J Intell Inf Syst (2015) 45:379–396
DOI 10.1007/s10844-014-0317-4

Optimizing text classification through efficient feature selection based on quality metric

Jean-Charles Lamirel · Pascal Cuxac · Aneesh Sreevallabh Chivukula · Kafil Hajlaoui

Received: 8 September 2013 / Revised: 10 January 2014 / Accepted: 23 March 2014 /
Published online: 23 May 2014

Abstract Feature maximization is a cluster quality metric which favors clusters with maximum feature representation as regard to their associated data. In this paper we show that a simple adaptation of such metric can provide a highly efficient feature selection and feature contrasting model in the context of supervised classification. The method is experienced on different types of textual datasets. The paper illustrates that the proposed method provides a very significant performance increase, as compared to state of the art methods, in all the studied cases even when a single bag of words model is exploited for data description. Interestingly, the most significant performance gain is obtained in the case of the classification of highly unbalanced, highly multidimensional and noisy data, with a high degree of similarity between the classes.

Keywords Feature maximization · Clustering quality index · Feature selection · Supervised learning · Unbalanced data · Text

J.-C. Lamirel (✉)
SYNALP Team - LORIA, INRIA Nancy-Grand Est, Vandoeuvre-les-Nancy, France
e-mail: lamirel@loria.fr
URL: http://www.loria.fr/

P. Cuxac
INIST-CNRS, Vandoeuvre-les-Nancy, France
e-mail: pascal.cuxac@inist.fr
URL: http://www.inist.fr/

A. S. Chivukula · K. Hajlaoui
Center For Data Engineering, International Institute of Information Technology,
Gachibowli Hyderabad, Andhra Pradesh, India

A. S. Chivukula
e-mail: aneesh.chivukula@gmail.com

K. Hajlaoui
e-mail: kafil_hajlaoui@yahoo.fr
URL: http://www.iiit.ac.in/

1 Introduction

Since the 1990s, advances in computing and storage capacity allow the manipulation of very large data. Whether in bioinformatics or in text mining, it is not uncommon to have description space of several thousand or even tens of thousands of features. One might think that classification algorithms are more efficient if there are a large number of features. However, the situation is not as simple as this. The first problem that arises is the increase in computation time. Moreover, the fact that a significant number of features are redundant or irrelevant to the task of classification significantly perturbs the operation of the classifiers. In addition, as soon as most learning algorithms exploit probabilities, probability distributions can be difficult to estimate in the case of the presence of a very high number of features. The integration of a feature selection process in the framework of the classification of high dimensional data becomes thus a central challenge.

The remaining of the paper is structured as follows. Section 2 presents usual approaches for feature selection. Section 3 presents our new feature selection approach. Section 4 provides more details on our experimental textual datasets. Section 5 compares the classification results with and without the use of the proposed approach on the said datasets. Section 6 draws our conclusion and perspectives.

2 Existing approaches

In the literature, three types of approaches for feature selection are mainly proposed: the integrated (embedded) approaches, the wrapper methods and the filter approaches. An exhaustive overview of the state-of-the-art techniques in this domain has been achieved by many authors, like Ladha and Deepa (2011), Bolón-Canedo et al. (2012), Guyon and Elisseef (2003) or Daviet (2009). We thus only provide hereafter a rapid overview of existing approaches and related methods.

The integrated approaches incorporate the selection of the features in the learning process (Breiman et al. 1984). The most popular methods of this category are the SVM-based methods and the neural based methods. SVM-EFR (Recursive Feature Elimination for Support Vector Machines) (Guyon et al. 2002) is an integrated process that performs the selection of features an iterative basis using a SVM classifier. The process starts with the complete feature set and remove the features given as the least important by the SVM. In an alternative way, the basic idea of the approaches of the FS-P (Feature Selection-Perceptron) family is to perform a supervised learning based on a perceptron neural model and to exploit the resulting interconnection weights between neurons as indicators of the features that may be relevant to provide a ranking (Mejía-Lavalle et al. 2006).

