Robustness Issues in Text Mining

Marco Turchi, Domenico Perrotta, Marco Riani, and Andrea Cerioli

Abstract. We extend the Forward Search approach for robust data analysis to address problems in text mining. In this domain, datasets are collections of an arbitrary number of documents, which are represented as vectors of thousands of elements according to the vector space model. When the number of variables v is so large and the dataset size n is smaller by order of magnitudes, the traditional Mahalanobis metric cannot be used as a similarity distance between documents. We show that by monitoring the cosine (dis)similarity measure with the Forward Search approach it is possible to perform robust estimation for a document collection and order the documents so that the most dissimilar (possibly outliers, for that collection) are left at the end. We also show that the presence of more groups of documents in the collection is clearly detected with multiple starts of the Forward Search.

Keywords: Cosine similarity, document classification, forward search.

1 Introduction

In text mining, where large collections of textual documents are analyzed by automatic tools such as document classifiers or indexers to help human beings to better understand their contents, the most used document representation schema is the vector space model (VSM), introduced by [11] in information

Marco Turchi · Domenico Perrotta
European Commission, Joint Research Centre
e-mail: marco.turchi@jrc.ec.europa.eu,
domenico.perrotta@ec.europa.eu

Marco Riani · Andrea Cerioli
University of Parma, Department of Economics
e-mail: {mriani,andrea.cerioli}@unipr.it

R. Kruse et al. (Eds.): Synergies of Soft Computing and Statistics, AISC 190, pp. 263–272.
springerlink.com

retrieval. This model transforms a text in a machine readable vector assigning words to numeric vector components. Datasets are collections with an arbitrary, sometimes large, number of units n (the documents) and each unit is identified by dozens of thousands of VSM variables (the v document word identifiers). In several text mining applications there is the need of estimating a centroid for a given document collection, and to define an ordering of the documents with respect to the centroid, from the most to the least representative one. This ordering can be used to identify documents which have a weak semantic relation with the dominant subject(s) in the collection.

Outlying documents are likely to be present in most text mining applications, either because they correspond to documents which are inconsistent with the rest of the collection, or because of human mistakes in document labeling. Three popular strategies for robust estimation in presence of outliers are the following (see, e.g. [7] for a review):

1. Use a reduced number of units in order to exclude outliers from the estimation process;
2. Down-weight each unit according to its deviation from the centroid;
3. Optimize a robust objective function.

Disadvantages of these approaches are the fact that the percentage of units to be discarded needs to be fixed in advance (strategy 1), that there is no universally accepted way to down-weight observations (strategy 2) and that optimization of complex functions may cause severe computational problems (strategy 3). In addition, these strategies cannot be easily extended to heterogeneous datasets, with the purpose of identifying subgroups of similar documents in the collection. A different approach is followed by a fourth robust strategy, the Forward Search (FS) [1, 3]: instead of choosing just one subsample, a sequence of subsets of increasing size is fit and a problem-specific diagnostic is monitored in order to reveal if a new observation is in agreement with those previously included. Outliers are left at the end of the subset sequence and the effect of each unit, once it is introduced into the subset, can be measured and appraised.

With VSM, where the number of variables v is so large and the dataset size n is perhaps smaller by order of magnitudes, none of the above strategies can use traditional metrics such as the Mahalanobis distance to measure the similarity between documents, as well as the distance from an estimated centroid of the collection and any of the documents. The same drawback also affects other robust distance-based methods for cluster analysis, like TCLUST [5]. In this work, we extend the FS method to VSMs by adopting the cosine similarity [14], a metric widely used in text mining. This metric is the cosine of the angle between two vectors, which is therefore non-negative and bounded between 0 and 1. It is also independent of the vector length. More precisely, we propose to monitor the progression of the complement to one of the minimum value of the cosine similarity between the subset centroid and all units outside the subset. We will refer to this diagnostic as

to the *minimum cosine dissimilarity*. Documents will be ordered with the FS in such a way that the most dissimilar (possibly outliers, for that collection) are left at the end of sequence. We will see that the extended FS algorithm preserves the good properties shown by the FS in more traditional statistical domains, such as regression [1], multivariate analysis and clustering [3].

