Feature Selection In Document Clustering Using Rough Set Theory

Mostafa Abdel Aziem Mostafa
Faculty of Engineering & Technology
Arab Acadmy of Science and Technology

Hoda Saleh Ahmed
Faculty of Science
Al Azhar University
Itegypt2000@hotmail.com

Abstract

One fundamental aspect of rough set theory is the search of subsets of attributes that provide the same information for classification purposes as the full set of attributes. In this paper, application of rough set theory to feature selection in document clustering is introduced. We emphasize the role of the basic constructs of rough set approach in feature selection, namely reducts. We propose a method of generating a best reduct of the data based on rough set theory to overcome the problems of generating all reducts. The application to a hierarchical clustering of document dataset is presented as an example. Finally, the paper presents a comparison of the clustering results based on the original data set and those based on the reduced data set.

Keywords: Rough set theory, feature selection, feature extraction, document clustering, and data reduction

1. Introduction

Document clustering is the fundamental enabling tool for efficient document organization, summarization, navigation and retrieval for very large datasets. The most critical problem for text clustering is the high dimensionality of the natural language text [1]. Two different approaches for feature dimensionality reduction are feature selection and feature extraction. The former is to find a set of input variables, from given set of input candidates, which really affect the output. The later reduces the original set of features into a linearly or nonlinearly transformed set. Feature selection allows discarding some of the irrelevant features and better performance may be achieved by discarding such features [2].

Features gathered to cluster objects are often not all equally informative; some of them may be redundant or irrelevant for the clustering. Often many candidate features are included since the relevant features are unknown a priori. The identification of the relevant features can save future measurement time and costs or can be used for an, e.g. physicochemical interpretation of the studied phenomenon [3]. In the case of multidimensional datasets, usually the number of features can be significantly reduced with feature selection.

In this paper, we represent an approach to dimensionality reduction by applying the concept of rough sets. Rough set theory will be used to construct reducts for unsupervised hierarchical clustering. Our aim is to preserve cluster structure of data while eliminating redundant and costly features. We are satisfied if the reduced variable set produces similar clustering results as the complete feature set because we know that the complete feature set results in a taxonomic sense making clustering.

The rest of the paper is organized as follows: In Section 2 we have reviewed the basic concepts of Rough Set Theory. Section 3 describes the proposed method for feature selection based on Rough Set Theory. Section 4 gives a clustering overview. Section 5 gives experimental results and presents a comparison of the clustering results based on the original data set and those based on the reduced data set. Finally, conclusion on the work found in Section 6.

2. Rough Set Theory

Rough set theory introduced by Pawlak [4] in the early 1980, is a technique for dealing with uncertainty and for identifying cause–effect relationships in databases as a form of data mining and database learning. It has also been used for improved information retrieval and for uncertainty management in relational databases [5]. In recent years we witnessed a rapid growth of interest in rough set theory and its applications [6]. The rough set methodology has a wide variety of applications. Besides information-preserving data reduction, for which it is described here, it can be used for representation of uncertain or imprecise knowledge, identification and evaluation of data dependencies, reasoning with uncertainty, approximate pattern classification, knowledge analysis, etc. It is especially useful for nominal or discrete data. The central concept in rough set theory is to approximate a target set through crisp partitions generated by equivalence relation, by a pair of exact sets called the lower and the upper approximations [2].

In this section some basic concepts of rough set theory are presented and illustrated by simple examples. Rough set theory operates on an information system, which is made up of objects for which certain characteristics are known [7].

The notion of information system, sometimes called data tables, attribute-value systems, knowledge representation systems, etc. provides a convenient tool for description of objects in terms of their attribute values. An information system is a pair (U, A), where U is a non-empty finite set of objects called the universe and A is a non-empty finite set of attributes, such that a: U → Va for any $a \in A$, where Va is called the domain of a . Objects with the same attribute values are grouped into equivalence classes called elementary sets. Each non-empty subset B $\subseteq$A determines an indiscernibility relation as follows:

$R_B=\{(x,y)\in U \times U : a(x)=a(y) \text{ for all } a\in B\}$.

