The Impact of Different Training Sets on Medical Documents Classification

Roberto Gatta[1], Mauro Vallati[2], Berardino De Bari[3], and Mahmut Ozsahin[3]

[1] Department of Radiation Oncology, Università Cattolica S. Cuore, Italy
roberto.gatta.bs@alice.it
[2] School of Computing and Engineering, University of Huddersfield, UK
m.vallati@hud.ac.uk
[3] Department of Radiation Oncology, Centre Hospitalier Universitaire Vaudois, Switzerland
name.surname@chuv.ch

Abstract. The clinical documents stored in a textual and unstructured manner represent a precious source of information that can be gathered by exploiting Information Retrieval techniques. Classification algorithms can be used for organizing this huge amount of data, but are usually tested on standardized corpora, which significantly differ from actual clinical documents that can be found in a modern hospital. The result is that observed performance are different from expected ones. Given such differences, it is unclear how should be the "right" training set, and how its characteristics affects the classification performance.
In this paper we present the results of an experimental analysis, conducted on actual clinical documents from a medical Department, which aims to evaluate the impact of differently sized and assembled training sets on well-known classification techniques.

1 Introduction

In modern hospitals a large amount of clinical documents are stored in a textual and unstructured manner; these documents are precious sources of knowledge that must be exploited rather than uselessly stocked. In order to exploit such knowledge, it is fundamental to classify the documents. Information Retrieval (IR) techniques provide an established way to distinguish the documents according to their general meaning (see, for instance [1]).

The traditional approach to IR envisages to exploit a large number of already classified documents — the ground truth — for training classification algorithms. Generally, we larger the training set, the better the expected performance. Usually, IR approaches are evaluated on standard corpora, that are significantly different from documents that can be found in real-world environments. In such environments, and especially in medical ones, several factors can affect the performance of IR classifiers, and limit the usefulness of extremely large training sets. Probably, the most critical one is the so-called documents obsolescence [4]. It refers to the fact that in a clinical context the turn-over of human resources, and the introduction of new techniques and methodologies, can quickly change the text style of medical reports; documents of the training set that include obsolete terms or structure can play the role of noise for the classification process.

Therefore, the usual approach based on exploiting large training sets could be not the best technique.

In this paper we perform an experimental analysis, on about 3,000 medical documents from a Radiotherapy Department, which aims to evaluate how classification performance are affected by (i) differently sized training sets, and (ii) the similarity of training documents with a given one.

2 Considered IR Algorithms

For the sake of this investigation, we considered three existing classifiers: Rocchio [6], ESA [4] and Naive Bayes [3]. Rocchio and Naive Bayse are well-known in literature, thus they represent the state-of-the-art. while ESA is a recent and somehow different classification algorithm.

Rocchio classifier uses a Vector Space Model (VSM) [7] to generate a multi-dimensional space where a document is represented as a vector, which components are functions of the frequencies of the terms. For each class of documents, a centroid is generated. New documents are classified as members of the class whose centroid is closer. Rocchio suffers of low accuracy while it has to classify documents that are close to the boundaries of a centroid. Our implementation adopted the tf-idf [8] technique to weight the terms in documents and used an Euclidean Distance metric to measure distances from centroids.

ESA is based on the idea of entropy, and exploits a two step training process. In the first step it selects a set of terms that better helps to predict the probability $p(t_i/c_j)$ that a document is classified as c_j given the fact that it contains the term t_i. In the second step, ESA calculates the entropy values associated to each term and discharges the terms which entropy is over a given threshold. For classifying a new document, the score $score(c_j)$ of each class is determined using Equation 1. The class with higher score is selected.

$$score(c_j) = \prod_{i=1}^{n} [1 - p(c_j/t_i)] \tag{1}$$

A Naive Bayes (NB) classifier uses a Bayesian approach to calculate the probability that a document is a member of every possible class. Even if it is based on the strong hypothesis of conditional independence between features, NB usually shows good performance; moreover it allows to estimate the uncertainty by evaluating the probability ratios between all the couples of possible classes.

3 Experimental Analysis

Clinical documents were collected from a Radiotherapy Department. It contributed with discharge “forms”. Each form is composed by 21 different documents, which should be classified according to the aspect of the patient they describe; the usage of tobacco and alcohol, allergies, medications, treatment plan, etc. The total number of document is about 3,000, written in French, that were divided in 21 classes, as previously stated.

