INFORMATION
SYSTEMS IN
MANAGEMENT

Information Systems in Management (2018) Vol. 7 (3) 213–224
DOI: 10.22630/ISIM.2018.7.3.6

CLUSTER ANALYSIS OF MEDICAL TEXT DOCUMENTS BY USING SEMI-CLUSTERING APPROACH BASED ON GRAPH REPRESENTATION

RAFAŁ WOŹNIAK, PIOTR OŻDŻYŃSKI, DANUTA ZAKRZEWSKA

Institute of Information Technology, Lodz University of Technology

The development of Internet resulted in an increasing number of online text repositories. In many cases, documents are assigned to more than one class and automatic multi-label classification needs to be used. When the number of labels exceeds the number of the documents, effective label space dimension reduction may significantly improve classification accuracy, what is a major priority in the medical field. In the paper, we propose document clustering for label selection. We use semi-clustering method, by considering graph representation, where documents are represented by vertices and edge weights are calculated according to their mutual similarity. Assigning documents to semi-clusters helps in reducing number of labels, further used in multi-label classification process. The performance of the method is examined by experiments conducted on real medical datasets.

Keywords: cluster analysis, semi-clustering, text mining

1. Introduction

Nowadays, especially in the area of medicine, there is a big need of automatic classification of text documents contained in large repositories. The documents are very often assigned to more than one class and then the application of multi-label classification is necessary. However, in many cases occurrence of big number of labels makes it difficult to obtain the required accuracy of multi-label classification task.

In the paper, we investigate the method of reducing number of labels, which can be used in the pre-processing step of multi-label classification task. We propose to consider the documents as a social network, where the social graph depicts relationships between documents represented by vertices, with edges indicating mutual similarities. The documents are assigned to semi-clusters, and groups of labels which occur together the most often are pointed out. The qualitative analysis of the experiments conducted on real medical text document datasets showed a potential of the proposed method in indicating the labels that mostly occur together.

The remainder of the paper is organized as follows. Relevant work concerning label space dimension reduction approaches is presented in the next section. Then the methodology is described, including the description of all its steps. In the following section the experiments conducted on two datasets are depicted and the results of qualitative analysis is presented. Finally, concluding remarks and future research are shortly described.

2. Related work

There exist different approaches to handle the problem of big amount of labels in multi-label classification tasks. From among them one should mention label subset selection, when the reduced number of labels is used in the classification process. Tsoumakas et al. [1] considered reducing label sets by using hierarchical algorithm for multi-label classification. Balasubramanian and Lebanon [2] proposed the method which is based on the assumption that for multidimensional variables there exists a small subset of dimensions, such that all the remaining ones may be expressed by their noisy linear combination. Read et al., in turn, eliminated rare label sets and thus reduced the number of labels [3]. Bi and Kwok [4] proposed randomized sampling with the probability of class labels reflecting their importance.

Hsu et al. [5] investigated Label Space Dimension Reduction (LSDR) approach. They used compressed sensing technique, taking into account sparsity of the label space and projecting it into a compressed space of lower dimensionality. LSDR techniques have been modified and investigated by many researchers. Lin et al. [6] proposed using feature-aware implicit label space encoding instead of explicit encoding function. They showed that their approach to LSDR gives superior classification performance. Another feature-aware approach to LSDR has been proposed by Chen and Lin [7]. They based their conditional label space transformation on minimizing an upper bound of Hamming Loss evaluation measure.

The broad review of the label space reduction has been presented in [8]. The research mainly concerned general cases not regarding data types.

3. Methodology

The considered methodology deals with reducing the number of labels in the tasks of multi-label classification of text documents. We use graph document representation to build groups that are not required to be separable and find out which labels occur together the most often.