R_B partitions U into a family of disjoint subsets U/R_B called a quotient set of U: U/R_B={$[x]_A$:x∈B} where $[x]_A$: denotes the equivalence class determined by x with respect to B ,i.e.,

$[x]_A$={y∈U: (x, y) ∈ R_B}.

Equivalence classes of the relation R_B are referred to as B-elementary sets. In the rough set approach the elementary sets are the basic building block of our knowledge about reality.

Let X⊆U,B⊆A, one can characterize X by a pair of lower and upper approximations [8]:

$$\underline{R_B}(X) = \{x \in U : [x]_B \subseteq X\}$$

$$\overline{R_B}(X) = \{x \in U : [x]_B \bigcap X \neq \Phi\}$$

The lower approximation $\underline{R_B}(X)$ is the set of objects that belong to X with certainty, while the upper approximation $\overline{R_B}(X)$ is the set of objects that possibly belong to X.

The most interesting feature of rough set theory for our approach is the construction of reducts: all possible minimal subsets of features that lead to the same partitioning (in elementary sets) as the whole set. The common part of all reducts is called core and represents the set of all indispensable features. In practice, to compute reducts and core, the discernibility matrix D can be used. This matrix has dimensions n * n, where n denotes the number of elementary sets. Element d_{ij} is computed as the set of features, which discerns the elementary sets i and j.

d_{ij}={a∈A/ a(s_i)≠a(s_j)} for i,j=1,…..n

The core is the set of all discernibility matrix entries containing only one feature. The reducts are the minimal feature subsets that have at least one common element with any nonempty entry in the discernibility matrix.

Research on reduct calculation is one of the fundamental investigations in rough set theory. There are two problems related to the notion of reduct, which have been intensively explored in rough set theory by many researchers. The first problem is related to searching for shortest reducts (i.e. reducts with minimal cardinality). The second problem is related to searching for all reducts. It has been shown that the first problem is NP-hard problem and the second is at least NP-hard. Moreover, the potential number of all reducts existing in a given information system, consisting with k attributes, is equal to . $N(K) = \begin{pmatrix} K \\ K/2 \end{pmatrix}$

These facts cause the high computational complexity of all reduct based rough set methods. In other words, selecting an optimal reduct from all subsets of features is not an easy work. Hence, various methods for finding the minimal set of attributes represent the data as the total set of attributes have been proposed. This explains the reason for which we search for the best reduct.

3. Feature selection Algorithm based on rough sets

```
Input:
    An information system S
    A set of attributes C over S
    A set of documents over S
Output:
    Best Reduct (Red)
Method:
Step 0:Randomly select a subset of documents with size m from the corpus
Step 1: Determine elementary sets
      For each document dj
            if document dj does not belong to any elementary set
                  Create new elementary set for it
            End
      End
Step 2: Construct the discernibility matrix
      For each elementary set$_i$
            For each elementary set$_j$
                  dij = features that discern elementary sets i and j
            End
      End
Step 3: Determine the core
      Core= attributes of cells ($d_{ij}$ ) with length equal to one
Step 4: Determine the reduct
      Let Red=Core
      For every entry in the discernibility matrix
      Count frequency of every attribute in the discernibility matrix
            Count frequency of every entry in the discernibility matrix
      End
      For every entry in the discernibility matrix
```

```
            Remove the elements in core
     End
   Reduce number of cells in discernibility matrix to be tested for each cell
            Remove cell if it contains all features of one of the previous cells
     End
   Sort the discernibility matrix according to the cardinality of every entry
     End
     For each entry $c_{ij}$ in the discernibility matrix to be tested
            if $c_{ij} \bigcap$ Red $= \Phi$
               Select attribute a with height frequency
                  Red=Red$\bigcup$ {a}
            End if
        End for
```

