

A Comparison of Two Document Clustering Approaches for Clustering Medical Documents

Fathi H. Saad¹, B. de la Iglesia¹, and G. D. Bell²

¹School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK.

²Endoscopy Unit, Norfolk and Norwich University Hospital, Colney Lane, Norwich NR4 7UY

Abstract— Medical data is often presented as free text in the form of medical reports. Such documents contain important information about patients, disease progression and management, but are difficult to analyse with conventional data mining techniques due to their unstructured nature. Clustering the medical documents into small number of meaningful clusters may facilitate discovering patterns by allowing us to extract a number of relevant features from each cluster, thus introducing structure into the data and facilitating the application of conventional data mining techniques. For this approach to work, it is essential to produce high-quality clustering. Thus, the main goals of this paper are (1) to experimentally evaluate the performance of six criterion functions in the context of partitional clustering approach, (2) to compare the clustering results of agglomerative approach and partitional approach for each of the criterion functions using real-world medical documents, and (3) to establish the right clustering algorithm to produce high quality clustering of real-world medical documents in order to discover hidden knowledge by analyzing the produced clusters. Our experimental results show that the clustering solutions produced by the agglomerative approach are consistently better than those produced by the partitional approach for all the criterion functions. Moreover, the results show that different criterion functions lead to substantially different results. In addition, we examine the quality of the features produced for each cluster for a classification task. The task involves discriminating between successful and unsuccessful procedures. The features extracted are used to produce an accurate classification of the data.

I. INTRODUCTION

Physicians' interpretations of images, signals, or any other clinical data, are written as unstructured free-text reports or documents. Such documents are very difficult to standardize and thus difficult to mine, even specialists from the same discipline cannot agree on unambiguous terms to be used in describing a patient's condition [1]. Developing methods or techniques to organizing large amount of unstructured clinical documents into a small number of meaningful clusters will help users to find what they are looking for, extract meaningful information and discovering trends and patterns hidden within these documents more effectively, because dealing with only the cluster that will contain relevant documents should improve effectiveness and efficiency [2]. The produced clusters contain groups of documents that are more similar to each other than to the members of any other group [3]. Therefore, the goal of

finding high-quality document clustering algorithms is to determine a set of clusters such that inter-cluster similarity is minimized and intra-cluster similarity is maximized. Since further knowledge extraction and data mining will be applied to the produced clusters, achieving high-quality clustering solution is important.

Document clustering has been investigated for usage in different areas such as browsing collections of document [5], improving the precision and recall in information retrieval systems [5], automatically generating hierarchical clusters of documents [6] etc. There are two clustering algorithms approaches based on the underlying methodology: agglomerative and partitional approaches [7]. There have been many conclusions derived in different studies that investigated the clustering performance of agglomerative and partitional approaches. The partitional clustering algorithms are well suited for clustering large documents datasets due to their relatively low computational requirements according to study conducted in [28]. In terms of clustering quality, work reported in [29] concluded that the partitional algorithms are actually inferior and less effective than their agglomerative counterparts. Using datasets from TREC and Reuters, Larsen and Aone [30] observed that agglomerative clustering outperformed various partitional clustering algorithms. None of these studies addressed the effect of the criterion functions. Criterion functions optimize the entire clustering process for both approaches. A recent study reported by Zhao and Kapyris [9] investigated the effect of the criterion functions to the problem of partitionally clustering documents and the results showed that different criterion functions lead to substantially different results. Another study reported in [7] investigated the effect of the criterion functions to partitional and agglomerative clustering algorithms using twelve document datasets obtained from various sources and their results showed that partitional algorithms always led to better clustering results than agglomerative algorithms.

Examination of the literature, show that there is no one strong conclusion recommending one approach. In addition, many of these have not been tested for clustering of medical text to find the best approach. Given the disparity of opinion and the importance of the real world problem being addressed, we set out to comprehensively evaluate the options available to prove empirically which approach is the best for clustering medical documents.

The main focus of this paper is to perform a practical

evaluation of various criterion functions in the context of the partitional approach, namely the repeated bisection clustering algorithm; then to compare the quality of the clusters produced by agglomerative and partitional algorithms from the perspective of different criterion functions; finally, to establish the right clustering algorithm to produce high quality clustering of real-world medical documents.