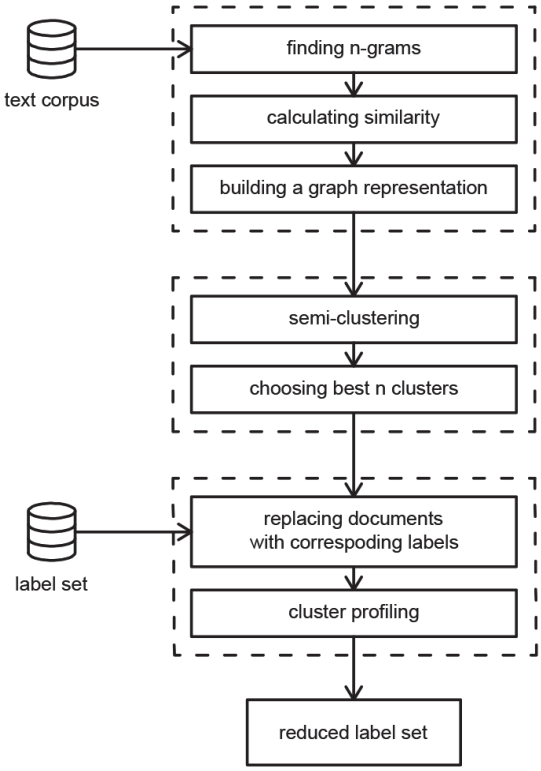

Figure 1. Scheme for the proposed algorithm

3.1. The method overview

Considering the document corpus, the proposed method consists of the three major steps:

- **Building graph representation**. For each pair of documents in the corpus, the similarity is calculated. Its value depends on the number of the same sequences in both documents. As the result, we obtain an undirected graph in which vertices represent documents and edge weights are equal to the similarity degree.

- **Semi-clustering.** The semi-clustering algorithm is performed on the graph. For each vertex, a list of semi-clusters is stored and the best one is chosen.
- **Label selection**. The content of clusters is transformed from documents to the corresponding labels. Finally, cluster profiles are created, what allows to select the best subset of labels.

The overview of the method is presented in Fig. 1, while the details of the steps are described in the following subsections.

3.2. Graph representation

For the semi-clustering purpose, documents are treated as social network, thus the graph reflecting their similarity can be constructed. In the considered solution, vertices stand for the text documents and edges denote *n*-grams which appear in the both connected documents. Since utilizing edge weights can provide better results in social network searching [9], we determine their values as the similarity degree of the two documents. The higher degree means higher similarity of the documents.

In the first step of constructing a graph representation, the list of all *n*-grams appearing in the text corpus, is built up. However, unigrams are excluded, as the minimum value of *n* is set to 2. For each sequence, a list of the documents in which it occurs is constructed and respective edges are created. Their weights $w(d_i,d_j)$ are calculated according to equation (1):

$$w(d_i,d_j) = \frac{(n-0{,}5)^2}{\max\{|d_i|,|d_j|\}} \tag{1}$$

where n is the size of the sequence and d_i, d_j represent the pair of documents, $i \neq j$. At the end, the edge score is normalized by the size of the bigger document.

Since the same two documents can have multiple edges, one for each *n*-gram, the values should be aggregated. The aggregation procedure is presented in Fig. 2. Even though an *n*-gram can be a part of another *n*-gram, only the longest one is taken into account during the calculations. On the contrary to social graphs [10], the proposed graph representation does not allow presence of loops, therefore process of their elimination is omitted.

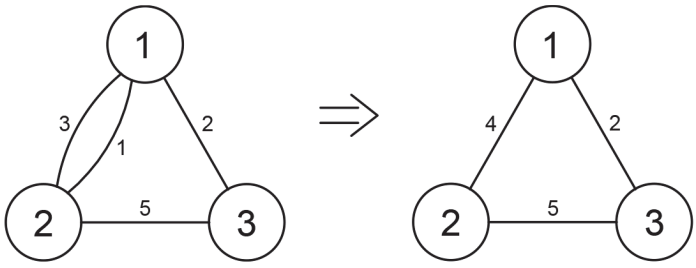

Figure 2. Aggregation of the multiple edges

3.3. Semi-clustering

Similarly to the social network of people, where a semi-cluster is defined as a group of people who have strong relations with each other and weak ones with people from outside, we will consider groups of similar documents. Namely, in the proposed method, a semi-cluster is a group of documents of the biggest weight values that was defined by (1). The major difference between clustering and semi-clustering is that vertices can be associated with more than one semi-cluster. Such approach is illustrated in Fig. 3.