4. Clustering Overview

Clustering is a useful technique in data mining for discovering interesting data distributions and patterns in the underlying data. Clustering is the non-trivial process of identifying implicit, previously unknown but potentially useful groups that may exist in any data set. The grouping is done based on some similarity function [9]. The main advantage of using this technique is that interesting structures or clusters can be found directly from the data without using any background knowledge. Clustering algorithms can be broadly classified into two categories: partitional and hierarchical. One popular approach in document clustering is agglomerative hierarchical clustering. Algorithms in this family follow a similar template: compute the similarity between all pairs of clusters and then merge the most similar pair. This process is repeated until all objects are joined in one cluster. Results can be shown in a tree or dendrogram, where the horizontal bars connecting clusters or objects represent the dissimilarity between them (can be read from the vertical axis). By selecting a cutoff dissimilarity value, a list of clusters can be generated, which resembles the output of nonhierarchical clustering methods. Different agglomerative algorithms may employ different similarity measuring schemes. It has shown that UPGMA (Un-weighted Pair Group Method with Arithmetic Mean) is the most accurate one in its category [10].

In UPGMA the similarity of two clusters is calculated as the average of the pair-wise similarity of documents from each cluster

$$similarity(cluster1, cluster2) = \frac{\sum_{\substack{d_1 \in cluster1 \\ d_2 \in cluster2}} Cosine(d_1, d_2)}{size(cluster1) * size(cluster2)}$$

Where d_1 and d_2 are, documents, respectively, in cluster1 and cluster2.

5. Experimental Result

5.1 Datasets

We used two document data sets in our experiment. Table1 describes the details of each data set. The first one DS1 consists of 185 documents. The documents are classified into ten different categories according to their content. The feature set includes 46 terms. The other data set DS2 contains 12 documents whose names indicate their topics in three categories. The feature set includes six terms flow, form, layer, patient, result and treatment.

Data set	**Docs**	**Classes**	**Docs/ Classes (Average)**
DS1	185	10	18.5
DS2	12	3	4

Table (1) Data sets Descriptions

5.2 Evaluation measurement

A commonly used measure, the F-measure is employed to evaluate the accuracy of the proposed clustering solutions. It is a standard evaluation method for both flat and hierarchical clustering structures [10]. The clustering accuracy can be evaluated using F-measure by comparing the results to the pre-classified classes. The definition of F-measure is derived from the definition of precision and recall in information retrieval [11]. For a cluster j with respect to class i, the precision and recall are defined as follows [12]:

Precision (i, ,j)=$\frac{N_{ij}}{N_i}$

Recall (i,j)=$\frac{N_{ij}}{N_j}$

Where N_{ij} is the number of elements which are contained both in class i and cluster j; N_j is the number of elements in cluster j and N_i is the number of elements in class i. F-measure for a class i with respect to cluster j is defined as

$$F(i,j) = \frac{2 \times precision(i,j) \times \mathrm{Re}\,call(i,j)}{precision(i,j) + \mathrm{Re}\,call(i,j)}$$

The cluster that maximizes the F-measure for one class is considered to be the best cluster solution for it. The overall quality of a clustering result is measured by the weighted sum of such maximum F-measure for all reclassified classes:

$$F = \frac{\sum_i (num_i \times \max F(i,j))}{\sum_i num_i}$$

Where num_i is the number of elements in class i. The range of F is [0,1]. The better the clustering method is, the higher F-measure will be obtained because clusters produced are more similar with the standard classes.

5.3 Results and discussions

In our study, we search for the best reduct generated by using rough set theory in order to overcome the problems described in section 2. Our proposed method to try solves these problems. We wanted to check whether the chosen reduct generated by using rough set theory can lead to clustering result that are comparable to the clustering based on the complete data set. It is desired that the chosen reduct should still able to distinguish the classes from each other, since we want to get the reduct preserving to the highest degree the hierarchal classification of the known data, hoping that it will be able to classify the unknown data. The experiments come with some Phases:

Phase 1. Create the best reduct of the system using the proposed method.

Phase 2. Apply UPGMA algorithm on the data with reduced number of attributes and on the data with total number of attributes.

Phase 3. Comparing with the use of unreduced features.