The paper is structured as follows. Section 2 introduces the practical motivations for the work and gives details of the data. For our demonstrations, we have used documents from a very rich source: the EuroVoc corpus. Then, since the work relies on two choices, the VSM to represent documents and the cosine similarity to measure their distance, Sect. 3 describes such choices and some related work. Section 4 provides the results and shows the potential of the FS for text mining applications. In particular, Sect. 4.1 contextualises the FS approach to text mining. All computations and simulations have been performed by extending the robust routines included in the FSDA toolbox of Matlab, downloadable from `http://www.riani.it/matlab.htm` and `http://fsda.jrc.ec.europa.eu` [10].

2 The EuroVoc Corpus

This research is driven by a real need in the development of the JRC EuroVoc Indexer (JEX) [13], a freely available multi-label categorization tool[1]. JEX is a system which automatically assigns a set of category labels from a thesaurus to a textual document. This software is based on the supervised profile ranking algorithm proposed by [9], which uses the EuroVoc thesaurus.

The EuroVoc thesaurus[2] is a multilingual, multidisciplinary thesaurus with currently about 6800 categories, covering all activities of the European Union (EU). EuroVoc's category labels have been translated one-to-one into currently 27 languages. It was developed for the purpose of manual (human) categorisation of all important documents in order to allow multilingual and cross-lingual search and retrieval in potentially very large document collections. As EuroVoc has been used to classify legal documents manually for many years, there are now tens of thousands of manually labelled documents per language that can be used to train automatic categorisation systems [9]. This collection of documents is available for download at `http://eur-lex.europa.eu/`.

The number of documents inside each category is highly unbalanced and follows the Zipf's law distribution: few categories contain more than 3000 documents, and a large number of categories has few documents. Categories belong to different domains and they can be very specific (e.g. *Fishery Management*) or very generic (e.g. *Radioactivity*). In both cases, we cannot exclude the presence of groups in the documents.

[1] `http://langtech.jrc.ec.europa.eu/JRC_Resources.html`

[2] `http://Eurovoc.europa.eu/`

Each English document of the corpus has to be preprocessed with an ordered series of operations. These include lowercasing each word (e.g. "*The White House, the*" → "*the white house, the*"), tokenizing the text ("*the white house, the*" → "*the white house , the*") and removing high frequent words (stopwords) using an external list of more than 2500 words ("*the white house , the*" → "*white house*"). This process reduces the vocabulary size and the sparseness in the data. Then we translate the documents into their VSM representations. For this purpose we count all the words in the full collection after pre-processing and we keep a variable for each corpus term, including those with zero frequency. The pre-processing work for the EuroVoc corpus thus results in a VSM vector of 119,112 variables, which is still sparse.

Models for thousands of categories are trained using only human labelled samples for each category. The training process consists in identifying a list of representative terms and associating to each of them a log-likelihood weight, with the training set used as the reference corpus. A new document is represented as a vector of terms with their frequency in the document. The most appropriate categories for the new document are found by ranking the category vector representations (called profiles), according to their similarity to the vector representation of the new document.

Despite the good performance provided by JEX, human label documents are affected by the presence of outliers: documents which are either wrongly assigned to a category or weakly correlated to the other documents into the category. The main motivation of the proposed extension of the FS is the automatic detection of these outliers, which have to be removed from the training data used by JEX.

3 Similarity in the Vector Space Model

In information retrieval the VSM was proposed to automatically retrieve documents which are similar to an input query [11]. In the VSM, a document d is represented in a high-dimensional space, in which each dimension corresponds to a term in the document. Formally, a document is a vector of v components $d = (t_1, t_2, ..., t_v)'$. A component, called *term weight*, measures how a term is important and representative. In general, v can be the vocabulary containing all terms of a natural language or all specific terms in a collection of documents. This representation produces very sparse vectors, which have only few non-zero terms.

Different options for the term weight are possible, most of which are discussed in [12]. The most used is the frequency count of a term in a document (*term frequency*). The higher the count the more likely it is that the term is a good descriptor of the content of the document. Other, more complex, approaches exist that take into account the distribution of a term in all the available documents. However, despite its limitations, the term

frequency measure is easy to compute and is still the most popular choice in text mining applications. Therefore, we restrict ourselves to a VSM where each component of d is defined as a frequency count. Similarly, in this work we do not explore possible extensions of the basic model, such as the Phrase-based VSM [8], or the Context VSM [4].