On their own side, wrapper methods explicitly use a performance criterion for searching a subset of relevant predictors (Kohavi and John 1997). More often it's error rate (but this can be a prediction cost or the area under the ROC curve). As an example, the WrapperSubsetEval method evaluates the feature sets using a learning approach. Cross-validation is used to estimate the accuracy of the learning for a given set of features. The algorithm starts with an empty set of features and continues until adding features does not improve performance (Witten and Frank 2005). Forman presents a remarkable work of methods' comparison in Forman (2003). As other similar works, this comparison clearly highlights that, disregarding of their efficiency, one of the main drawbacks of embedded and wrapper methods is that

they are very computationally intensive. This prohibits their use in the case of highly multi-dimensional data description space. A potential alternative is thus to exploit filter methods in such context.

Filter approaches are selection procedures that are used prior and independently to the learning algorithm. They are based on statistical tests. They are thus lighter in terms of computation time than the other approaches and the obtained features can generally be ranked regarding to the testing phase results.

The Chi-square method exploits a usual statistical test that measures the discrepancy to an expected distribution assuming that a feature is independent of a class label (Ladha and Deepa 2011). The information gain is also one of the most common methods of evaluation of the features. This univariate filter provides an ordered classification of all features. In this approach, selected features are those that obtain a positive value of information gain (Hall and Smith 1999).

In the MIFS (Mutual Information Feature Selection) method, a feature f is added to the subset M of already selected features if its link with the target Y surpasses its average connection with already selected predictors. The method takes into account both relevance and redundancy. In a similar way, the CFS method (Correlation-based Feature Selection) uses a global measure of "merit" of a subset M of m features. Then, a relevant subset consists of features highly correlated with the class, and lowly correlated one to another (Hall and Smith 1999).

The CBF (Consistency-based Filter) method evaluates the relevance of a subset of features by the resulting level of consistency of the classes when learning samples are projected onto that subset (Dash and Liu 2003).

The MODTREE method is a correlation-based filtering method that relies on the principle of pairwise correlation. The method operates in the space of pairs of individuals described by co-labeling indicators attached to each original feature. For that, a pairwise correlation coefficient that represents the linear correlation between two features is used. Once the pairwise correlations are tabulated, the calculation of partial correlation coefficients allows performing a stepwise feature selection (Lallich and Rakotomalala 2000).

The basic assumption of the Relief feature ordering method is to consider a feature as relevant if it discriminates well an object in the positive class from its nearest neighbor in the negative class. The score of the features is a cumulative score computed thanks to a random selection of objects. ReliefF, an extension of Relief, adds the ability to address multiclass problems. It is also more robust and capable of handling incomplete and noisy data (Kononenko 1994). This latter method is considered as one of the most efficient filter-based feature selection technique.

Like any statistical test, filter approaches are known to have erratic behavior for very low features' frequencies (which is a common case in text classification) (Ladha and Deepa 2011). Moreover, we show in this paper that, despite their diversity, all the existing filter approaches also fail to successfully solve the feature selection task in case they are faced with highly unbalanced, highly multidimensional and noisy data, with a high degree of similarity between the classes.

On their own side, resampling methods aim at correcting class imbalance by either adding new artificial samples to the minority classes (oversampling) or suppressing some samples of the majority classes (undersampling) (Good 2006). As an example, Chawla et al. (2002) proposed the successful SMOTE oversampling technique in 2002 whose main principle is to synthesize new minority class examples between several minority examples that lie together, rather than simply duplicating them as in random over-sampling.

However, we show in this paper that in such complex context as the one of the classification of textual data with highly imbalanced and similar classes, the ability of all the above mentioned techniques to precisely detect the right class is curtailed by the high class to class similarity. We alternatively propose a new feature selection and contrasting approach based on the recently developed feature maximization metric and we compare its performance with standard techniques in the patents validation assistance context. We then extend the scope of our comparison to usual reference datasets.