II. DOCUMENTS CLUSTERING ALGORITHMS

A. Hierarchical Agglomerative Clustering Algorithm

The agglomerative algorithm is bottom-up because it begins with the objects (documents in our case) as individual clusters and then repeatedly merges two clusters that are most similar until a single all-inclusive cluster is obtained [18, 20]. The main role of different clustering criterion functions that will be studied in the experiments is to determine the pairs of clusters to be merged at each step. There are two main computationally expensive steps in agglomerative clustering. The first step is the computation of the pairwise similarity between all the documents in the dataset. The second step is the repeated selection of the pair of clusters that best optimizes the criterion function [7].

TABLE I
SUMMARY OF DATA SETS USED IN THE EXPERIMENTS

Data	# of Documents	# of Distinct words
Colo_1	2158	1593
Colo_2	4837	4004
Endo_1	2113	825
Endo_2	3151	1004
Endo_3	3006	1157

B. Partitional Clustering Algorithm

There are two approaches to computing a k -way clustering of a set of documents in partitional clustering algorithms, either directly or via a sequence of repeated bisection [7]. The results of the study conducted in [9] show that the clustering solutions obtained via repeated bisections are better than those produced via direct clustering, and their computational requirements are much smaller. For those two reasons, we will use repeated bisection approach in all our experiments to compute partitional clustering solution. In the repeated bisection approach, a k -way clustering solution is obtained by first bisecting the entire collection. Then one of the two clusters is selected and it is further bisected, leading to a total of three clusters. The process of selecting and bisecting a particular cluster continues until k clusters are obtained. Each of these bisections is performed so that the resulting two-way clustering solution optimizes a particular criterion function [9]. The main step in this approach is the scheme used to select which cluster to bisect next; different approaches were described in [31, 32]. In our experiments we used the largest cluster, because this obtained reasonably good and balanced clustering solution

when we practically compare it with other schemes.

III. CLUSTERING CRITERION FUNCTIONS

As we mentioned earlier, six criterion functions will be included in our experiments I_1 , I_2 , ϵ_1 , G_1 , H_1 and H_2 . Those criterion functions can be classified in to four groups: *internal*, *external*, *graph-based* and *hybrid*. The internal criterion functions (I_1 , I_2) do not take into account the documents assigned to different clusters. They focus on producing a clustering solution that optimizes a particular criterion function that is defined over the documents that are part of each cluster [10]. I_1 is an example of internal criterion function that maximizes the sum of the average pairwise similarities between the documents assigned to each cluster; the size of each cluster determines the weight. Measuring the similarity between two documents using the cosine function, then the clustering solution must optimize the criterion function [7, 17]. I_2 is another example of internal criterion function. I_2 tries to find the clustering solution that maximizes the similarity between each document and the centroid of the cluster that is assigned to [9, 11]. The external criterion functions focus on optimizing a function that is based on how the various clusters are different from each other [11]. ϵ_1 tries to minimize the cosine between the centroid vector of each cluster to the centroid vector of the entire collection, to increase the angle between them as much as possible. The contribution of each cluster is weighted based on the cluster size [9]. The combination of different clustering criterion functions can define a set of *hybrid* criterion functions that simultaneously optimize multiple individual criterion functions. For example, H_1 is obtained by combining criterion I_1 with ϵ_1 , and H_2 is obtained by combining I_2 with ϵ_1 [17]. Finally, graph-based is an alternative way of viewing the relations between the documents. The motivation behind this criterion function is that the clustering process can be viewed as that of partitioning the documents into groups by minimizing the edge-cut of each partition. An example of graph-based criterion function is G_1 [11]. Detailed description of these criterion functions can be found in [9].

IV. EXPERIMENTS SETUP

A. Document Set

For these experiments, we used six real medical datasets from the Gastroenterology unit of a local hospital. The general characteristics are summarized in Table I. All data sets will be included in the comparison except “Colo_2” which will be used by the best clustering solution for the purpose of extracting features. We used different datasets to ensure diversity. The data sets “Colo_1” and “Colo_2” contain information about colonoscopy procedures. Colonoscopy refers to the passage of the colonoscope from the lowest part (anus and rectum) right around the colon to

