Comparative Study of Feature Selection Methods for Medical Full Text Classification

Carlos Adriano Gonçalves[1,3], Eva Lorenzo Iglesias[1], Lourdes Borrajo[1], Rui Camacho[2,3(✉)], Adrián Seara Vieira[1], and Célia Talma Gonçalves[4,5]

[1] Computer Science Department, University of Vigo, Escola Superior de Enxeñería Informática, Ourense, Spain

[2] Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
rcamacho@fe.up.pt

[3] LIAAD - INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

[4] CEOS.PP/ISCAP-P.PORTO, Rua Jaime Lopes Amorim s/n, 4465-004 Porto, Portugal

[5] LIACC, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

Abstract. There is a lot of work in text categorization using only the title and abstract of the papers. However, in a full paper there is a much larger amount of information that could be used to improve the text classification performance. The potential benefits of using full texts come with an additional problem: the increased size of the data sets.

To overcome the increased the size of full text data sets we performed an assessment study on the use of feature selection methods for full text classification. We have compared two existing feature selection methods (Information Gain and Correlation) and a novel method called k-Best-Discriminative-Terms. The assessment was conducted using the Ohsumed corpora. We have made two sets of experiments: using title and abstract only; and full text.

The results achieved by the novel method show that the novel method does not perform well in small amounts of text like title and abstract but performs much better for the full text data sets and requires a much smaller number of attributes.

Keywords: Text classification · Feature selection · Medical texts corpus

1 Introduction

The increasing and overwhelming amount of scientific research documents available in bio-medicine scientific databases such as MEDLINE, requires the development of tools to help researchers to keep up with all the relevant work being done

I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11466, pp. 550–560, 2019.
https://doi.org/10.1007/978-3-030-17935-9_49

all over the world. Moreover the number of corpora of full scientific texts publicly available is also increasing rapidly. The availability of full texts may increase the chances for better analyses but also requires processing larger amounts of data. We argue that new and more powerful tools are required.

Our approach to the increase in the amount of information to process is a novel feature selection algorithm that achieves better results than existing competitors with a much smaller number of attributes. We have empirically assessed the performance of several feature selection algorithms conducting a series of experiments using: Information Gain, Correlation and the newly developed feature selection algorithm.

Text classifiers can adequately be used to extract medical/biological information from very large scientific papers repositories as shown in [20]. Techniques, like the one presented in this paper, can contribute to build better text classifiers and therefore extract better papers from those large repositories.

The rest of the paper is organized as follows. Section 2 makes an introduction to the feature selection methods focusing on the ones used in the present study. Section 3 presents the new algorithm for feature selection, k-Best-Discriminative-Terms. Section 4 will highlight the feature selection work applied to biomedical full-text document for classification. In Sect. 5 we present the results regarding our study and finally in Sect. 6 we draw the conclusions of the work described in the paper.

2 Feature Selection Methods for Text Classification

According to [2] the feature selection process or attribute reduction is the process of selecting a subset of features that best represents by itself all the data. The rational of feature selection in the context of text classification is to represent a document with a reduced number of highly representative/discriminative attributes.

The full-text document classification, specially in the biomedical domain, involves the manipulation of very large data sets. This brings several well-known problems such as the increase of the computation time. Besides that, not all the attributes are relevant and important for the classification task, which is another well-known problem that disturbs the performance of the classifiers.

We have adopted the bag-of-word approach, original documents are seen as a vector containing a huge number of words. Since we are working with a large collection of documents, the number of words increases quite dramatically, which entails memory and time restrictions to run learning algorithms. Due to the exposed situation it is seriously important to select the most important and relevant attributes for the classification process. That is the objective of Feature Selection algorithms.

According to [11] there are two main reasons for selecting some features over others. The first reason is related to the algorithm's performance, e.g., algorithms produce better results when not considering all the attributes. This is due to some attributes do not add more information instead they add noise,

and removing them makes the classifier to perform better. The second reason is due to scalability, once a huge number of attributes demands computation power, memory, network bandwidth, storage, etc.) thus running a smaller subset decreases the computation time.

We have assessed the performance of three different feature selection methods:

1. Information Gain (IG)
2. Correlation (Corr)
3. *k*-Best-Discriminative-Terms (k-BDT)

The first two methods are now described and the k-BDT is presented in Sect. 3.

2.1 Information Gain

Information Gain (IG) is used to determine which attribute in a given set of training feature vectors is most useful for discriminating the class values to be learned [4,5].

IG is a "synonym" for *KullbackLeibler* divergence [14] and it is often used for ranking individual features [15].