Phase 1: Creating the best Reduct of the system

It is important to show that the use of features selected does not significantly reduce the classification accuracy as compared to the use of the full set of original features. The proposed algorithm based on Rough set theory was applied to the text documents in order to find the best reduct and core of the data. We have found that for the first dataset DS1 the best reduct of the full data set of 185 documents and 46 features contains 33 features. The core contains 19 elements. On the other hand the number of objects is reduced to 146 objects. For the other dataset DS2 the best reduct of the full data set of 12 documents and 6 features contains 5 features. The core contains 3 elements and the number of objects is reduced to 8 objects.

Phase 2: Apply data clustering using UPGMA algorithm

UPGMA algorithm is used to inductively cluster the documents. At the first time we will apply the algorithm on the data using the total number of attributes. After that we will apply the algorithm on the reduced number of attributes.

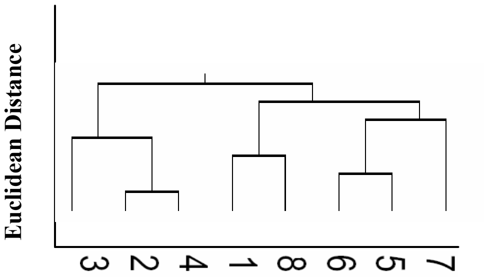

Figure1 Dendrogram for DS2 based on total data

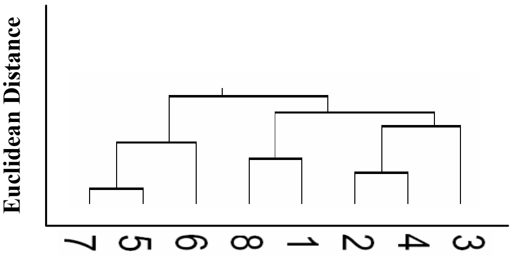

Figure 2 Dendrogram for DS2 based on reduced data

The dendrogram for the data set DS2 based on the complete data is shown in figure 1 and the dendrogram based on the reduct is in figure 2. High similarity between the two dendrograms can be seen, but the visual inspection is difficult, certainly for the larger dendrograms.

Phase 3: Comparing with the use of unreduced features.

Table 2 shows the F-measure results of using UPGMA algorithm on the two datasets DS1 and DS2.

Dataset	No. of clusters	Total data	Reduced data
DS1	2	0.18397	0.17251
	5	0.15697	0.16066
	18	0.14252	0.14489
	Avg	0.16115	0.15935
DS2	2	0.57828	0.82260
	5	0.76805	0.76805
	6	0.79583	0.79583
	Avg	0.71405	0.79549