the caecum and in some cases into the terminal ileum via the ileo-caecal valve. The aim of colonoscopy is to check for medical problems such as bleeding, colon cancer, polyps, colitis, etc [22]. The data sets “Endo_1”, “Endo_2” and “Endo_3” contain information about upper GI endoscopy procedures. Sometimes called EGD . Endoscopy is a visual examination of the upper intestinal track using an endoscope. The aim of endoscopy is to investigate swallowing difficulties, abdominal pain, chest pain, etc. [23]. The last data set “Sigmoid” contains information about the sigmoidoscopy procedure. Sigmoidoscopy is the visual examination of inside the rectum and sigmoid colon (the lower third of the colon) using an endoscope. Sigmoidoscopy is used to diagnose the cause of certain symptoms such as bleeding, diarrhea, pain, etc. [24]. After each colonoscopy, endoscopy or sigmoidoscopy procedure, the endoscopist writes a detailed report about the current status of the examined part and the procedure itself. The information contained in this report is extremely valuable for clinical purposes but difficult to handle with standard data mining techniques due to the lack of structure. The class labels of all different document sets are generated by Doc2Mat [25]. The largest document set contains 4,876 documents and the smallest document set contains 2,105 documents.

B. Document Pre-Processing

Not all the words in the documents are important, so they may degrade the classifier’s performance. In addition, representing small set of documents that may have hundreds of different words using *bag-of-words* approach will generate a huge feature space and thus will increase the processing time. To solve these problems, approaches to reduce the feature space dimension are needed. We used three approaches bellow as the same sequence:

- 1) As a result of consulting an expert in the domain field, we removed unhelpful sentences from the documents such as “Informed consent was obtained with the benefits, risks and alternatives for the procedure explained”, which is found in all reports;
- 2) We have removed stop words from all data sets using stop-lists containing common words such as “the”, “a”, “an”; the stopwords used are corpus-based.
- 3) We stemmed the words using Porter’s suffix-stripping algorithm [26]. Words are considered the same if they share the same stem.

C. Text Representation

The kind of linguistic features used in this paper to represent documents are single words. Single words are the structural units of language made up of one individual term [13]. The most frequently used method to represent text is bag-of-words representation where all words from the set of documents are taken and no ordering of words or any structure of text is used [26]. The different clustering algorithms used in our experiments use vector-space model [16] to represent each document. In this model, the

document d in the term space is considered to be a vector. The vector represents each document, $d_{tf} = (tf_1, tf_2, \dots, tf_n)$, where tf_i is the frequency of the i th term in the document. This model is refined by weighting each term based on its inverse document frequency. This weighting is needed because terms that appear frequently in many documents have limited discrimination power, so these terms must be de-emphasized. The new refined model is called term frequency-inverse document frequency (tf-idf) [16].

D. Performance Measures

A good clustering algorithm produces high-quality clusters such that inter-cluster similarity (external similarity) is minimized and intra-cluster similarity (internal similarity) is maximized. So the performance of a clustering algorithm is dependant on the quality of the produced clusters. Two cluster quality measures, *entropy* and *purity*, were used to measure the quality of the produced clusters of different algorithms. A perfect clustering solution will be the one that leads to clusters that contain documents from only a single class, in which case the entropy will be zero, and the purity will be 1.

Entropy - The best clustering solution will be the one that leads to clusters that contain documents from only a single class, in which case the entropy will be zero. In general, the smaller the entropy values, the better the clustering solution is [1, 4, 7, 12]. Entropy measures how various classes of documents are distributed within each cluster. First, the class distribution is calculated for each cluster, then this class distribution will be used to calculate the entropy for each cluster according to the following formula

$$E_j = - \sum_i p_{ij} \log(p_{ij}) \quad (1)$$

where p_{ij} is the probability that a member of cluster j belongs to class i and then the summation is taken over all classes. After the entropy is calculated, the summation of entropy for each cluster is calculated using the size of each cluster as weight. In other words, the entropy of all produced clusters is calculated as the sum of the individual cluster entropies weighted according to the cluster size, and defined as

$$E_{sc} = \sum_{j=1}^m \frac{n_j * E_j}{n} \quad (2)$$

where n_j is the size of cluster j , n is the total number of documents, and m is the number of clusters.