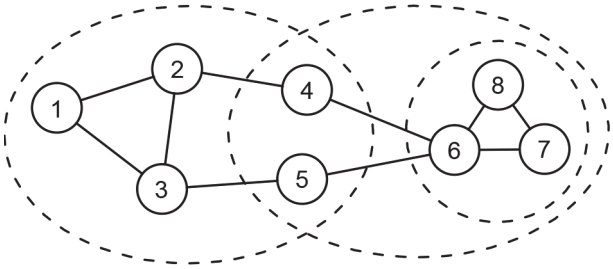

Figure 3. Result of the semi-clustering algorithm

In our solution, the vertex-centric iterative model from Pregel [11] is adopted. Each iteration is called a super-step and works on the results of the previous phase. The input of the algorithm is a weighted, undirected graph that is created in the first step of the proposed method and the output is a list of maximum $C_{\max}$ semi-clusters containing maximum $V_{\max}$ vertices, generated for each document in the graph. Both $C_{\max}$ and $V_{\max}$ are user-defined parameters, while each semi-cluster score S_c should be calculated according to the following formula (2):

$$S_c = \frac{I_c - f_B B_c}{V_c(V_c - 1)/2} \tag{2}$$

where I_c is the sum of internal edge weights, f_B is the user-specified boundary edge score factor, B_c is the sum of boundary edge weights and V_c denotes the number of vertices in the semi-cluster. The value S_c is normalized with the number of edges in a clique of size V_c, so that larger semi-clusters are not preferred.

In the initial super-step, each vertex adds itself to an empty semi-cluster, so its size increases to 1. Afterwards, vertices send themselves to their neighbors which check if they are already included in the semi-clusters. Otherwise, a new vertex is added and a semi-cluster score is calculated. If any of the existing semi-clusters has lower score, a new one is added to the list or replaces the worst one. Those lists are sorted by semi-cluster scores and sent to another neighbors. The algorithm stops if there is no improvement or when the user-defined number of iterations is reached.

It should be stressed out that vertex will not be added to the semi-cluster if the number of vertices equals V_{max}. The same goes with semi-clusters and C_{max}.

3.4. Label selection

The result of the previous step is a collection of semi-clusters, each containing a group of vertices representing text documents. Since it frequently happens that its size is too large, it is highly recommended to choose only the best semi-clusters, as their scores are known.

When the semi-clustering part is finished, there is no more strict interest in the documents, as we concentrate on the labels corresponding to them. For each chosen semi-cluster, text documents are replaced with their assigned labels. The operation of label grouping is presented in Fig. 4.

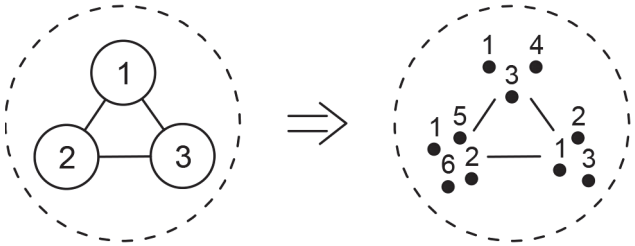

Figure 4. Grouping of the labels

Each document is given one or more labels so the size of clusters increases or hardly ever stays the same. The final phase is to perform the cluster profiling which selects the most relevant labels.

3.5. Datasets

In order to evaluate the proposed method, there will be considered two subsets of a real text dataset of medical abstracts. The OHSUMED corpus [12] includes the first 20,000 documents from 50,216 medical abstracts of the year 1991. The unique task was to categorize them into 23 categories of cardiovascular diseases. After that subset selection process, the number of documents was reduced to 13,929.

