}_1 = \sum_{r=1}^{k} n_r \left(\frac{1}{n_r^2} \sum_{d_i, d_j \in S_r} cos(d_i, d_j) \right) \tag{3}$$

The second internal criterion function maximizes the similarities between each document in the same cluster too.

$$\mathcal{I}_2 = \sum_{r=1}^{k} \sqrt{\sum_{d_i, d_j \in S_r} cos(d_i, d_j)} \tag{4}$$

3.1.2 External Criterion Functions

This type of criterion functions measure the *external* similarity (i.e. similarity between documents which are in different clusters). The only function of this type is defined as follows:

$$\mathcal{E}_1 = \sum_{r=1}^{k} n_r \frac{\sum_{d_i \in S_r, d_j \in S} cos(d_i, d_j)}{\sqrt{\sum_{d_i, d_j \in S_r} cos(d_i, d_j)}} \tag{5}$$

3.1.3 Hybrid Criterion Functions

Hybrid criterion functions compute how well a clustering solution is by taking account both internal similarity and external similarity. Hybrid criterion functions evaluated in this study are constructed by combining internal and external criterion functions explained in previous subsections. Two functions of this type are as follows:

$$\mathcal{H}_1 = \frac{\mathcal{I}_1}{\mathcal{E}_1} \tag{6}$$

$$\mathcal{H}_2 = \frac{\mathcal{I}_2}{\mathcal{E}_1} \tag{7}$$

3.1.4 Graph Based Criterion Functions

Unlike criterion functions that are explained in previous subsections, for this type of criterion functions, graphs are used to represent documents and similarities between them. There is only one criterion function of this type used in experiments. For that function, the graph is document to document similarity graph. Each node represents a document and the weight of edges between nodes represents the similarity between documents. It minimizes the edge-cut of each partition (i.e. cluster) and maximizes the similarity of documents that are in the same cluster. It can be formulated as follows:

$$\mathcal{G}_1 = \sum_{r=1}^{k} \frac{\text{cut}(S_r, S - S_r)}{\sum_{d_i, d_j \in S_r} cos(d_i, d_j)} \tag{8}$$

3.2 Clustering Methods

Three different clustering methods are evaluated. They are all partitional clustering algorithms which divide the set of documents into non-overlapping subsets (clusters) such that each document is in exactly one subset, unlike hierarchical clustering. They are all greedy approaches and they make some criterion function locally optimum at each step.

3.2.1 Direct k-way Clustering

Steps of the direct k-way clustering method are as follows:

1. Randomly select k documents as *seed* documents of k clusters.
2. Assign each document to the most similar seed document.

3. In a random order, assign each document to another cluster if it makes an improvement for the criterion function.

4. Repeat step 3 until no change.

Note that, the method *locally* optimizes the clustering criterion function at each iteration of the third step. Also at first step, seed documents are selected randomly. So, at the end of the method, criterion function may not be globally optimum. To handle the problem of possible bad results which are not close to the optimum solution, each experiment is performed 10 times and the one that makes the criterion function optimum most is selected.

3.2.2 k-way Clustering with Repeated Bisections

In the beginning of this method, entire set of documents can be considered as one cluster. k clusters are obtained by repeating the operation of bisecting a cluster, $k-1$ times. At each step, a cluster is selected for bisection such that it makes the criterion function optimum. Like direct k-way clustering, each experiment is repeated 10 times and the best one is selected.

3.2.3 k-way Clustering with Repeated Bisections followed by k-way Refinement

This method consists of two main steps. The first step is repeated bisecting method described in Section 3.2.2. After that, at the second step, the solution is refined as in Section 3.2.1, step 3. Like the previous methods, each experiment is repeated 10 times and the best one is selected as the result.

4 Experiments

4.1 Experiment Setup

For clustering experiments, CLUTO[1] was used. Criterion functions and clutering methods used in this study are implemented on this tool. All experiments were run on a system with Intel Pentium M Processor 1.60GHz and 512MB of memory.