Purity- measures to which extend each cluster contained documents from primarily one class. In other words, it measures the largest class for each cluster. In general, the larger the values of purity, the better the clustering solution is [4, 7]. In similar way as entropy, the purity of each cluster is calculated as

$$P(S_r) = \frac{1}{n_r} \max_i (n_i^r) \quad (3)$$

where S_r is a particular cluster of size n_r . The purity of all produced clusters is computed as a weighted sum of the individual cluster purities and is defined as

$$Purity = \sum_{r=1}^k \frac{n_r}{n} P(S_r) \quad (4)$$

E. Experimental Methodology

The first set of experiments was focused on evaluating the quality of the clustering solutions produced by the various criterion functions in the context of repeated bisection algorithm. The second set of experiments was focused on comparing the quality of the produced clusters via agglomerative approaches. Then the clusters that were obtained via the best clustering solution were analyzed in order to extract features and discover hidden knowledge.

When a document clustering algorithm is used, the similarity between two documents must be measured. There are many similarity measures such as Tanimoto [27], cosine [12, 14], correlation coefficient [4, 11], Euclidean distance [6, 4] and extended Jaccard coefficient [4, 11]. We experimentally managed to evaluate cosine, correlation coefficient, Euclidean distance and extended Jaccard coefficient in order to find the best document similarity measure that produce the highest cluster quality and we found cosine is the best. This might be because cosine works well when documents are viewed using vector-space model as it explained in [7, 11]. For partitional clustering methods there are many schemes for selecting the clusters to be bisected next. These 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. We also managed to investigate the three schemes practically and the results show that “large” is the best. The detailed results of these experiments were omitted due to space limitation.

In these experiments, for each one of different datasets we obtained a 5-, 10-, 15-, and 20-way clustering solution that optimizes the various clustering criterion functions, in order to investigate the performance of the selected criterion functions to produce different number of clusters.

V. RESULTS AND ANALYSIS

The results of entropy and purity obtained via repeated bisection to evaluate various criterion functions are summarized by looking at the average performance of each criterion function over entire set of datasets. There are two ways to calculate the average. The first way is the simple averaging, which is calculated by summing the internal similarities, entropies or purities of a particular criterion function for the 5 data sets and then dividing by 5 (the number of data sets), but using simple averaging is not recommended by [9] because it may distort the overall results. The second way of averaging recommended by [9] is calculated by dividing, for example, the entropy obtained by a particular criterion function for each dataset and value of k (5-, 10-, 15- or 20) by the best entropy result which is the smallest entropy value obtained for that particular dataset and value of k over the different criterion functions. The

degree to which a particular criterion function performed worse than the best criterion function is represented by the calculated ratios, which will be referred as *relative entropies*. After the relative entropies are calculated for each criterion function and value of k , the average of these relative entropies over the various datasets is computed. A criterion function that has *average relative entropy* close to 1 will indicate that this function did the best for most of the datasets.

The inverse ratio is used to calculate the *averaged relative purity*. Since the higher values of purity are better. The inverse ratio is calculated by dividing the highest purity value (the best purity) by a particular purity value for each criterion function and value of k , and then averaged them over the various datasets. The average relative purity will be interpreted in a similar manner as those of the average relative entropy (they are good if they are close to 1 and they are getting worse as they become greater than 1).

A. Evaluating Criterion Functions via Partitional Clustering

The detailed results of the calculated relative entropy and purity of different datasets for each criterion functions for the clustering solution obtained via repeated bisection and agglomerative clustering algorithms for 5-, 10-, 15- and 20-way clustering solutions were omitted due to space limitation.

The calculated averaged relative entropies and averaged relative purities for the 5-, 10-, 15-, and 20-way clustering solutions produced via repeated bisection algorithms are shown in Table II. The columns labeled “Avg.” contain the simple average of these averaged relative values over the four sets of k -way clustering solutions.

A number of observations can be made by analyzing the results shown in Table II. In terms of entropy measures, the I_1 , I_2 and G_1 criterion functions lead to clustering solutions that are worse than the solutions obtained using the other criterion functions. They lead to solutions that are 11%–22% worse than the best solution. The E_1 and the H_2 criterion functions lead to the best solutions irrespective of the number of clusters. Over the entire set of experiments, these methods are either the best or always within 2% of the best solution. The H_1 criterion function always performs somewhere in the middle. It is on the average 7% worse when compared to the best scheme.

On the other hand, in terms of purity measures, all the observations that were made based on entropy measures are true when the quality of the clustering solution evaluated using the purity measures. The I_1 , I_2 and G_1 are the worst, they lead to solutions that are 3%–9% worse than the best solution. The E_1 and the H_2 criterion functions lead to the best solutions irrespective of the number of clusters. Over the entire set of experiments, these methods are either the best or always within 2% of the best solution. The H_1 criterion function always performs somewhere in the middle.