To ensure qualitative analysis of the proposed method small subsets of medical abstracts have been chosen. The first one includes 10 abstracts that are assigned to the category C01 (*bacterial infections and mycoses*). The second one contains 10 abstracts which are given category C04 (*neoplasms*).

4. Experiment results

The performance of the proposed method has been evaluated by experiments conducted on the datasets described in Section 3.5. For building a graph representation, there was implemented the system in Java programming language. The graph processing was done by using the open source Okapi library [13].

The input of the semi-clustering algorithm should consist of a weighted and undirected graph. Since, in Okapi library, the direction of the edges should be specified, each edge has been considered twice, once in each direction.

For the both of the datasets, user-defined parameters were experimentally chosen and have the following values:

- the maximum number of iterations k is 10 (default),
- the maximum number of clusters to shape C_{max} is 100,
- the maximum number of vertices a cluster can have V_{max} is 10,
- the boundary edge score factor f_B is 0.5 (default).

4.1. Documents of category C01

In the first step, pairs of text documents were identified. The minimum value of n was set to 2 according to Section 3.2. The pairs of documents with the same n-grams and their edge weight values are presented in Table 1.

Table 1. Pairs of documents with the same n-grams in the first subset

Documents		n-grams	Edge weight
0	8	haemophilu influenza type b	0.070809
2	9	human immunodefici viru hiv infect	0.056667
4	6	gram neg	0.022959
1	9	compar control	0.015000
2	7	small intestin	0.014151
1	8	significantli lower	0.013006

As it can be noticed, there are 6 edges between 8 documents and the strongest connections occur if there is more than one n-gram or the size of the sequence is bigger than 2. The length of the documents and the graph properties are also taken into account. The respective graph representation is presented in Fig. 5.

Two of the documents (3 and 5) that are not similar to others are omitted when it comes to the next step. Then, for each text document the semi-cluster with the highest score was suggested. Afterwards, the corresponding labels were taken into account. Since their minimum number in the semi-cluster is equal to 5, only those labels that occurred more than 20 percent times were chosen. The detailed semi-clustering results are presented in Table 2. The second column contains semi-

clusters for documents identified in the first one. The third column comprises labels connected with documents contained in the respective semi-cluster. Finally, in the last column the profiles of the chosen labels are indicated.

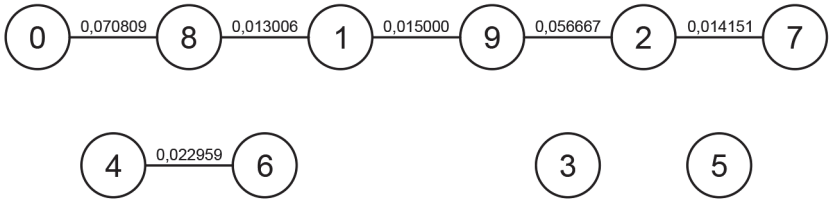

Figure 5. The graph representation of the first subset

Table 2. Results of the semi-clustering algorithm for the first subset

Document	Semi-cluster	Labels	Profile
0	0 8	C01 C10 C01 C10 C18 C23	C01 C10
1	0 1 8	C01 C10 C01 C06 C23 C01 C10 C18 C23	C01 C10 C23
2	2 9	C01 C02 C06 C20 C01 C15 C20	C01 C20
4	4 6	C01 C05 C17 C19 C01	C01
6	4 6	C01 C05 C17 C19 C01	C01
7	2 7 9	C01 C02 C06 C20 C01 C23 C01 C15 C20	C01 C20
8	0 8	C01 C10 C01 C10 C18 C23	C01 C10
9	2 9	C01 C02 C06 C20 C01 C15 C20	C01 C20

While the size of the best semi-clusters found by the system equals to 2 or 3, the number of the labels assigned to them varies from 5 to 9. As the result, 4 exclusive cluster profiles were recognized. The two of them are represented by 2 labels, the other one by 3, and the last one by the single label.