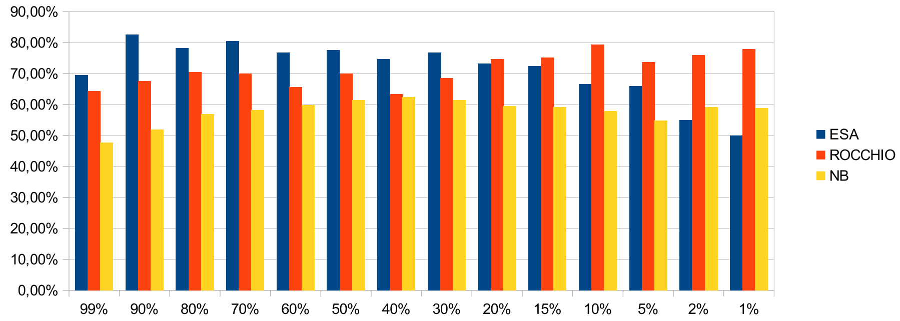

Fig. 1. Accuracy of ESA, Rocchio and Naive Bayes (NB), with regard to the considered percentage of ordered documents from the training set.

The documents are generally short (94 is the average number of words) and their structure can be significantly different, since no guidelines or “standard sentences” are proposed to physicians by the input system. We observed that different physicians wrote documents in very different ways, both from structural and syntactic point of view.

Out of the available 3,000 documents, 2,700 have been considered for training, while the rest for testing the different algorithms. Since the focus of the analysis is on the differences between large and small training sets, rather than on the testing accuracy itself, we decided to exploit a very large amount of available data for training. We considered different percentages of the learning set, ranging from 1% to 99%. In order to evaluate how quality of training instances affect classification performance, we decided to select training documents which are “close” to the given one. In other words, the given document plays the role of ”centroid” for a kNN [5] extraction which aim is to build the training set. The rationale is that, in order to limit the detrimental effect of obsolescence on classification performance, looking at documents that have been already classified and are similar to the given one should provide useful information. Similar documents can either be written in the same period or from the same physician. Distance of documents have been quantified by evaluating the euclidean distance between the tf-idf of the documents on all the words. Higher percentages of considered training documents imply that less similar documents are exploited for training IR algorithms and, potentially, that more noise is introduced.

Figure 1 shows how the considered percentages of ordered (w.r.t. tf-idf) documents of the training sets affect the accuracy performance of the IR algorithms. Remarkably, the three algorithms show different behaviours. Rocchio shows the best accuracy while using only the 10% of the available training set; its performance are then monotonically decreasing when the number of training documents growth. Naive Bayes accuracy performance remain stable while considering training sets with a size between 1 and 60% of all the available training documents. After that the accuracy decreases quickly. Finally, ESA is the only considered approach in which accuracy proportionally increases with the size of the exploited training set. It is worth noticing that all the algorithms

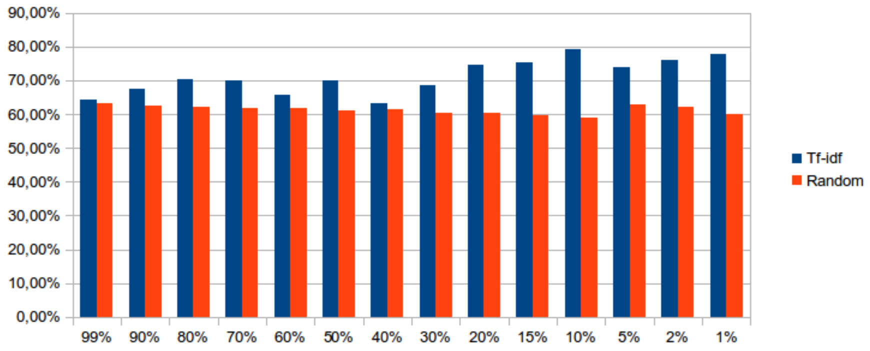

Fig. 2. Accuracy of Rocchio exploiting tf-idf based training documents selection and a random selection.

show somehow good performance, considering that the documents can be classified in 21 different classes, also when exploiting a very small number of training problems. This is probably due to the good quality of training sets, which derives from using the tf-idf technique for selecting them.