It is on the average 2% worse when compared to the best scheme. The relative performance of the various criterion functions remains more-or-less the same for both the entropy- and the purity-based evaluation methods. The only change is that the relative differences between the various criterion functions as measured by entropy are somewhat greater when compared to those measured by purity. This should not be surprising, as the entropy measure takes into account the entire distribution of the documents in a particular cluster and not just the largest class as it is done by the purity measure.

TABLE II
AVERAGED RELATIVE ENTROPIES AND PURITIES OVER DIFFERENT DATASETS FOR DIFFERENT CRITERION FUNCTIONS FOR THE CLUSTERING SOLUTION OBTAINED VIA REPEATED BISECTION CLUSTERING ALGORITHM FOR 5-, 10-, 15- AND 20-WAY CLUSTERING

Averaged Relative Entropy					
	5	10	15	20	Avg.
I_1	1.319	1.294	1.325	1.395	1.333
I_2	1.179	1.225	1.233	1.207	1.211
E_1	1.108	1.11	1.128	1.11	1.114
G_1	1.137	1.23	1.225	1.26	1.213
H_1	1.201	1.165	1.148	1.194	1.177
H_2	1.095	1.124	1.097	1.061	1.094
Averaged Relative Purity					
	5	10	15	20	Avg.
I_1	1.158	1.135	1.129	1.125	1.137
I_2	1.079	1.08	1.073	1.049	1.070
E_1	1.032	1.04	1.037	1.027	1.034
G_1	1.057	1.091	1.079	1.07	1.074
H_1	1.087	1.072	1.053	1.059	1.068
H_2	1.051	1.062	1.05	1.021	1.046

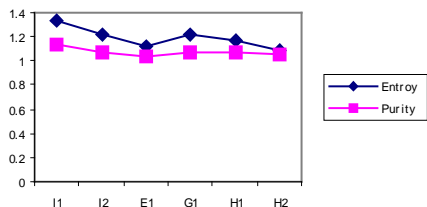


Fig. 1. The averaged relative entropy and averaged relative purity results for the six clustering criterion functions

Figure 1 illustrates graphically the averaged relative entropies and averaged relative purities results for the six clustering criterion functions. The axis x and y represent the criterion functions and the averaged relative values respectively. All the observations that were made based on the entropy and purity results can be seen clearly in Figure 1.

B. Comparison of Partitional vs. Agglomerative Clustering Solutions

The calculated averaged relative entropies and averaged relative purities for the 5-, 10-, 15-, and 20-way clustering solutions produced via agglomerative clustering algorithms are shown in Table III. The columns labelled “Avg.” contain the simple average of these averaged relative values over the four sets of k -way clustering solutions. Table IV shows the

averaged relative entropies and purities over different datasets for different criterion functions for the clustering solution obtained via agglomerative and partitional clustering algorithms. By analyzing the results shown in Table IV, a number of observations can be made. First, the clustering solutions produced by the agglomerative approach are consistently better than those produced by the partitional approach for all the criterion functions. Second, the relative performance of various criterion functions in agglomerative clustering algorithms does differ from the relative performance in partitional clustering algorithms as we found in [7] and [9]. Third, the improvements vary among the criterion functions. For criterion functions that perform poorly in agglomerative approach, but perform well in partitional approach, the improvement is considerable. For example, the E_1 criterion function performed the best in the partitional approach; it is improved by 1.6% in terms of entropy and by 2.1% in terms of purity. The column titled “Improvement %” in Table IV illustrates the detailed improvement for each criterion function for both average relative entropy and averaged relative purity. Figure 2 shows the averaged relative entropies and purities obtained by agglomerative and partitional approaches respectively for all criterion functions. The x axis in Figure 2 represents the criterion functions, and the y axis represents the averaged relative values. The same Figure also shows clearly that the agglomerative approach outperform the partitional approach for all the criterion functions in terms of entropy and purity.

