Genetic algorithm for clustering MEDLINE abstracts

Dhekra BEN SASSI
Higher Institute of Management of Tunis
Member of RIADI Laboratory in the National School of Computer Sciences, Manouba
bensassi.dhekra@yahoo.fr

Wahiba BEN ABDESSALEM KARAA
Higher Institute of Management of Tunis,
Member of RIADI Laboratory in the National School of Computer Sciences,
Manouba
wahiba.time@yahoo.fr

ABSTRACT. Text mining is widely applied in the biomedical domain. This domain combines a variety of research area which makes the amount of available data enormous and uncountable. MEDLINE, the most known biomedical data base, suffers from a lack of text mining techniques and this makes information retrieval a more difficult challenge. In our research, we will focus on one of text analysis tasks which is clustering. Many works have focused on enhancing MEDLINE abstracts clustering but it stills a research area where researchers are trying to propose new approaches for a better clustering. In this article, we propose a new approach for clustering MEDLINE abstracts based on an evolutionary algorithm which is genetic algorithm.

KEYWORDS. MEDLINE, Clustering, Text mining, Genetic algorithm.

1. Introduction

The quantity of electronic documents is in permanent growth [FIS 09] which makes manual knowledge exploration and retrieval difficult or almost impossible. However, the use of computer’s technology strength helps people to extract knowledge from large amounts of semi-structured or unstructured text. This technology is called Text Mining [FEL 95]. Text mining is widely applied in the biomedical domain [ZHU 09]. This domain combines a variety of research area which makes the amount of available data enormous and uncountable. Hence, the need of automatic helps. Many researches are related to this domain which
cooperates in its improvement such as document clustering, document ranking and document indexing... MEDLINE, the most known biomedical data base, suffers from a lack of text mining techniques and this makes information retrieval a more difficult challenge. Several works were developed in this context but clustering MEDLINE abstracts is still a field where researchers are trying to propose new approaches for a better clustering solution.

In this paper, we propose a novel approach for clustering MEDLINE abstracts based on genetic algorithm. This paper is organized as follows; Section 2 defines our work’s area and mentions related works to this paper. In section 3 we present our proposed algorithm. An empirical study is conducted in section 4 to measure the effectiveness of our proposed approach and finely we will conclude in section 5.

2. Related works

Clustering is an unsupervised classification technique, which consists on partitioning a data set into a set of clusters without knowing the classes’ labels [VES 00]. Based on similarity concepts, this method tries to maximize the intra-cluster similarity and minimize the inter-cluster similarity [TAY 10, JAI 99, DUD 01].

2.1. Textual data clustering

The concept document clustering is so related to data clustering. It is a specific kind of clustering where data are not structured which makes it difficult. For these reasons the task of documents clustering is intensively studied in literature. Many algorithms which were applied in data clustering problem have been adapted to deal with textual data structure such as K-means [SEL 86] and the branch and bound [MAU 00].

As being an unsupervised classification, clustering can be considered as a particular kind of NP-hard grouping problem [FAL 98]. Evolutionary algorithms are considered as effective on resolving NP-hard problem because of their ability to found the best clustering solution in reasonable time [HRU 09]. Under this assumption, a large number of evolutionary algorithms have been adopted to solve clustering problems such as Genetic Algorithm [SAN 95], Self-Organizing Maps[RAG 79], Particle Swarm Optimization [MER 02] etc

2.2. Medline Clustering

MEDLINE is a largest biomedical literature database. It is daily updated with 200 to 4,000 citations. This permanent growth induces the need of a good MEDLINE abstracts clustering to accelerate the procedure of research and
information retrieval. In a general way, clustering biomedical documents has been effected by the “vector space model” [SAL 83], where each document is considered as a “bag of words” and represented by a weighted vector to facilitate the similarity computation [HUA 11]. Recently, the semantic information of MeSH thesaurus (Medical Subject Heading thesaurus) is being applied to clustering MEDLINE documents by mapping documents into MeSH concept vectors to be clustered. [ZHA 07, YOO 06]. However, those works had two limitations: First, automatically mapped MeSH concept vectors might not reflect the semantic information of the original document accurately. Second, important content information in the original text might be lost. To deal with these problems, Zhu et al. [ZHU 09] developed a strategy which includes three key points. First, they developed a sound method for measuring the semantic similarity between two documents over the MeSH thesaurus. Second, they combined both the semantic and content similarities to generate the integrated similarity matrix between documents. Third, they applied a spectral approach to clustering documents over the integrated similarity matrix. In the same year a text based search engine have been developed for retrieving documents from Medline and PubMed biomedical databases [MAN 09].