In document classification, IG measures the number of bits of information gained, with respect to deciding the class to which a document belongs, by using each word frequency of occurrence in the document. However, IG only evaluates features in an individual manner.

IG is a feature selection method used prior and independent from the learning process, e.g., a filter method compares the computation score of each attribute and then selects the best attributes according to the highest scores [6].

Based on their comparative study of filter methods, [7] and [11] concluded that IG and Chi-Square (CHI) are among the most effective methods of feature selection for classification.

2.2 Correlation

According to the correlation algorithm an attribute is very relevant if it is highly correlated with the class, otherwise it is irrelevant [12].

We have used the WE*k*A *CorrelationAttributeEval* functionality, that evaluates the worth of an attribute by measuring its correlation (Pearson's Correlation) with the class. The WE*k*A *CorrelationAttributeEval* technique used requires a Ranker Search Method, that evaluates each attribute and lists the results in a ranked order.

3 k-Best-Discriminative-Terms

The rational of the k-BDT method is to find the best k terms[1] in the corpus that best discriminate the two classes of documents (assuming a binary classification

[1] We have used single words in our study but the k-BDT can also be used with other groupings of words like n-grams (n > 1), NERs, etc.

problem). In an informal description the documents are first separated by class value and, for each class value, the metric Tf×Df is computed for each term. This metric represents the average term frequency in the class value multiplied by the document frequency. The justification for Df is that we aim at terms that are frequent in all documents of each class value but infrequent in the "other class value". The k-BDT method is described in detail in Algorithm 1. The documents of the two class values are separate in lines 2 and 3. The documents from one of the class value (let say POS) are processed between lines 5 and 11. First the term frequency (Tf - line 7) and document frequency (Df - line 8) are computed for each term and document and then the Tf×Df is computed (line 9). Finally we compute the average Tf×Df for each term in lines 10–11. We repeat the same procedure for the other class value (NEGS) (lines 12–18). The "final values" are the difference between the Tf×Df of POS and the corresponding Tf×Df of the NEGS (lines 19–21). The final values are sorted by descending order (line 22) and the k first terms are returned in line 23. In Algorithm 1 (line 21) we have used the *abs* function but, as described in the next paragraph, we have also considered an alternative procedure.

Algorithm 1. k-Best-Discriminative-Terms algorithm

```
1: procedure K-BDTPROCEDURE(Corpus, k)
2:     Pos ← relevantTexts(Corpus)
3:     Negs ← irrelevantTexts(Corpus)
4:
5:     for doc in POS do
6:         for term in doc do
7:             Tf_{term,doc} ← termFrequency(term, doc)
8:             Df_{term} ← docFrequency(term, POS)
9:             Ti_{value} = Tf_{term,doc} × Df_{term}
10:    for term in allTerms do
11:        T_{posval} ← average(Ti_{value}, POSdocs)
12:    for doc in NEGS do
13:        for term in doc do
14:            Tf_{term,doc} ← termFrequency(term, doc)
15:            Df_{term} ← docFrequency(term, NEGS)
16:            Ti_{value} = Tf_{term,doc} × Df_{term}
17:    for term in allTerms do
18:        T_{negsval} ← average(Ti_{value}, NEGSdocs)
19:    T_{values} = ∅
20:    for term in (POS ∪ NEGS) do
21:        T_{values} ← T_{values} ∪ { absValue(T_{posval} - T_{negsval}) }
22:    sortInDecreasingOrder(T_{values})
23:    Return truncate(T_{values} , k)
```

Two Alternative Implementations

To choose the best k discriminative terms we have adopted and evaluated (Section 5) two alternative methods of computing the "final value" of each term (line 21). One approach, designated *abs*, sorts, in decreasing order, the absolute value of the difference between the Tf×Df value in the positives minus the value for the same term in the negatives, and chooses the best k of them. An alternative approach called *half-k* also computes, for each term, the differences of Tf×Df between the corresponding positive term and in the negative term but does not take the absolute value of their difference. It then chooses the $k/2$ terms achieving the most positive values (appear most frequently in relevant documents) and also the $k/2$ terms achieving the most negative values (appear most frequently in irrelevant documents). This late approach aims at making sure that representative terms from *positive* texts and *negative texts* are chosen.

Given a set of labeled examples, the goal of a classifier is to discriminate the elements of the different classes. K-BDT is based on a similar principle. K-BDT identifies terms that discriminate *relevant* documents from the *non-relevant* ones. It does that differently from the traditional $tf \times idf$ approach. In traditional text classification $tf \times idf$ promotes terms that are highly represented in a single document independent of the class it belongs. K-BDT promotes terms that are highly represented in a large number of documents of one of the classes[2] and at the same time rare in the documents of the other class. K-BDT promotes terms that are good at discriminating the two classes. Since K-BDT looks for terms highly represented in the whole set of documents of each class the experimental results show that we often need a small number of such terms to build a good classifier. This feature seems to be an advantage over the traditional $tf \times idf$ approach.