Interestingly, the average accuracy among the three algorithms monotonically increases between 1 and 20%, monotonically decreases from 70% upward, and remains almost the same in between. While a lower accuracy with very small training set was expected, it is surprising that also very large sets lead to reduced accuracy. This suggests that the approach “the more the better” should be revised and improved with better selection techniques of training sets.

For a better comprehension of the actual impact of a good selection of training instances, we compared the accuracy performance of Rocchio exploiting (i) the aforementioned tf-idf based technique and (ii) a random selection of documents from the training set. Figure 2 shows the results of such comparison. Remarkably, the performance gap is significant. A good selection of training instances lead to an evident performance improvement. Moreover, while exploiting a random selection of training documents, the size of the training set does not significantly affect the classification performance.

4 Conclusions

In modern hospitals a large amount of clinical documents, which represent a precious source of information, are stored in a textual and unstructured manner. In order to exploit such knowledge, it is fundamental to classify the documents using IR algorithms. Traditional IR approaches are tested and evaluated on standard corpora, that usually have characteristics which are very different from those of real-world documents. One of the aspects, observed in clinical documents but not in standard corpora, that has a remarkable impact on IR performance is the obsolescence, which refers to the fact that turn-over of human resources, and introduction of new techniques and methodologies, can quickly change the text style of reports. The presence of such sudden changes in real-word text corpora makes the standard learning approach — the larger the training set, the better — questionable; documents of the training set that include obsolete terms or structure, w.r.t. the current document to classify, can play the role of noise.

In this work we experimentally evaluated how differently sized and differently assembled training sets affect the classification performance of three IR approaches on clinical documents from a Radiotherapy Department. The take-home messages that can be synthesised are: (i) the size of the training set does not significantly affect the classification performance; (ii) a good selection of training instances can boost the accuracy, i.e. selecting training instances which are similar to the one to classify, according to some metrics. While the latter is intuitive, the former result is astonishing. It clearly indicates that focusing on collecting large amount of training documents is not always the best strategy for achieving good performance, at least on considered algorithms.

We see several avenues for future work. Concerning the documents, we are interested in performing a larger experimental evaluation on documents from different departments. It can be expected that, in departments that support physicians through guidelines or "standard sentences", selecting training instances will not have remarkable impact on IR performance. Moreover, we will extend the set of considered classification algorithms by including more well-known approaches (e.g., KNN [5]) and ensemble methods [2]. It will be useful to understand if such methods, which combine different classification algorithms, suffer noticeably documents obsolescence, and how different training sets affect their performance.

References

1. Bratsas, C., Koutkias, V., Kaimakamis, E., Bamidis, P.D., Pangalos, G.I., Maglaveras, N.: Knowbasics-m: An ontology-based system for semantic management of medical problems and computerised algorithmic solutions. Computer methods and programs in biomedicine 88(1), 39–51 (2007)
2. Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple classifier systems, LBCS-1857. pp. 1–15. Springer (2000)
3. Frank, E., Bouckaert, R.R.: Naive bayes for text classification with unbalanced classes. In: In Proc 10th European Conference on Principles and Practice of Knowledge Discovery in Databases. pp. 503–510 (2006)
4. Gatta, R., Vallati, M., De Bari, B., Pasinetti, N., Cappelli, C., Pirola, I., Salvetti, M., Buglione, M., Muiesan, M.L., Magrini, S., et al.: Information retrieval in medicine: an extensive experimental study. In: the 7th International Conference on Health Informatics (HealthInf) (2014)
5. Guo, G., Wang, H., Bell, D.A., Bi, Y., Greer, K.: An knn model-based approach and its application in text categorization. In: In Proc 5th International Conference on Computational Linguistics and Intelligent Text Processing. pp. 559–570 (2004)
6. Rocchio, J.J.: Relevance feedback in information retrieval. In: The Smart retrieval system - experiments in automatic document processing, pp. 313–323. Englewood Cliffs, NJ: Prentice-Hall (1971)
7. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (1975)
8. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. In: Information processing and management. pp. 513–523 (1988)