Each proposed approach has its limitation and its advantages. Until now, MEDLINE suffers from a lack of text mining technique and this makes information retrieval a more difficult challenge. Our aim is to found a better clustering solution.

3. The proposed algorithm

In the last few years, evolutionary algorithms have been widely applied to clustering problems. As an evolutionary algorithm, genetic algorithm is able to avoid local optimal solutions and converge to a global one. From this came our first idea to apply genetic algorithm on clustering problem while trying to find the best clustering solution. This idea was supported especially when, first comparing genetic algorithm clustering results with those obtained by hill-climbing search methods [FRA 00, GAR 04, MAR 06, MIT 04, KRI 99, KUN 98,], second doing the literature review concerning Medline abstracts’ clustering and finding that genetic algorithm has never been applied. Since genetic algorithm does not fit with the textual data structures, a data pre-processing is needed. To do it, we have applied the most straightforward and popular approach in representing textual data which is the vector space model (VSM).

We have integrated the ability of the vector space model to deal with textual data and the simplicity of agglomerative algorithms with the capability of genetic algorithm in avoiding local optima for developing the proposed approach. The following paragraphs will go into details of describing the key elements of our approach.
3. 1. **Individual encoding**

This step consists on the adaptation of our problem to the genetic algorithm such as each point of the state space is encoding in a chromosome. For a given document set \( D = \{D_1, D_2, \ldots, D_n\} \), each chromosome is a sequence of integer numbers representing the class label of \( n \) documents. This means that each chromosome is a possible clustering of the documents. As an illustration let us consider the following example: For given 4 documents \( D_1, D_2, D_3 \) and \( D_4 \); we assume that:

**EXAMPLE 1.**

– \( D_1 \) and \( D_2 \) belong to cluster 1, \( D_3 \) and \( D_4 \) belong to cluster 2

The encoding of this clustering is as follows:

<table>
<thead>
<tr>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

3.2. **Population initialization**

The choice of the initial population is important because it may make more or less rapid convergence towards the global optimum. In several works the initial populations are randomly generated, but in this case we are not sure that we have explored all the search space thus, we may miss a solution.

The best mechanism for generating the initial population consists on applying a clustering algorithm which allows the restriction of the search space while being sure that the subspace to be explored contains the optimal solution. When choosing the right algorithm, we must take into account that the number \( K \) of cluster is unknown. For these two reasons we have applied an agglomerative algorithm (ALGORITHM 2.–), one of the most widely used clustering strategies [LUK 79]

ALGORITHM 2.–

**Input:** Set of \( n \) documents

**Output:** a tree of clustering

- Create \( N \times N \) doc-doc similarity matrix
- Each document starts as a cluster of size one
- Do until there is only one cluster:
  - combine the two clusters with the greatest similarity
  - update the doc-doc matrix
At this stage, we notice that calculating the similarity requires a pre-processing of the document set to represent them in a structured manner and to subsequently apply mathematical formulas to compute the similarity between them. For this reason, we used the vector space model which is an algebraic model for representing text documents, where each document is considered as a “bag of words” and represented by a weighted vector \( d \) such as \( d_i = (w_{1,i}, w_{2,i}, \ldots, w_{t,i}) \) and the term specific weight \( w_{t,d} \) in the document vectors represents the weight of the term \( t \) in the document \( d \). The principle is more explained in the following figure.

![Figure N.1. Vector space model](image)

After modeling each document as a vector, we compute the similarity between documents to create the document similarity matrix of the agglomerative algorithm as shown in figure 2. The most known similarity measure applying in vector space model is cosine similarity, \([SAL\ 83]\). Given two documents’ vectors \( d_1 \) and \( d_2 \), this similarity is represented by this formula:

\[
\cos(\theta) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}
\]  

[1]
To generate the initial population, we will launch the agglomerative algorithm and we will consider the output of each iteration as being an individual as shown in the following example with a set of 4 documents.