The k-BDT technique is suitable to be applied to a text classification problem in any domain and text corpus. There are no domain restrictions to the application of the technique.

4 Related Work on Feature Selection for Attribute Reduction in Full-Text Documents Classification

In the literature the work in feature selection is quite extensive, so we will highlight the feature selection work applied to the biomedical full-text document for classification purpose.

The recent work of [8] presents a study of the impact of feature selection on medical document classification. This study uses two data sets containing MEDLINE documents and makes a comparison between two different feature selections methods: the Gini Index and the Distinguish Feature Selector through two base classifiers: C4.5 decision tree and the Bayesian network. The authors also used documents from ten different disease categories for the experiments.

[2] It has been used in binary text classification but can also be adapted to non binary classification problems.

The authors concluded that the best accuracy results are a combination of the two proposed feature selection methods.

The authors in [9] present a novel method for attribute reduction using a data set of PubMed articles. The authors claim that achieved better results with their new method in terms of accuracy. The process involves a first phase of pre-processing the documents through the application of the tokenization, stemming and stop words removal. This new method is a variation of the Global Weighting Schema (GRW), that extracts unique terms from documents and these terms are weighted through the global weighting schema proposed.

The authors in [10] propose a group of scoring measures for feature selection using an SVM classifier and applied it to the OHSUMED corpus. The authors claim that the results achieved mixing their proposed scoring measures outperformed both Information Gain and Tf×IDf in some cases. According to the authors the proposed measures are more dependent of the distribution of the terms through the categories and also of the documents over the categories.

The work proposed in [16] presents a novel feature selection method to reduce the dimension of terms which takes into a new semantic space, between terms, based on the latent semantic indexing method. The idea is to appropriately capture the underlying conceptual similarity between terms and documents, which is helpful for improving the accuracy of text categorization.

Xu et al. [18] describe a work based on a very simple technique called Document Frequency thresholding (DF) that has shown to be one of the best methods in either Chinese or English text data. To improve DF Xu added the Term Frequency (TF) factor. The extended method called TFDF was tested on Reuters-21578 and OHSUMED corpora showed better results than the original DF method. Although we also use document frequency (Df), Xu approach is still quite different from the novel method reported in this paper. In Xu's work there is no concern to use directly a method that discriminates the class values by performing separate computations on each class value set of documents. Document Frequency thresholding (DF) is also a different definition than the Df used and defined in this paper.

An extensive survey on text categorization techniques can be found in [19].

5 Experimental Work and Results

Methods

The empirical evaluation was done using the OHSUMED corpus [13]. We have used five OHSUMED data sets for which we manage to collect the full texts. With that corpus we have "created" two corpus: the original corpus with the full text papers; and a corpus with the same papers but with just title and abstract. For each corpus there are five data sets (c04, c06 c14, c20 and c23) that are characterized in Table 1 for title and abstract and Table 2 for full text.

We have performed three sets of experiments. We first conducted an experiment to estimate the best values of k for the title and abstract data sets and for the full text data sets. Secondly and with the best values of k for title and

Table 1. Characterization of the data sets in the Ohsumed corpus (Title+Abstract).

Data set id	Number of relevant papers	Number of non-relevant papers
c04	2630	7755
c06	1220	8430
c14	2550	8030
c20	1220	8239
c23	3952	6778

Table 2. Characterization of the data sets in the Ohsumed corpus (full text).

Data set id	Number of relevant papers	Number of non-relevant papers
c04	5598	5598
c06	256	1582
c14	343	399
c20	239	1553
c23	683	719

abstract we have compared the performance of the three feature selection algorithms in the title and abstract corpus. Lastly and using the best values of k for full text, we have compared the algorithms in the full data set corpus.

For the experiments we have used the Support Vector Machine (SVM) algorithm from Weka [17]. A 10-fold Cross Validation procedure was used as the evaluation method. The values used for k were set to 10, 50 and 100 for the title and abstract corpus and 50, 100, 500, 1000 and 1500 for the full text corpus. Both alternative implementation (*abs* and *half-k*) were used in the assessment of the novel approach.

For the purpose of our work and concerning the Information Gain feature selection method we have used a threshold of $1.00e^{-10}$ that was a value used in a previous work [1]. In the Normalize component we have used the nominal representation.