TABLE III
AVERAGED RELATIVE ENTROPIES AND PURITIES OVER DIFFERENT DATASETS FOR DIFFERENT CRITERION FUNCTIONS FOR THE CLUSTERING SOLUTION OBTAINED VIA AGGLOMERATIVE CLUSTERING ALGORITHMS FOR 5-, 10-, 15- AND 20-WAY CLUSTERING

Averaged Relative Entropy					
	5	10	15	20	Avg.
I_1	1.053	1.041	1.063	1.060	1.054
I_2	1.045	1.072	1.067	1.063	1.062
E_1	1.057	1.104	1.094	1.097	1.088
G_1	1.059	1.047	1.054	1.059	1.055
H_1	1.059	1.077	1.067	1.056	1.065
H_2	1.038	1.093	1.087	1.089	1.077
Averaged Relative Purity					
	5	10	15	20	Avg.
I_1	1.079	1.041	1.043	1.030	1.049
I_2	1.060	1.045	1.026	1.017	1.037
E_1	1.037	1.028	1.033	1.033	1.033
G_1	1.063	1.024	1.020	1.020	1.032
H_1	1.070	1.050	1.038	1.023	1.045
H_2	1.016	1.030	1.026	1.025	1.024

C. Cluster Examination

The agglomerative algorithm was applied to “Colo_2” dataset to extract features from the data. “Colo_2” contains reports from 4,837 colonoscopies performed at the local hospital. For each colonoscopy, the report details information on the preparation of the bowel prior to the procedure, what was seen, any diagnoses, etc. A procedure is considered successful, if the colonoscope is inserted to the

caecum, ileum, terminal ileum, or if surgical anastomosis,

TABLE IV

AVERAGED RELATIVE ENTROPIES AND PURITIES OVER DIFFERENT DATA SETS FOR DIFFERENT CRITERION FUNCTIONS FOR THE CLUSTERING SOLUTION OBTAINED VIA AGGLOMERATIVE AND PARTITIONAL CLUSTERING ALGORITHMS

	Agglomerative		Partitional		Improvement %	
	Entropy	Purity	Entropy	Purity	Entropy	Purity
I_1	1.054	1.049	1.333	1.137	20.9	7.7
I_2	1.062	1.037	1.211	1.070	12.3	3.1
E_1	1.088	1.033	1.114	1.034	2.3	0.1
G_1	1.055	1.032	1.213	1.074	13.0	3.9
H_1	1.065	1.045	1.177	1.068	9.5	2.2
H_2	1.077	1.024	1.094	1.046	1.6	2.1

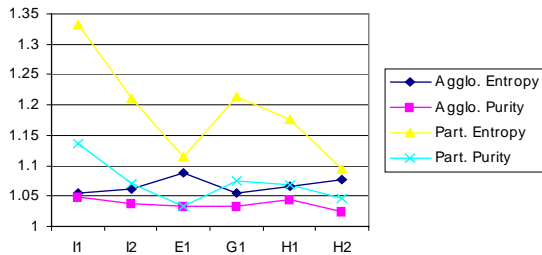


Fig. 2. The averaged relative entropies and purities obtained by agglomerative and partitional approach for all criterion functions

ileal anastomosis or ileocolic anastomosis is found. The percentage of caecal intubation gives a ‘crude’ success rate. There are cases in which the procedure are classified as unsuccessful but the reason for the surgeon not to get around the colon is an obstructing tumour which precludes further passage of the colonoscope. We would like to reclassify this as successful per protocol. There is an increased need to examine the outcomes of colonoscopy due to the introduction of a programme of colorectal cancer screening in the UK. At present, it is difficult for doctors to analyse their success/failure rate or the factors affecting this, as the information is buried in a large number of reports which are difficult to analyse due to their unstructured nature.

We performed an initial classification of successful intubation rate for the 4,837 reports. We did this using a combination of automatic methods, searching for regular expressions, and using experts to sift through the documents and provide a classification. The number of failures for this data set was 600 which represents a 12.4% failure rate. Table V shows the results achieved by applying the agglomerative clustering algorithm to this data. The tables contain information about cluster id (CID), the percentage of successful (S%) and unsuccessful (U%) colonoscopy procedures documents in each cluster and descriptive terms for each cluster, respectively. As can be observed in Table V, the clusters can be labelled as successful or unsuccessful based on the values of S% and U%, for example, clusters 12 and 15 could be labelled as unsuccessful. Also note that the percentages S% and U% much higher than on the whole data set. Furthermore, clusters can be labelled as successful or unsuccessful based on the meaningful descriptive terms and this has been confirmed by an expert in the domain field.