Summing up, in the examined subset, there are 11 unique labels assigned to 8 documents that have at least one edge in the graph. Qualitative analysis showed that the process of label selection allows to ignore some of them and choose the most relevant ones. The mutual relations between *bacterial infections and mycoses* (C01), *nervous system diseases* (C10), *immunologic diseases* (C20) and *pathologi-*

cal conditions, signs and symptoms (C23) were revealed by the proposed approach. There were also 7 labels reduced, such as *virus diseases* (C02), *musculoskeletal diseases* (C05), *digestive system diseases* (C06), *hemic and lymphatic diseases* (C15), *skin and connective tissue diseases* (C17), *nutritional and metabolic diseases* (C18) and *endocrine diseases* (C19).

4.2. Documents of category C04

Similarly to Section 4.1, all the pairs of medical abstracts with the same *n*-grams are identified for the second subset. Their data are respectively contained in Table 3. Also in this case, the minimum value of *n* was set to 2.

Table 3. Pairs of documents with the same *n*-grams in the second subset

Documents		*n*-grams	Edge weight
16	17	patient year ag	0.044014
10	11	malign transform t cell	0.025862
13	14	hospit patient	0.023438
13	16	patient ag	0.023438
13	19	gastric cancer	0.021845
16	18	ag year	0.017717
17	18	review patient	0.015845
12	15	monoclon antibodi	0.014423

There are 8 edges between 10 abstracts. Once again, the strongest connections occur if the size of the sequence is bigger than 2 or if more than one *n*-gram can be found. As it can be noticed in the Fig. 6, the graph structure is slightly different.

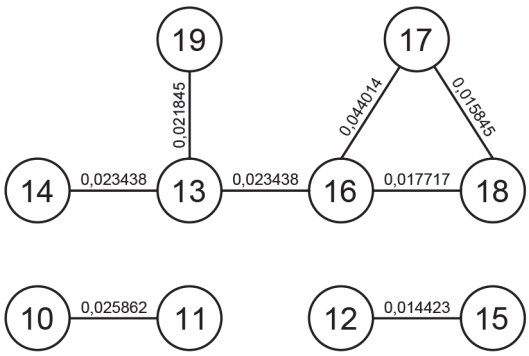

Figure 6. The graph representation of the second subset

For each text document the semi-cluster with the highest score was suggested and the corresponding labels were taken into account. Similarly to the case of the

previous dataset, we selected those labels that account for more than 20 percent. All the results of the algorithm are presented in Table 4.

Table 4. Results of the semi-clustering algorithm for the second subset

Document	Semi-cluster	Labels	Profile
10	10 11	C04 C20 C23 C04 C16 C23	C04 C23
11	10 11	C04 C20 C23 C04 C16 C23	C04 C23
12	12 15	C04 C04 C12	C04 C12
13	13 16 17 18	C04 C23 C04 C12 C04 C12 C04	C04 C12
14	13 14 16 17 18	C04 C23 C04 C12 C23 C04 C12 C04 C12 C04	C04 C12
15	12 15	C04 C04 C12	C04 C12
16	16 17 18	C04 C12 C04 C12 C04	C04 C12
17	16 17 18	C04 C12 C04 C12 C04	C04 C12
18	16 17 18	C04 C12 C04 C12 C04	C04 C12
19	13 14 19	C04 C23 C04 C12 C23 C04	C04 C23

While the size of the best found semi-clusters varies from 2 to 5, the number of the labels assigned to them is ranged from 5 to 10. As the result, 2 unique cluster profiles were recognized and both of them are represented by 2 labels.

Qualitative analysis for the second subset showed that there are 5 unique labels assigned to the 10 documents. The mutual relations between *neoplasms* (C04), *urologic and male genital diseases* (C12) and *pathological conditions, signs and symptoms* (C23) were revealed by the proposed method. It should be stressed out that prostate cancer was the most popular among men in the year of 1991 [14].