The metric used for the evaluation of the classifiers performance was the F-measure. The F-measure value combines precision and recall, where precision is the percentage of classifications that are correct and recall is the percentage of classifications actually made by the classifier. F-measure is computed as the harmonic average of the precision and recall. The best performance of a classifier on a classification task is when the F-measure has value 1 (perfect precision and recall) and its worst performance is when the F-measure is 0.

Results

Table 3 shows the results of the experiments to assess the impact of parameter k and the two alternative methods to choose the attributes in k-BDT method. We can see from those results that the novel method does not perform well in data sets that have a small number of terms. Looking at the term-doc matrix we see a very large amount of zeros, the matrix is very sparse. There is a low probability to find a frequent term common to a large number of documents of each class value.

Table 4 shows the results of the experiments to assess the impact of parameter k and the two alternative methods to choose the attributes in k-BDT method on the full text data sets. The results are completely the opposite of the results with title and abstract. The f-measure values are well above the reference values in all data sets.

Concerning the second set of experiments we have obtained the results shown in Table 5. The results in the table show that in the case of using only title and abstract the novel method is much worse than its competitors.

Concerning the third set of experiments we have obtained the best results with the novel method in all data sets. Table 6 shows the best results of the experiments to compare the study's feature selection methods on the full text corpus. We can see that in all data sets the novel method achieves performances

Table 3. Choosing the values of k together with the best of *abs* or *half-k* alternatives. The title and abstract corpus was used. k = 100 was the best value among all alternatives tested for both *abs* and *half-k*.

Data set id	k value	method	F-measure
c04	**base line value**		0,899(0,009)
c04	100	abs	0,703(0,01)-
c04	100	half-k	0,703(0,01)-
c06	**base line value**		0,936(0,008)
c06	100	abs	0,82(0,002)-
c06	100	half-k	0,82(0,002)-
c14	**base line value**		0,907(0,009)
c14	100	abs	0,681(0,002)-
c14	100	half-k	0,681(0,002)-
c20	**base line value**		0,92(0,008)
c20	100	abs	0,823(0,005)-
c20	100	half-k	0,812(0,001)-
c23	**base line value**		0,715(0,012)
c23	100	abs	0,528(0,009)-
c23	100	half-k	0,528(0,009)-

Table 4. Choosing the values of k together with the best of *abs* or *half-k* alternatives. The full text corpus was used. k values are the best ones for each *abs* and *half-k* among other values tested.

Data set id	k value	method	F-measure
c04	**base line value**		0,888(0,01)
c04	1500	abs	0,965(0,005)+
c04	1500	half-k	0,965(0,005)+
c06	**base line value**		0,856(0,02)
c06	1000	abs	0,945(0,014)+
c06	100	half-k	0,951(0,014)+
c14	**base line value**		0,799(0,049)
c14	1000	abs	0,944(0,029)+
c14	1500	half-k	0,941(0,028)+
c20	**base line value**		0,873(0,021)
c20	1500	abs	0,951(0,016)+
c20	1500	half-k	0,95(0,017)+
c23	**base line value**		0,629(0,033)
c23	1500	abs	0,83(0,03)+
c23	1500	half-k	0,826(0,034)+

Table 5. Comparison of the feature selection methods on the corpus using only title and abstract. Cells of the table contain the average and standard deviation of F-measure of a 10-fold cross validation. IG stands for information Gain. k-BDT stands for k Best Discriminative Terms. '+' means that the value is statistically significantly better than the base line value. Base line values can be found in Table 3.

Data set id	IG	Correlation	k-BDT
c04	0,915(0,009)+	0,893(0,009)-	0,703(0,01)-
c06	0,951(0,007)+	0,936(0,008)∼	0,82(0,002)-
c14	0,923(0,008)+	0,907(0,009)∼	0,681(0,008)-
c20	0,941(0,007)+	0,92(0,008)	0,823(0,005)-
c23	0,752(0,01)+	0,715(0,012)∼	0,528(0,009)-

well above the base line value and better than the competitors. In data set c06 and using the half-k version of the k-BDT method we need only 100 attributes to achieve a very good performance.

Table 6. Comparison of the feature selection methods on the data sets using full text. Cells of the table contain the average and standard deviation of F-measure of a 10-fold cross validation. IG stands for information Gain. k-BDT stands for k Best Discriminative Terms. '+' means that the value is statistically significantly better than the base line value. Base line values can be found in Table 4.