For example cluster 12 has descriptive terms such as “limited”, “inadequate” and “hepatic flexure” which are associated with failure of the colonoscopist to get around the colon. The clustering solution has imposed structure in the data which should aid in the classification and characterization of colonoscopies as successful or unsuccessful.

TABLE V

CLASSIFICATION OF PROCEDURES TO SUCCESSFUL OR FAILED FOR 20 CLUSTERS OBTAINED BY APPLYING AGGLOMERATIVE ALGORITHM TO “COLO_2” DATASET

CID	S%	U%	Descriptive terms
0	100	0	Terminal, ileum, appendiceal, orifice and caecal
1	100	0	Normal, repeat, insert, caecum and bowel
2	88.3	11.4	Ulcer, mucosa, granular, mucopurulent and exudate
3	97.7	2.3	Excised, retrieved, polyp, sessile and pedunculate
4	98.4	1.6	Physical, precluding, history, patient and colonoscopy
5	85	15	biopsy, colon, mucosa and normal
6	100	0	appendiceal , orifice, caecal and tri-radiate
7	97.3	2.7	rescope, plan, throughout, appear and mucosa
8	96.6	3.4	Poor, caecal, normal, caecum and ileo
9	92	8	Polyp, sessile, retrieved, excised and hot
10	100	0	Neo, terminal, ileum, normal and diverticula
11	100	0	caecal , transillumination , tri-radiate and fold
12	17.6	82.4	Limited, examination, flexure, hepatic and inadequate
13	99.3	0.7	Difficulty, caecal, appendiceal, orifice and picolax
14	50	50	Tumour, polypoid, encounter, pathologist and fungating
15	0	100	Unsuccessful, intubation, faeces, severe and obstruct
16	96.9	3.1	mucosa, granular, congest, erythematous and rectum
17	100	0	Normal, picolax, ileo, caecum and valve
18	96.1	3.9	Visualised, rest, diverticula, sigmoid and evident
19	97	3	diverticula , multiple, sigmoid, distal and colon

VI. CONCLUSION AND FUTURE WORK

In this paper we experimentally evaluated six different criterion functions for clustering large real-world medical data sets using the partitional approach and compared the clustering results obtained via partitional approach with those obtained via agglomerative approach for each one of the clustering criterion functions. Our experimental results showed that different criterion functions lead to substantially different results. The experimental results also showed that the clustering solutions produced by the agglomerative approach are consistently better than those produced by the partitional approach for all the criterion functions. In addition, the produced clusters facilitate examining the features produced for each cluster for a classification task. The task involves discriminating between successful and unsuccessful procedures. The features extracted used to produce an accurate classification of the data. The application of this to a real medical database of reports on colonoscopy procedures has shown the potential of the clustering algorithm as a tool to infer structure into the data and extract relevant terms. The application area is very important due to the forthcoming introduction in the UK of a colorectal cancer-screening program which will require improved outcomes for the procedure.

In future, we will continue to work in the interpretation of the clustering results and on other methods of extracting features from the data and introducing structure so that free text data can be analyzed usefully in conjunction with other

details of patients and outcomes kept in structured databases. In addition, there are new clustering techniques that could be investigated such as the constrained clustering technique introduced recently in [9].