Four different document sets were used in experiments. These document datasets are summarized in Table 1. The largest dataset is 20News which has 18821 documents and the smallest one is WebKB which has 4199 documents. 20Newsgroups[2] consists of 1000 Usenet articles for 20 different newsgroups. The second and third ones are preprocessed versions of the dataset Reuters21578[3], R8 and R52[4] which are collections of single-labeled documents. R8 has 8 different classes and R52 has 52 different classes. The last dataset, The 4 Universities Data Set[5] contains WWW-pages collected from computer science departments of various universities. In the version of this dataset used in this study, two of seven classes are

[1] http://glaros.dtc.umn.edu/gkhome/views/cluto/

[2] 20Newsgroups dataset is available at http://archive.ics.uci.edu/ml/databases/20newsgroups/20newsgroups.html. The same dataset but cleaned-up, stemmed and stop-word-free version is available at http://web.ist.utl.pt/~acardoso/datasets/. This version was used.

[3] The original dataset is available at http://www.daviddlewis.com/resources/testcollections/reuters21578/

[4] Available at http://web.ist.utl.pt/ acardoso/datasets/

[5] Available at http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

discarded and also the class "other" which contains very different documents is discarded[6]. For each experiment, documents in datasets are stemmed and do not contain any stop words.

Dataset	# of documents	# of classes
20News	18821	20
R8	7674	8
R52	9100	52
WebKB	4199	4

Table 1: Summary of dataets used

Six different clustering criterion functions were evaluated in this study. Three different clustering methods were used. Each experiment were run for four different k values, where k is the number of desired clusters. Experiments were run for $k = 5$, $k = 10$, $k = 15$ and $k = 20$.So the total number of experiments are $6 \times 3 \times 4 \times 4 = 288$. Results of each of them are given in Section 4.

The first measure used for evaluation in this study is entropy. For a cluster S_r of size n_r, the entropy of it is defined as follows:

$$E(S_r) = -\frac{1}{logq} \sum_{i=1}^{q} \frac{n_r^i}{n_r} log \frac{n_r^i}{n_r} \tag{9}$$

where q is the number of classes in dataset and n_r^i is the number of documents of the class i in the rth cluster. The entropy of the entire clustering solution is defined as

$$Entropy = \sum_{r=1}^{k} \frac{n_r}{n} E(S_r) \tag{10}$$

Second measure used is purity. For a cluster S_r of size n_r, the purity of it is defined as

$$P(S_r) = \frac{1}{n_r} \max_i (n_r^i) \tag{11}$$

where n_r^i is the number of documents of class i assigned to the rth cluster. The overall purity is

$$Purity = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r) \tag{12}$$

Smaller values of entropy and larger values of purity means better clustering.

4.2 Experiment Results and Analysis

As stated in [2], simply averaging the entropy and purity results may distort the overall results. The reason is that for different k values and for different datasets, the clustering quality differs. So, the same method applied in [2] is used. Average relative entropy and average relative purity values are used for summarizing results. **Relative entropy** of a clustering criterion function for a particular k value and a particular dataset is the entropy of that clustering criterion function divided by the smallest entropy over all clustering criterion

[6]This version is available at http://web.ist.utl.pt/ acardoso/datasets/

functions for that k and dataset. It is the degree to which a criterion function performed worse than the criterion function performed best. **Relative purity** of a criterion function for a particular k value and dataset is calculated by dividing the highest purity value over all criterion functions for that k and dataset (the best-achieved purity) by the purity of that clustering criterion function. A criterion function whose relative entropy and relative purity values closer to 1.0 is considered as successful.