[12] arguments that the similarity between two documents may be obtained, as a first approximation, by applying the standard dot product formula on the boolean vector representation of the two documents. This representation would measure the number of terms that jointly appear in the two documents. In practice, it is preferable to use weights lying in the range $[0, 1]$, in order to provide a more refined discrimination among terms, with weights closer to 1 for the more important (frequent) terms. This naturally yields to take as a similarity measure the cosine of the angle between two VSM vectors:

$$\cos(d_1, d_2) = \frac{\sum_{i=1}^{v} d_1(i) d_2(i)}{\sqrt{\sum_{i=1}^{v} d_1^2(i)} \sqrt{\sum_{i=1}^{v} d_2^2(i)}}. \quad (1)$$

Index (1) is called the *cosine similarity* between d_1 and d_2, while $1 - \cos(d_1, d_2)$ represents the cosine dissimilarity. The value of $\cos(d_1, d_2)$ is 0 if the two vectors are orthogonal, and 1 if they are identical. By definition, the numerator takes into account only the non-zero terms of both vectors, while the denominator is affected by all components of the vectors. Note that the cosine similarity between large documents in general results in small values, because they have poor similarity values (a small scalar product and a large dimensionality).

Since its introduction, the cosine similarity has been the dominant document similarity measure in information retrieval and text mining (see e.g. [6]). A key factor for its success is its capacity of working with high-dimensional vectors, as it projects the vectors into the first quadrant of the circle of radius one. This goes at the expenses of the information lost in the drastic reduction of dimensionality. A potential drawback of working with the pairwise measure (1) is its lack of invariance under different correlation models for the v term frequencies appearing in d_1 and d_2. However, a Mahalanobis-type approach is unfeasible in text mining applications, except in very particular situations. This is the price to pay when we work with $v \gg n$.

4 Data Analysis with the Forward Search

4.1 Steps of Forward Search for Text Mining

The FS builds subsets of increasing size m, starting from a small number of units (the VSM vectors), e.g. $m_0 = 5$, until all units are included. The subsets

are built using this ordering criterion: at step m, compute the centroid of the m units in the subset and select for the next subset the $m+1$ units with smaller cosine dissimilarity from the centroid. Then, as m goes from m_0 to n, we monitor the evolution of the minimum cosine dissimilarity. In absence of outliers we expect a rather constant or smoothly increasing statistic progression. On the contrary the entry of outliers, which by construction will happen in the last subsets, will be revealed by appreciable changes of the minimum cosine dissimilarity trajectory. A similar behaviour is observed in presence of different groups when we look at the data from the perspective of a centroid fitted to one group. While for outlier detection a single forward search from a good starting subset is sufficient to reveal possible isolated outliers, for cluster identification many searches are needed. Those starting in a same group, will reveal the group presence in the form of converging group trajectories, such as those highlighted in Fig. 2. The precise identification of outliers and groups, with given statistical significance, is possible using confidence envelopes for the cosine dissimilarity, that can be found along the lines of [2]. Refer to [2] also for details on the key concepts recalled in this section.

4.2 Synthetic Data

The distribution of the terms in a corpus, which typically follows a power law (Zipf's distribution), can be easily estimated once the documents are translated into their VSM representation. Based on the estimated distribution parameters of the EuroVoc corpus, we have built synthetic datasets of 100 units and 119112 variables having cosine similarity for each pair of vectors around 0.8. Such synthetic datasets are used to study the properties of the proposed statistical analysis for a collection of documents with features mimicking those of the EuroVoc corpus.

The left panel of Fig. 1 shows the monitoring of the minimum cosine dissimilarity trajectories of 500 randomly started forward searches, for one of these synthetic datasets. A prototype trajectory is displayed by a black solid line. It is uneventful and well included within the bootstrap bands obtained by random selection of the starting point. Therefore, this plot provides evidence of what we can expect from the FS under the null hypothesis of an homogeneous collection of documents.