3 Feature maximization for feature selection

Feature maximization is an unbiased cluster quality metrics that exploits the features of the data associated to each cluster without prior consideration of clusters profiles. This metrics has been initially proposed in Lamirel et al. (2004). Its main advantage is to be independent altogether of the clustering method and of its operating mode. When it is used during the clustering process, it can substitute to usual distances during that process (Lamirel et al. 2011). In a complementary way, whenever it is used after learning, it can be exploited to set up overall clustering quality indexes (Lamirel et al. 2010) or for cluster labeling (Lamirel and Ta 2008).

Feature maximization is a metric which favours clusters with maximum feature F-measure. *Feature F-measure* (FF) is the harmonic mean of *Feature recall* (FR) and *Feature precision* (FP) which in turn are defined as:[1]

$$FR_g(f) = \frac{\sum_{x \in g} W_x^f}{\sum_{g \in G} \sum_{x \in g} W_x^f} \qquad FP_g(f) = \frac{\sum_{x \in g} W_x^f}{\sum_{f \in F_g, x \in g} W_x^f} \tag{1}$$

where W_x^f represents the weight of the feature f for element x,[2] F_g designates the set of features associated with the data occuring in cluster g which is associated to a given prototype p_g and G represents the global set of clusters of the clustering. A feature is then said to be maximal for a given cluster iff its feature F-measure is higher for that cluster than for any other cluster. Finally the feature F-measure FF_g of a cluster $g \in G$ is the average of the feature F-measures of the maximal features for c:

$$FF_g = \frac{\sum_{f \in F_g} FF_g(f)}{|F_g|} \tag{2}$$

An important application of feature maximization metric is related to clusters labeling whose role is to highlight the prevalent features of the clusters associated to a clustering model at a given time. Labeling can thus be used altogether for visualizing or synthesizing clustering results and for optimizing the learning process of a clustering method (Attik et al. 2006). It can rely on endogenous data features or on exogenous ones. Endogenous data features represent the ones being used during the clustering process. Exogenous data

[1] Since *Feature recall* is equivalent to the conditional probability $P(g|p)$ and *Feature precision* is equivalent to the conditional probability $P(p|g)$, this former strategy can be classified as an expectation maximization approach with respect to the original definition given by Dempster et al. (1977). Harmonic mean provides an additional influence to the lowest of the two values in the combination of *feature recall* and *feature precision*.

[2] See Section 4 for more details on usual weighting schemes exploited on textual data.

features represent either complementary features or specific validation features. Exploiting feature maximization metric for cluster labeling results is a parameter-free labeling technique (Lamirel and Ta 2008). As regards to this approach, a feature is then said to be maximal or prevalent for a given cluster iff its Feature F-measure is higher for that cluster than for any other cluster. Thus the set L_g of prevalent features of a cluster g can be defined as:

$$L_g = \left\{ f \in F_g \mid FF_g(f) = \max_{g' \in G} \left(FF_{g'}(f) \right) \right\} \tag{3}$$

Whenever it has been exploited in combination with hypertree representation, this technique has highlighted promising results, as compared to the state-of-the-art labeling techniques, like Chi-square labeling, for synthetizing complex hierarchical clustering output issued from the management of highly multidimensional data (Lamirel and Ta 2008). Additionally, the combination of this technique with unsupervised Bayesian reasoning resulted in the proposal of the first parameter-free fully unsupervised approach for analyzing the textual information evolving over time. Exhaustive experiments on large reference datasets of bibliographic records have shown that the approach is reliable and likely to produce accurate and meaningful results for diachronic scientometrics studies (Lamirel 2012).

Taking into consideration the basic definition of feature maximization metric presented above, its exploitation for the task of feature selection in the context of supervised learning is a natural process, as soon as this generic metric can apply on data associated to a class as well as to those associated to a cluster. The feature maximization-based selection process can thus be defined as a parameter-free class-based process in which a class feature is characterized using both its capacity to discriminate a given class from the others ($FP_c(f)$ index) and its capacity to accurately represent the class data ($FR_c(f)$ index). The set S_c of features that are characteristic of a given class c belonging to an overall class set C results in:

$$S_c = \left\{ f \in F_c \mid FF_c(f) > \overline{FF}(f) \text{ and } FF_c(f) > \overline{FF}_D \right\} \text{ where} \tag{4}$$

$$\overline{FF}(f) = \Sigma_{c' \in C} \frac{FF_{c'(f)}}{|C_{/f}|} \text{ and } \overline{FF}_D = \Sigma_{f \in F} \frac{\overline{FF}(f)}{|F|} \tag{5}$$

and $C_{/f}$ represents the restriction of the set C to the classes in which the feature f is represented. Finally, the set of all the selected features S_C is the subset of F defined as $S_C = \cup_{c \in C} S_c$.