We note that more the number of document is high more the initial population is diverse and wide.

### 3.3. Fitness Computation

The fitness score is the function to be optimized. In the literature, many measures are implemented to compute it. Maulik and Bandyopadhyay [MAU 00] used the Euclidean distances between the objects and their respective cluster centers. The Davies-Bouldin index has been utilized for computing the fitness of individual by Song et al [SON 09]. However, we cannot qualify one of them as the best.

In our case, since it is textual data clustering, we had recourse to the vector space model to represent each document as a vector containing the frequencies of terms.
For this reason we choose Cosine similarity, which is a measure of similarity between two vectors, to compute our fitness score. Mathematically, the representation of the fitness function is as follows:

\[
F = \sum_{i=1}^{K} F_i, \text{ such } K \text{ is the number of clusters } C_i \tag{2}
\]

and:

\[
F_i = \sum_{x_j, x_u \in C_i} \frac{x_j \times x_u}{||x_j|| \times ||x_u||}; \forall x_j, x_u \text{ documents' vectors } \in C_i \text{ and } j \neq u \tag{3}
\]

In other words, we will run through the \( K \) clusters \( C_i \) (1 \( \leq \) i \( \leq \) k) one by one, and for each of them, we will compute the Cosine similarity between each two documents \((x_j, x_u)\) of the same cluster, and the result will be the sum of the calculated Cosine similarities. The goal is to maximize the intra-cluster similarity hence the objective function \( F \).

3.4. Crossover

Crossover is a probabilistic process that exchanges information between two parent chromosomes in order to generate two child chromosomes. In our study, the one-point crossover operator is used, because of its simplicity, with a probability \( P_c=0.6 \). This type of cross consists on randomly draw a common position \( i \) in both parents, then exchange the two terminal substrings from each chromosome which produces two children.

3.5. Mutation

In order to maintain diversity through generations, a mutation operator is used. It consists on selecting a random gene in the chromosome and replacing it with a random value. We use the mutation operator with a probability \( P_m=0.1 \).

3.6. Selection

Genetic algorithms use methods of selecting individuals from the population. These individuals form an intermediate generation (mating pool) to reproduce and eventually mutate, thus forming a new generation. Since selected individuals pass their genes to the next generation it is necessary to select good individuals. Selection methods are divided into two classes: selection by fitness score and selection by rank. In our approach the Roulette wheel selection is applied. It belongs to the first class of selection methods. Its principle consists on associating with each individual \( \theta_i \) (1 \( \leq \) i \( \leq \) n) a probability \( P_i \) proportional to its fitness \( f(\theta_i) \) in the population. \( P_i \) is computing as follow:
\[ P_i = \frac{f(\theta_i)}{\sum_{i=1}^{n} f(\theta_i)} \] Where \( n \) is the population’s size

3.7. Termination criterion

Since the objective is to maximize the intra-cluster similarity, we impose that the algorithm is terminated when the number of consecutive iterations without improvement reaches \( N_{\text{max}} \) which will be experimentally fixed. The best chromosome which has the highest fitness value is returned by the last iteration.

The pseudo code of the proposed approach is summarized in (ALGORITHM 1.). As explained before, we will start by modeling the abstracts into vectors with VSM, second we will compute the similarity between them using cosine similarity, and then we will obtain a population of candidate clustering solutions by executing the agglomerative algorithm. An objective and fitness function is associated with each chromosome that represents the degree of fitness. The best individuals will be selected to reproduce through the crossover operator and eventually mutate through mutation operator, thus forming a new generation. The fitness function of the new clustering solution will be calculated then compared with the fitness of the worst clustering solution in the population. If it is lower than that of the worst clustering solution in the population, then the worst clustering solution is replaced by the new clustering solution. Selection, crossover and mutation operators iterate until the termination criterion is satisfied. The best clustering solution will be returned at the end.
4. Experimental result

In this section, we will focus on the technical details related to the implementation of our algorithm, then we will present the obtained results and finely an evaluation of the developed system will be carried out.