The first experiment was run by using direct k-way clustering method. The entropy and purity values for $k = 5$, $k = 10$, $k = 15$ and $k = 20$ are given in Table 2. Average relative entropy and average relative purity values for experiments run with direct k-way clustering method are shown in Table 3. The function $\mathcal{E}_1$ has the best performance when $k = 5$. $\mathcal{I}_1$ performs worst for all k values. Functions $\mathcal{I}_2$ and $\mathcal{H}_2$ has the best enropy values where $\mathcal{I}_2$ performs better in average. The same two clustering criterion functions $\mathcal{I}_2$ and $\mathcal{H}_2$ has also the best purity values. For purity, $\mathcal{H}_2$ performs best for almost all k values and also in average. Again, the worst one is $\mathcal{I}_1$.

Entropy

	5-way clustering						10-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.748	0.611	0.632	0.618	0.648	0.625	0.631	0.460	0.465	0.531	0.496	0.450
R8	0.385	0.295	0.260	0.301	0.313	0.267	0.309	0.181	0.227	0.213	0.215	0.195
R52	0.393	0.320	0.280	0.330	0.332	0.310	0.294	0.234	0.236	0.265	0.244	0.226
WebKB	0.663	0.592	0.600	0.609	0.630	0.578	0.625	0.612	0.611	0.579	0.615	0.587
	15-way clustering						20-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.598	0.431	0.435	0.464	0.447	0.425	0.565	0.406	0.392	0.431	0.410	0.375
R8	0.267	0.170	0.213	0.157	0.174	0.185	0.221	0.154	0.192	0.166	0.176	0.174
R52	0.225	0.186	0.222	0.232	0.219	0.219	0.222	0.193	0.212	0.200	0.199	0.206
WebKB	0.659	0.597	0.618	0.586	0.604	0.616	0.653	0.625	0.598	0.616	0.616	0.599

Purity

	5-way clustering						10-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.205	0.251	0.250	0.248	0.239	0.251	0.340	0.449	0.456	0.405	0.420	0.465
R8	0.673	0.759	0.800	0.738	0.771	0.793	0.745	0.868	0.828	0.829	0.843	0.861
R52	0.569	0.655	0.686	0.634	0.634	0.686	0.695	0.743	0.714	0.696	0.736	0.729
WebKB	0.644	0.682	0.667	0.659	0.646	0.687	0.661	0.655	0.647	0.664	0.659	0.672
	15-way clustering						20-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.384	0.517	0.539	0.522	0.508	0.549	0.431	0.585	0.631	0.572	0.593	0.657
R8	0.788	0.874	0.833	0.906	0.879	0.868	0.853	0.889	0.852	0.886	0.874	0.870
R52	0.755	0.785	0.727	0.732	0.734	0.745	0.744	0.778	0.736	0.786	0.763	0.748
WebKB	0.632	0.666	0.659	0.667	0.649	0.662	0.630	0.639	0.678	0.637	0.642	0.672

Table 2: Entropy and purity values obtained using *direct k-way* clustering

The second experiment was run by using k-way clustering with repeated bisections. Results showing entropy and purity values for each test case are shown in Table 4. Average relative entropy and average relative purity values for each k value and criterion function are shown in Table 5. Results are similar to the results obtained by using direct k-way clustering. But, in this experiment, $\mathcal{I}_2$ performs best for all k values for both entropy and purity. The second best performing criterion function is $\mathcal{H}_2$. The worst function is again $\mathcal{I}_1$.

The third experiment was run by using k-way clustering with repeated bisections, but with k-way refinement after bisections. All entropy and purity values for each k, criterion function and dataset are given in Table 6. Also, average relative entropy and average relative purity results are given in Table 7. Like results obtained by using direct k-way method, when

Average relative entropy

k	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
5	1.313	1.075	**1.018**	1.100	1.135	1.039
10	1.372	1.028	1.096	1.132	1.107	**1.022**
15	1.360	**1.028**	1.157	1.084	1.092	1.101
20	1.295	**1.031**	1.097	1.073	1.072	1.049
avg	1.335	**1.040**	1.092	1.097	1.101	1.052

Average relative purity

k	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
5	1.171	1.027	1.008	1.055	1.058	**1.002**
10	1.154	1.015	1.036	1.068	1.041	**1.006**
15	1.168	**1.025**	1.049	1.031	1.052	1.026
20	1.174	1.048	1.038	1.054	1.052	**1.020**
avg	1.116	1.028	1.032	1.052	1.050	**1.013**

Table 3: Average relative entropy and average relative purity values obtained using *direct k-way clustering*. Bold and underlined values represent best performances.