REFERENCES

- [1] J. C. Krzysztof, G. W. Moore, "Uniqueness of Medical Data Mining", *Artificial Intelligence in Medicine journal*, 26(1-2), 1-24, 2002
- [2] C. J. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, L. M. Hage, W. E. Hammond, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse". American Medical Informatics Association Annual Fall Symposium (formerly SCAMC). pp. 101-5. 1997.
- [3] A. K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review". *ACM Computing Surveys*, Vol. 31, No. 3, September 1999.
- [4] SAS Institute Inc. *SAS Enterprise Miner*. 2004.
- [5] R. D. Cutting, D. R. Karger, J. O. Pedersen, J. W. Tukey, "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections", *SIGIR '92*, pp. 318 – 329, 1992.
- [6] D. Koller, M. Sahami, "Hierarchically classifying documents using very few words", *Proceedings of the 14th International Conference on Machine Learning (ML)*, Nashville, Tennessee, pp. 170-178, 1997.
- [7] 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.
- [8] A. Casillas, M. Lena, R. Martinez, "Partitional Clustering Experiments with News Documents". *Lecture Notes in Computer Science*, # 2588, ISSN 0302-9743, Springer-Verlag, pp. 615–618. 2003.
- [9] 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. <http://cs.umn.edu/~karypis/publications>.
- [10] M. Steinbach, G. Karypis, V. Kumar, "A Comparison of Document Clustering Techniques". In *KDD Workshop on Text Mining*, 2000.
- [11] Y. Zhao, G. Karypis, "Evaluation of Hierarchical Clustering Algorithms for Document Datasets". In *Proceedings of the 11th International conference on Information and knowledge management*, Virginia, USA, pp. 515-524, 2002.
- [12] U. Yong, J. R. Mooney, "Text Mining with Information Extraction". *AAAI Symposium on Mining Answers from Texts and Knowledge Bases*, Stanford, CA, 2002.
- [13] D. B. Aronow, F. Feng. "Ad-Hoc Classification of Electronic Clinical Documents". *D-Lib Magazine*. ISSN 1082-9873. 1997.
- [14] E. Claude, "A Mathematical Theory of Communication", *Bell System Technical Journal*, Vol. 27, pp. 379-423. 1948.
- [15] G. Karypis. "CLUTO: A Clustering Toolkit". Technical Report: #02-017. University of Minnesota, Department of Computer Science. November 28, 2003. Available at <http://www.cs.umn.edu/~karypis>
- [16] G. Salton, M. McGill, "Introduction to Modern Information Retrieval". McGraw-Hill. 1983.
- [17] 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>
- [18] S. Guha, R. Rastogi, K. Shim. "CURE: An efficient clustering algorithm for large databases". In *Proc. of 1998 ACM SIGMOD Int. Conf. On Management of Data*, 1998.
- [19] Y. Zhao, G. Karypis. "Hierarchical Clustering Algorithms for Document Datasets". Technical Report #03-027, Department of Computer Science, University of Minnesota, Minneapolis, MN, 2002. Available at <http://www.cs.umn.edu/~karypis>
- [20] R. Ali, U. Ghani, A. Saeed. "Data Clustering and Its Applications". Available at http://members.tripod.com/asim_saeed/paper.htm
- [21] A. K. Jain, R. C. Dubes. "Algorithms for Clustering Data". Prentice Hall, 1988.
- [22] C. J. Bowles, R. Leicester, C. Romaya, E. Swarbrick, C. B. Williams, O. Epstein. "A Prospective Study of Colonoscopy Practice in the UK today: are we Adequately Prepared for National Colorectal Cancer Screening Tomorrow?" *International Journal of Gastroenterology and Hepatology*. Jun 2003.
- [23] Jackson Gastroenterology, *Upper GI Endoscopy*. 2002. <http://www.gicare.com/pated/epdgs18.htm>
- [24] National Institute of Diabetes and Digestive and Kidney Diseases. *Flexible Sigmoidoscopy*. National Institutes of Health. Bethesda, MD. <http://digestive.niddk.nih.gov/ddiseases/pubs/upperendoscopy/index.htm>
- [25] G. Karypis. "Doc2Mat: Converting Documents into vector-space format". Program.
- [26] M. F. Porter, "An algorithm for suffix stripping", *Program; automated library and information systems*, 14(3), 130-137, 1980.
- [27] K. Lin, R. Kondadadi, "A Word-Based Soft Clustering Algorithm for Documents". Department of Mathematical Sciences, The University of Memphis, Memphis, TN 38152, USA. Available at http://www.cs.memphis.edu/~linki/_mypaper/CATA01.doc
- [28] 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.
- [29] P. Willett. "Recent Trends in Hierarchic Document Clustering: A Critical Review". In *Information Processing and Management*, 24(5): 577-597, 1988.
- [30] B. Larsen, C. Aone. "Fast and Effective Text Mining Using Linear-time Document Clustering". In *Proc. of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p 16-22. 1999
- [31] G. Karypis, E. H. Han. "Concept Indexing: A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval & Categorization". Technical Report TR-00-016, University of Minnesota, Minneapolis, 2000. Available on <http://www.cs.umn.edu/~karypis>.
- [32] M. Steinbach, G. Karypis, V. Kumar, "A Comparison of Document Clustering Techniques". In *KDD Workshop on Text Mining*, 2000.
- [33] H. Benbrahim, M. A. Barmer, "Neighborhood Exploitation in Hypertext Categorization". In *Research and Development in Intelligent Systems XXI*. Springer-Verlag, 2005.