Before the implementation of the genetic clustering algorithm, a pre-processing step is needed. It is a crucial and common step for all text mining process and this is due to the nature of textual data which are unstructured and mostly available in natural language form. In our case the pre-processing consists on the selecting and the cleaning steps. The selection step consists of the extraction of 100 abstracts from MEDLINE data base. These abstracts must be of different themes, hence belong to different classes. To do it, we make use of JabRef, a program that lets user fetch in

**Algorithm 1.**

<table>
<thead>
<tr>
<th>Input: a set of n documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output: final clustering $S^*$</td>
</tr>
</tbody>
</table>

1. Vector space model
2. Cosine similarity
3. Population initialization $P_i$
   - Agglomerative algorithm
4. Compute fitness score of all individual in the population $F(P_i)$

While stop condition do
5. Parents = Selection operator ($P_i, F$)
6. Child1, Child2 = crossover (parent1, parent2)
7. Child3 = mutation (parent)
8. For i = 1; i < children’s number
   - Compute fitness score (Child$_i$)
9. if Child chromosome fitness score > the worst chromosome fitness score
   - Swap Child1, Child2, Child3 with the worst chromosomes
end if
end while
return best chromosome
MEDLINE and other database in order to build his own biomedical abstracts database. The selected abstracts must be cleaned because they contain much useless information that affect the clustering performance. The cleaning step consists firstly on defining the English stop-words list, then scanning all documents and eliminating these collected words from the abstracts if ever they exist. The collected abstracts are ready now to be the input of our clustering algorithm. Processing results are presented in the next section.

As mentioned before, to evaluate our system we used a corpus of 100 MEDLINE abstracts selected from 10 different themes such as each 10 abstracts belong to one different class i.e. at the beginning, the abstracts were selected according to a known class label and thus to detect errors through a comparison between the obtained clusters and the initial ones. After selection, the classes labels are hidden and the abstracts are represented by numbers and thus to ensure the unsupervised aspect in the clustering process. Documents are ready now to be clustered by the proposed genetic algorithm. The population undergoes the selection operation according to the fitness score. The selected individuals will make either a crossover or a mutation according to the probability of each operator. These probabilities are handled according to the following principle: We associate to each probability a segment of length proportional to its value (0.1+0.6= 0.7). Then we draw a random number from this interval and we select the corresponding segment. To ensure the effectiveness of our program, we conducted a set of tests:

- To perform the first test, we used 10 abstracts which have no common term such as one from each theme and after 20 iterations the program returns the best clustering which is a single clustering with a null value of fitness. The execution time = 0.3s.

- Then we extend the basis and we retest with 50 abstracts where our algorithm performs well in an execution time = 1.33 min.

- Finally, we test our algorithm performs on the whole abstracts’ sample. The error rate is 3%.

To evaluate the effectiveness of our system, we choose to calculate precision, recall and f-measurement to the whole of the clustering solution, but also to each cluster separately, to detect mistakes made. Results are summarized in Table 1:

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Disease</td>
<td>0.83</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td>Genetic</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Anatomy</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Physiology</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Psychology</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pharmacology</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dentistry</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
The previous table shows that the “Disease” cluster and the “Nursing” cluster contain errors. This is because it contain several terms in common.

5. Conclusion

In this paper, we dealt with the concept of textual data clustering and more precisely MEDLINE abstract clustering. In summary, this work proposes a new clustering technique based on genetic algorithm. The proposed algorithm uses vector space model to deal with textual data structure and an agglomerative algorithm for initial population generation. In the experimental study, we used a sample of 100 abstracts and the effectiveness measures return were of good value. For the proposed algorithm, after the documents’ modeling with the vector space model, we considered each term as a numerical weighed data i.e. we do not treat the semantic aspect of textual data. It will be of interest to extend the proposed algorithm and take into account the semantic between terms. This step will certainly improve the information retrieval exercise.

References

[FEL 95] FELDMAN R., DAGAN I., “Knowledge discovery in textual databases (KDT)”, In proceedings of the First International Conference on Knowledge Discovery and Data Mining KDD ’95, Montreal, 20-21 August 1995, Canada, AAAI Press, p.112-117.


“Hierarchical Agglomerative Clustering Procedure”, *Pattern Recognition*, vol. 11, 1979, p. 365-381.


Genetic algorithm for clustering MEDLINE abstracts