Data set id	IG	Correlation	k-BDT
c04	0,895(0,009)+	0,888(0,01)+	0,96(0,005)+
c06	0,913(0,018)+	0,856(0,02)∼	0,951(0,014)+
c14	0,877(0,033)+	0,799(0,049)∼	0,944(0,025)+
c20	0,919(0,019)+	0,873(0,021)∼	0,951(0,016)+
c23	0,742(0,036)+	0,629(0,033)∼	0,83(0,03)+

6 Conclusions

In this paper we have presented and empirically evaluated a novel feature selection method. The method is based on the idea of finding terms that are frequent in the documents of one of the class values and infrequent in the other class values. We have compared the novel method with too other feature selection approaches for title and abstract and for full-text document classification.

The results of the novel method are much better than its competitors in all full text data sets used. However, the novel method seems to be inadequate for data sets using title and abstract only.

The results suggest that the novel method requires a very small number of attributes to achieve good performances. In one of the data sets used in the study, the novel method just need 100 attributes to achieve the best performance among the competitors.

Acknowledgements. This work was supported by the Consellería de Educación, Universidades e Formación Profesional (Xunta de Galicia) under the scope of the strategic funding of ED431C2018/55-GRC Competitive Reference Group. This work was also partially funded by the ERDF through the COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT as part of project UID/EEA/50014/2013.

References

1. Gonçalves, C.A., Iglesias, E.L., Borrajo, L., Camacho, R., Vieira, A.S., Gonçalves, C.T.: LearnSec: a framework for full text analysis. In: de Cos Juez, F., et al. (eds) Hybrid Artificial Intelligent Systems HAIS 2018, vol. 10870, pp. 502–513. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92639-1_42
2. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**, 2507–2517 (2007)
3. Markov, A.A., Nitussov, A.Y., Voropai, L., Link, D., Custance, G., Mahoney, M.S.: Classical Text in Translation: An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains (2006)

4. Borasem, P.N., Kinariwala, S.A.: Image re-ranking using information gain and relative consistency through multigraph learning (2016)
5. Vieira, A.S., Iglesias, E.L., Borrajo, L.: An HMM-based text classier less sensitive to document management problems. Bioinformatics **11**, 503–515 (2016)
6. Mladenic, D., Grobelnik, M.: Feature selection for unbalanced class distribution and Naive Bayes. In: 16th International Conference on Machine Learning (ICML), pp. 258–267. Morgan Kaufmann Publishers, San Francisco (1999)
7. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Fourteenth International Conference on Machine Learning, pp. 412–420. Morgan Kaufmann Publishers Inc., San Francisco (1997)
8. Parlak, B., Uysal, A.K.: The impact of feature selection on medical document classification. In: 11th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–5 (2016)
9. Imambi, S.S., Sudha, T.: Article: a novel feature selection method for classification of medical documents from Pubmed. Int. J. Comput. Appl. **26**(9), 29–33 (2011)
10. Monta, E., Ranilla, J., Fernandez, J., Combarro, E.F., Diaz, I.: Scoring and selecting terms for text categorization. IEEE Intell. Syst. **20**, 40–47 (2005)
11. Forman, G.: Feature selection for text classification. In: Liu, H., Motoda, H. (eds.) Computational Methods of Feature Selection, Data Mining and Knowledge Discoveries Series, pp. 257–276. Chapman and Hall/CRC, Boca Raton (2007)
12. Hall, M.A., Smith, L.A.: Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In: Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, pp. 235–239. AAAI Press (1999)
13. Hersh, W.R., Buckley, C., Leone, T.J., Hickam, D.H.: Ohsumed: an interactive retrieval evaluation and new large test collection for research. In: 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press (1994)
14. Zdravevski, E., Lameski, P., Kulakov, A., Filiposka, S., Trajanov, D., Boro, J.: Parallel computation of information gain using Hadoop and MapReduce. In: Federated Conference on Computer Science and Information Systems (2015)
15. Shang, C., Li, M., Feng, S., Jiang, Q., Fan, J.: Feature selection via maximizing global information gain for text classification. J. Know.-Based Syst. **54**, 298–309 (2013)
16. Wang, F., Li, C., Wang, J., Xu, J., Li, L.: A two-stage feature selection method for text categorization by using category correlation degree and latent semantic indexing. J. Shanghai Jiaotong Univ. (Sci.) **20**(1), 44–50 (2015)
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
18. Xu, Y., Wang, B., Li, J.T., Jing, H.: An extended document frequency metric for feature selection in text categorization. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) AIRS 2008. LNCS, vol. 4993, pp. 71–82. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-68636-1_8
19. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**, 1–47 (2002)
20. Talma Gonçalves, C., Camacho, R., Oliveira, E.: BioTextRetriever: a tool to retrieve relevant papers. Int. J. Knowl. Discov. Bioinform. **2**(3), 21–36 (2011)