Table 2 F-measure results

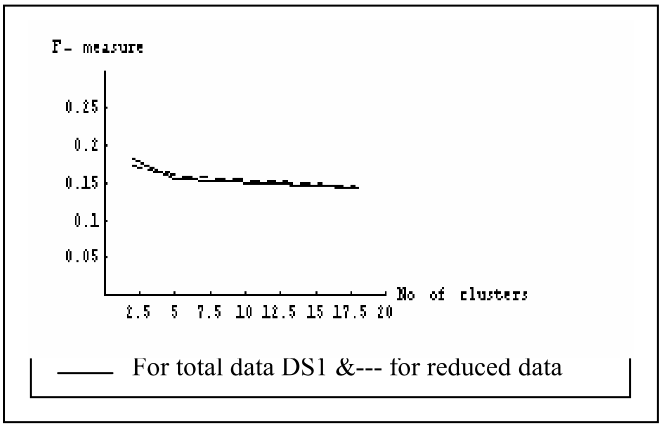

Figure 3 F-measure comparison for total and reduced data set DS1

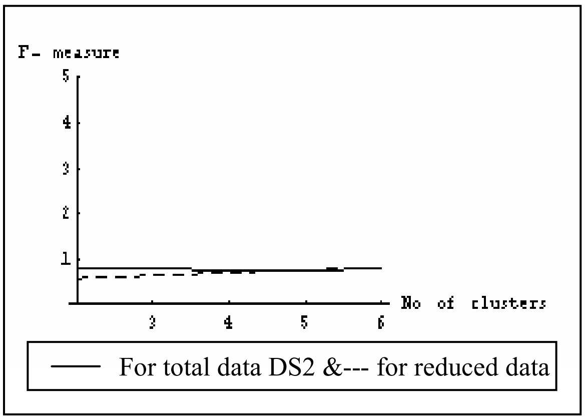

Figure 4 F-measure comparison for total and reduced data set DS2

Figure 3, and figure 4 illustrate the F-measure results variation along with number of clusters increasing from 1 to 18, for reduced data and for total data. By comparing the clustering solution of these two data sets, it is apparent to see that our proposed approach produces accurate clusters in most cases. This can be seen from the results tested on DS1 and DS2. From figure 3 we can conclude that the F-measure results for both the original data and the corresponding reduced data are nearly the same, not only this, but also it appears that the F-measure results for the reduced data is relatively exceed the F-measure results for the original data in most regions of clustering. In conclusion the clustering result based on reduced feature set and that one based on the original feature set are very similar.

6. Conclusion

This study demonstrates how rough set can be used for feature selection in unsupervised hierarchical clustering. In this paper we start with the definition of feature selection. Then, we have been presented a rough set method and its foundations. Rough set method has shown ability to reduce significantly the dimensionality. We provided a method based on rough set theory to find the best set of attributes of any given dataset. We tried to overcome the problems faced by using rough set theory when generating the total set of reducts in a large data set. A hierarchical clustering algorithm is applied on two different datasets. Finally, we present a comparison of the clustering results based on the original dataset and those based on the reduced dataset. The clustering result based on the original features set and that one based on the reduced features set are very similar which led us to know that our proposed method for generating a best reduct is an acceptable method.

7. Reference

[1] Bin Tang, Michael Shepherd, Malcolm I. Heywood, Xiao Luo, Comparing Dimension Reduction Techniques for Document Clustering, Canadian Conference on AI 2005, 2005, 292-296.

[2] Rajen B. Bhatt and M. Gopal, On the compact computational domain of fuzzy-rough sets, Pattern Recognition Letters, 26(11), 2005, 1632-1640.

[3] F. Questier, B. Walczak, D. L. Massart, C. Boucon and S. de Jong, Feature selection for hierarchical clustering, Analytica Chimica Acta, 466(2), 2002, 311-324.

[4] Shailendra Singh and Lipika Dey, A new customized document categorization scheme using rough membership, Applied Soft Computing, 5, (4), 2005, 373-390.

[5] Theresa Beaubouef, Frederick E. Petry and Roy Ladner, Spatial data methods and vague regions: A rough set approach, Applied Soft Computing, In Press, Corrected Proof, Available online 18 January 2006.

[6] Roman W. Swiniarski and Andrzej Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters, 24(6), 2003, 833-849.

[7] Malcolm Beynon, Reducts within the variable precision rough sets model: A further investigation, European Journal of Operational Research, 134(3), 2001, 592-605.

[8] Ju-Sheng Mi, Wei-Zhi Wu and Wen-Xiu Zhang, Approaches to knowledge reduction based on variable precision rough set model, Information Sciences, 159(3-4), 2004, 255-272.

[9] S. Asharaf, S. K. Shevade and M. Narasimha Murty, Rough support vector clustering, Pattern Recognition, volume 38(10), 2005, 1779-1783.

[10] Fung, B. C. M., Wang, K., and Ester, M., Hierarchical document clustering using frequent itemsets, Proc. of the 3rd SIAM International Conference on Data Mining (SDM 2003), San Francisco, CA, 2003, 59-70.

[11] K.M. Hummouda and M. S. Kamel, phrased based document similarity based on an index graph model, Proc. of 2002 IEEE International Conference on Data Mining, 2002, 203-210.

[12] Ling Zhuang, Honghua Dai, Maximal Frequent Itemset Approach for Web Document Clustering, Proc. of The 2004 International Conference on Computer and Information Technology (CIT'04) Wuhan, China, 2004.