$k = 5$ the function $\mathcal{E}_1$ performs best. For other k values, functions $\mathcal{I}_2$ and $\mathcal{H}_2$ are the best ones. As two previous experiments, $\mathcal{I}_1$ has worst performance. For entropy and purity, $\mathcal{I}_2$ has the best average performance, whereas its purity performance is same with $\mathcal{E}_1$.

5 Conclusion

In this study, six clustering criterion functions were compared on four different datasets for four different k values by using three different clustering methods. Detailed experiment results were given and these experiments showed that criterion functions $\mathcal{I}_2$ and $\mathcal{H}_2$ performs consistently well when compared with other ones. Also, $\mathcal{I}_1$ which seems similar to $\mathcal{I}_2$ has the worst entropy and purity values.

Results obtained are consistent with results of experiments run in [2] and [3], which compare some clustering criterion functions used in this study.

References

[1] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques, 2000.

[2] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis, 2001.

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

Entropy

	5-way clustering						10-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.752	0.631	0.645	0.675	0.653	0.643	0.668	0.487	0.511	0.533	0.534	0.508
R8	0.501	0.292	0.334	0.306	0.333	0.307	0.313	0.200	0.247	0.251	0.238	0.212
R52	0.397	0.321	0.322	0.327	0.367	0.325	0.314	0.247	0.238	0.269	0.262	0.230
WebKB	0.656	0.596	0.564	0.611	0.616	0.607	0.648	0.570	0.553	0.586	0.612	0.560
	15-way clustering						20-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.609	0.446	0.463	0.498	0.506	0.451	0.585	0.413	0.424	0.448	0.455	0.418
R8	0.214	0.156	0.187	0.203	0.196	0.188	0.193	0.125	0.172	0.158	0.183	0.173
R52	0.284	0.200	0.213	0.237	0.214	0.223	0.225	0.190	0.208	0.215	0.202	0.197
WebKB	0.635	0.565	0.542	0.581	0.595	0.553	0.624	0.554	0.537	0.577	0.591	0.551

Purity

	5-way clustering						10-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.201	0.246	0.243	0.229	0.239	0.245	0.297	0.428	0.423	0.382	0.396	0.410
R8	0.623	0.763	0.730	0.751	0.719	0.771	0.753	0.838	0.801	0.786	0.815	0.844
R52	0.561	0.656	0.663	0.644	0.615	0.655	0.607	0.713	0.713	0.696	0.705	0.743
WebKB	0.646	0.678	0.688	0.645	0.667	0.671	0.650	0.679	0.689	0.645	0.667	0.671
	15-way clustering						20-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.354	0.534	0.536	0.480	0.426	0.536	0.401	0.589	0.610	0.545	0.528	0.605
R8	0.875	0.890	0.866	0.836	0.872	0.855	0.891	0.926	0.882	0.883	0.872	0.880
R52	0.670	0.783	0.749	0.711	0.754	0.749	0.781	0.795	0.755	0.742	0.766	0.774
WebKB	0.656	0.682	0.690	0.655	0.674	0.686	0.660	0.690	0.690	0.657	0.679	0.686

Table 4: Entropy and purity values obtained using method *k-way clustering with repeated bisections*

Average relative entropy

k	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
5	1.326	**1.014**	1.042	1.054	1.102	1.039
10	1.368	**1.026**	1.079	1.144	1.133	1.028
15	1.332	**1.010**	1.075	1.168	1.139	1.087
20	1.326	**1.007**	1.124	1.138	1.182	1.114
avg	1.338	**1.014**	1.080	1.126	1.139	1.067

Average relative purity

k	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
5	1.177	**1.008**	1.017	1.049	1.052	1.010
10	1.211	**1.015**	1.026	1.082	1.050	1.017
15	1.187	**1.003**	1.018	1.083	1.085	1.023
20	1.155	**1.008**	1.025	1.072	1.067	1.023
avg	1.182	**1.008**	1.021	1.071	1.063	1.018

Table 5: Average relative entropy and average relative purity values obtained using method *k-way clustering with repeated bisections*. Bold and underlined values represent best performances.