Two labels were identified as eliminated: *neonatal diseases and abnormalities* (C16) and *immunologic diseases* (C20).

5. Concluding remarks

In the paper the method of reducing number of labels assigned to text documents has been proposed. The presented technique consisted in considering social network of documents and using graph document representations. Such approach allows to build semi-clusters of documents and find out which groups of labels occur together the most often. The research is focused on medical documents, where the number of assigned categories is big, what makes difficult to obtain the required performance of multi-label classification task. The qualitative analysis of the two subsets of real medical text documents datasets showed the good potential of the proposed method.

Future research will consist in incorporating the proposed method as the preprocessing step of multi-label classification technique that aims at reducing the number of considered labels. Investigations will concern effectiveness in obtaining the results of the required performance.

REFERENCES

[1] Tsoumakas G., Katakis I., Vlahavas I. (2008) *Effective and Efficient Multilabel Classification in Domains with Large Number of Labels*, Proceedings of ECML/PKDD Workshop on Mining Multidimensional Data, MMD'08, 30-44.

[2] Balasubramanian K., Lebanon G. (2012) *The Landmark Selection Method for Multiple Output Prediction*, Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, 983-990.

[3] Read J., Pfahringer B., Holmes G. (2008) *Multi-label Classification Using Ensembles of Pruned Sets*, Proceedings of 8th IEEE International Conference on Data Mining, 995-1000.

[4] Bi W., Kwok J. (2013) *Efficient Multi-label Classification with Many Labels*, Proceedings of the 30th International Conference on International Conference on Machine Learning 28, Atlanta, Georgia, USA, III-405-III-413.

[5] Hsu D,. Kakade S.M., Langford J., Zhang T. (2009) *Multi-label Prediction via Compressed Sensing*, Bengio Y., Schuurmans D., Lafferty J.D., Williams C.K.I., Culotta A. [eds]: Advances in Neural Information Processing Systems 22, Curran Associates Inc., 772-780.

[6] Lin Z., Ding G., Hu M., Wang J. (2014) *Multi-label Classification via Feature-aware Implicit Label Space Encoding*, Proceedings of the 31st International Conference on International Conference on Machine Learning 32, Beijing, China, II-325-II-333.

[7] Chen Y.-N., Lin H.-T. (2012) *Feature-aware Label Space Dimension Reduction for Multi-label Classification*, Proceedings of the 25th International Conference on Neural Information Processing Systems 1, Nevada, USA, 1529-1537.

[8] Herrera F., Charte F. Rivera A. J., del Jesus M.J. (2016) *Multilabel Classification. Problem Analysis, Metrics and Techniques*, Springer Switzerland.

[9] Hangal S., MacLean D., Lam M.S., Heer J. (2010) *All Friends are Not Equal: Using Weights in Social Graphs to Improve Search*, Proceedings of the 4th ACM Workshop on Social Network Mining and Analysis, Washington, USA, 1-7.

[10] Andersen J.S., Zukunft O. (2016) *Semi–Clustering that Scales: An Empirical Evaluation of GraphX*, Proceedings of the 2016 IEEE International Congress on Big Data, San Francisco, USA, 333-336.

[11] Malewicz G., Austern M.H., Bik A.J.C., Dehnert J.C., Horn I., Leiser N., Czajkowski G. (2010) *Pregel: A System for Large-Scale Graph Processing*, Proceedings of the 2010 International Conference on Management of Data, New York, USA, 135-146.

[12] http://disi.unitn.it/moschitti/corpora.htm (accessed November 20, 2017)

[13] http://grafos.ml/okapi.html (accessed November 20, 2017)

[14] Boring C.C., Squires T.S., Tong T. (1991) *Cancer statistics, 1991*, CA: A Cancer Journal for Clinicians, 41(6), 19-36.