On the right panel of Fig. 1 five units of the same dataset have been shuffled. In the VSM this corresponds to considering 5 documents with completely different cosine similarity values from the rest of the documents in the collection. Outlyingness of these observations is clearly reflected in the plot by the large peak at the end of the searches, when the anomalous units enter into the fitting subset regardless of the actual starting point. It may also occasionally happen that a search is randomly initialised with one of such units,

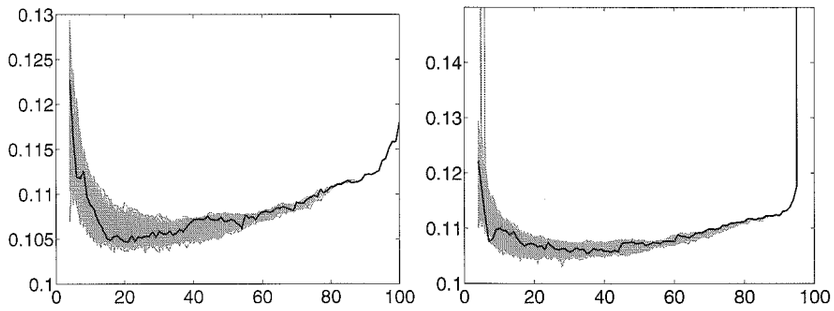

Fig. 1 500 random start forward searches for a synthetic dataset, homogeneous (left panel) and with 5 shuffled units (right panel)

but then the algorithm is immediately able to recover and to substitute the anomalous observations in the fitting subset with uncontaminated ones. In the parlance of the FS, we say that *interchanges* have occurred in the first steps of the algorithm.

4.3 EuroVoc Data

Figure 2 shows the minimum cosine dissimilarity trajectories of 500 randomly started forward searches, for two EuroVoc datasets. The left panel is about category C7, formed by 26 units and 119112 variables. The structure of this plot is very different from what we have seen in Figure 1, both for the case of

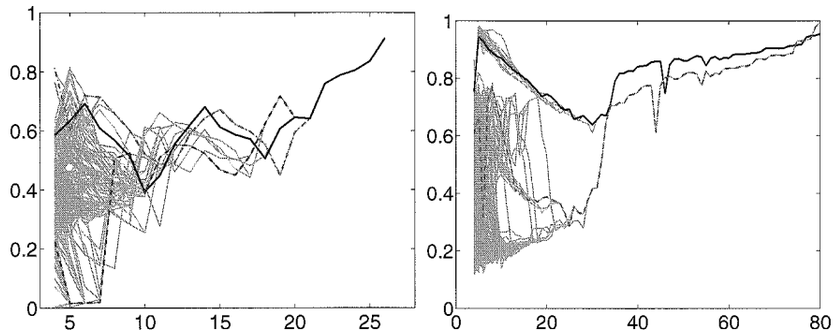

Fig. 2 500 random start forward searches for two EuroVoc datasets, classified by professional librarians to categories identified with C7 (left panel) and C174 (right panel)

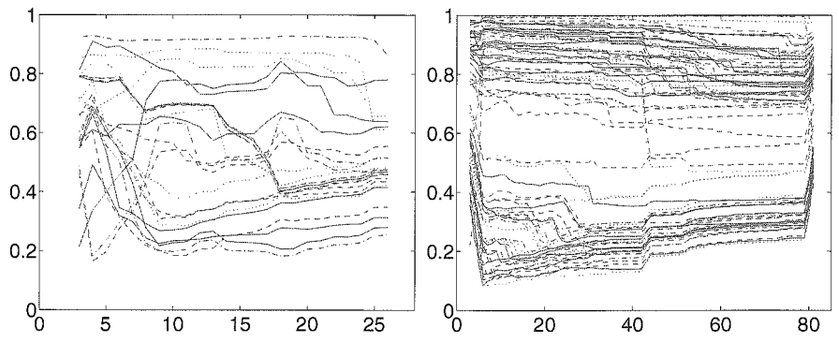

Fig. 3 Individual trajectories of the cosine dissimilarity measures for all the documents from their overall mean in one of the runs that clarifies the group structure in Figure 2. Left panel: category C7; right panel: category C174.

uncontaminated data and when outliers are present. Specifically, at step 18 there are only three groups of 418, 15 and 67 trajectories (respectively from the top to the bottom one). Each group is formed by trajectories that, starting from different initial subsets of documents, converge to the same path. This behaviour provides clear evidence of a cluster structure, because the searches that start in individual groups continue to add observations from the group until all observations in that cluster have been used in estimation. There is then a sudden change in the cosine dissimilarity measure as units from other clusters enter the subset used for estimation. We conclude that category C7 of the EuroVoc corpus cannot be considered homogeneous, but displays three different substructures. This analysis can be supplemented by a forward plot of the individual trajectories of the cosine dissimilarity measures for all the documents from their overall mean in one of the runs that clarifies the group structure in Figure 2. This plot is shown in the left panel of Figure 3. Despite the reduced sample size, three groups of trajectories with different shapes emerge, with one of them "crossing" the other two.