Entropy

	5-way clustering						10-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.756	0.622	0.617	0.631	0.631	0.630	0.674	0.460	0.463	0.525	0.493	0.483
R8	0.490	0.295	0.264	0.300	0.323	0.271	0.310	0.203	0.216	0.235	0.223	0.203
R52	0.395	0.319	0.308	0.330	0.354	0.309	0.319	0.250	0.242	0.262	0.255	0.230
WebKB	0.663	0.609	0.575	0.601	0.630	0.591	0.642	0.587	0.595	0.601	0.611	0.582
	15-way clustering						20-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.622	0.405	0.413	0.483	0.464	0.406	0.586	0.371	0.375	0.443	0.398	0.379
R8	0.225	0.155	0.187	0.162	0.189	0.195	0.195	0.130	0.177	0.164	0.180	0.163
R52	0.289	0.207	0.208	0.231	0.218	0.223	0.240	0.194	0.203	0.215	0.197	0.195
WebKB	0.640	0.613	0.596	0.557	0.608	0.596	0.644	0.597	0.606	0.581	0.607	0.605

Purity

	5-way clustering						10-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.201	0.248	0.246	0.245	0.243	0.247	0.299	0.449	0.451	0.408	0.421	0.430
R8	0.641	0.759	0.800	0.738	0.732	0.792	0.765	0.840	0.838	0.803	0.832	0.856
R52	0.568	0.654	0.689	0.634	0.651	0.688	0.611	0.705	0.728	0.699	0.713	0.738
WebKB	0.642	0.669	0.679	0.653	0.658	0.657	0.656	0.675	0.651	0.663	0.669	0.668
	15-way clustering						20-way clustering					
Dataset	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
20News	0.350	0.573	0.581	0.502	0.460	0.580	0.402	0.639	0.656	0.558	0.590	0.636
R8	0.869	0.890	0.868	0.869	0.867	0.858	0.890	0.917	0.874	0.908	0.877	0.888
R52	0.681	0.770	0.750	0.723	0.744	0.746	0.774	0.784	0.755	0.769	0.774	0.774
WebKB	0.652	0.654	0.655	0.679	0.669	0.655	0.642	0.659	0.652	0.661	0.660	0.649

Table 6: Entropy and purity values obtained using *k-way clustering with repeated bisections followed by k-way refinement*

Average relative entropy

k	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
5	1.379	1.055	**1.000**	1.068	1.122	1.019
10	1.370	1.023	1.036	1.117	1.082	**1.012**
15	1.383	**1.025**	1.075	1.088	1.127	1.101
20	1.356	**1.006**	1.115	1.140	1.129	1.080
avg	1.372	**1.027**	1.056	1.103	1.115	1.053

Average relative purity

k	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{E}_1$	$\mathcal{G}_1$	$\mathcal{H}_1$	$\mathcal{H}_2$
5	1.188	1.030	**1.002**	1.055	1.050	1.024
10	1.216	1.017	1.018	1.061	1.036	**1.014**
15	1.214	**1.013**	1.022	1.061	1.084	1.026
20	1.176	**1.007**	1.025	1.051	1.042	1.023
avg	1.198	**1.016**	**1.016**	1.057	1.053	1.021

Table 7: Average relative entropy and average relative purity values obtained using method *k-way clustering with repeated bisections*. Bold and underlined values represent best performances.