Our analysis is repeated for category C174. Here the pictures are even clearer, thanks to the increased sample size ($n = 81$ documents, again on 119112 variables) and to the presence of only two groups. These clusters are identified, at step 40, by two bunches of 39 and 461 trajectories, respectively, of the minimum cosine dissimilarity in the right panel of Figure 2. They are also clearly visible in the right panel of Figure 3, where the individual trajectories from the two groups are well separated.

The practical relevance of these results consists in being able to distinguish between rather homogeneous sets of documents, possibly contaminated by isolated outliers, and sets formed by different groups. Depending on the final application, outliers and subgroups can be treated differently. For instance,

groups can be used to build a committee of classifiers rather than a single one for the entire dataset.

5 Summary

In this paper we have extended the Forward Search approach for robust data analysis to address some relevant issues in text mining, such as the detection of outlying documents or the identification of possible clusters in the data. This achievement has been reached by replacing the traditional Mahalanobis metric of multivariate analysis, which cannot be applied in situations where the sample size is smaller by order of magnitudes than the number of variables, with the cosine dissimilarity measure.

It is well known that when using the VSM, documents can talk about the same theme even using very different set of terms, resulting in low cosine similarity. In this case our approach would identify different groups in the set of documents. This effect can be limited by the adoption of more sophisticated text representation schemes such as the Concept VSM, where each component of the numeric vector represents a concept that is identified by a group of semantically similar terms. As the cosine similarity is a reasonable distance also for concept vectors, our Forward Search extension to text mining would be still applicable.

References

1. Atkinson, A.C., Riani, M.: Robust Diagnostic Regression Analysis. Springer, Berlin (2000)
2. Atkinson, A.C., Riani, M.: Exploratory tools for clustering multivariate data. Comput. Stat. Data Anal. 52, 272–285 (2007)
3. Atkinson, A.C., Riani, M., Cerioli, A.: Exploring Multivariate Data with the Forward Search. Springer, Berlin (2004)
4. Billhardt, H., Borrajo, D., Maojo, V.: A context vector model for information retrieval. J. Am Soc. Inf. Sci. Tec. 53, 236–249 (2002)
5. Garcia-Escudero, L., Gordaliza, A., Matran, C., Mayo-Iscar, A.: A general trimming approach to robust cluster analysis. Ann. Stat. 36, 1324–1345 (2008)
6. Huang, A.: Similarity measures for text document clustering. In: Proc. of the 6th New Zealand Computer Science Research Student Conference, Christchurch, New Zealand, pp. 49–56 (2008)
7. Hubert, M., Rousseeuw, P.J., Van Aelst, S.: High-breakdown robust multivariate methods. Stat. Sci. 23, 92–119 (2008)
8. Mao, W., Chu, W.W.: Free-text medical document retrieval via phrase-based vector space model. In: Proc. of the AMIA Symposium, p. 489 (2002)
9. Pouliquen, B., Steinberger, R., Ignat, C.: Automatic annotation of multilingual text collections with a conceptual thesaurus In: Proc. of the Workshop Ontologies and Information Extraction at the EUROLAN 2003, Bucharest, Romania (2003)

10. Riani, M., Perrotta, D., Torti, F.: FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. Chemometr. Intell. Lab. 116, 17–32 (2012)
11. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18, 613–620 (1975)
12. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inform. Process. Manag. 24, 513–522 (1988)
13. Steinberger, R., Ebrahim, M., Turchi, M.: JRC Eurovoc Indexer JEX — A freely available multi-label categorisation tool. In: Proc. of the 8th Int. Conf. on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey (2012)
14. Yates, R.B., Neto, B.R.: Modern Information Retrieval. Addison-Wesley, Reading (1999)