Features that are judged relevant for a given class are the features whose representation is altogether better than their average representation in all the classes including those features and better than the average representation of all the features, as regard to the feature F-measure metric. In the specific framework of the feature maximization process, a contrast enhancement step can be exploited complementary to the former feature selection step. The role of this step is to adapt the description of each data to the specific characteristics of its associated class which have been formerly highlighted by the feature selection step. In the case of our metric, it consists in modifying the weighting scheme of the data specifically to each class by taking into consideration the "information gain" provided by the Feature F-measures of the features, locally to that class.

Thanks to the former strategy, the "information gain" provided by a feature in a given class is considered as proportional to the ratio between the value of the Feature F-measure of this feature in the class $FF_c(f)$ and the average value of the Feature F-measure of the said feature on all the partition $\overline{FF}(f)$. For a given data and a given feature describing this data, the resulting gain acts as a contrast weight factorizing with any existing feature weight

that can be issued from data preprocessing. For a feature f belonging to the set of selected features S_c of a class c, the gain $G_c(f)$ is expressed as:

$$G_c(f) = \left(\frac{FF_c(f)}{\overline{FF}(f)} \right)^k \tag{6}$$

where k is a magnification factor that can be optimized based on the resulting accuracy.

Active features of one class are those for which information gain is greater than 1 in that class. As soon as resulting method is a class-based feature selection and contrasting method, the average number of active variables per class is consequently comparable to the overall number of selected variables with global selection methods.

4 Experimental datasets

Our main ressource is a collection of patent documents related to pharmacology domain and issued from the QUAERO[3] context. The bibliographic citations in the patents are extracted from the Medline database.[4] The source data contains 6387 patents in XML format, grouped into 15 subclasses of the A61K class (medical preparation). 25887 citations have been extracted from 6387 patents (Hajlaoui et al. 2012). Then the Medline database is queried with extracted citations for related scientific articles. The querying gives 7501 articles. Each article is then labeled by the first class code of the citing patent. The set of labeled articles represents the final document set on which the training is performed. The final document set is highly unbalanced, with smallest class containing 22 articles (A61K41 class) and largest class containing 2500 articles (A61K31 class). Inter-class similarity computed using cosine correlation indicates that more than 70 % of classes' couples have a similarity between 0.5 and 0.9. Thus the ability of any classification model to precisely detect the right class is curtailed. A common solution to deal with unbalance in dataset is undersampling majority classes and oversampling minority classes. However, resampling that introduces redundancy in dataset does not improve the performance in this dataset, as it has been shown in Hajlaoui et al. (2012). We thus propose hereafter to prune irrelevant features and to contrast the relevant ones as an alternative solution.

The abstract of each article is processed and converted into a bag of words model (Salton 1971) using the TreeTagger tool (Schmid 1994) developed by the Institute for Computational Linguistics of the University of Stuttgart. This tool is both a lemmatizer and a tagger. As a result, each article is represented by a vector of terms which have been extracted from the abstract. In such vector, terms are represented by their frequency in the abstract. The description space generated by the tagger has dimensionality 31214. To reduce noise generated by the TreeTager tool, a frequency threshold of 45 (i.e. an average threshold of 3/class) is applied on the extracted descriptors. It resulted in a thresholded description space of dimensionality 1804. A final step of Term Frequency-Inverse Document Frequency (TF-IDF) weighting (Salton and Buckley 1988) is applied on resulting articles' descriptions. We abbreviate the resulting dataset as PAT-QUAERO.

[3]The QUAERO project was initiated to meet multimedia content analysis requirements for consumers and professionals facing the rapid increase of accessible digital information. This collaborative research and development project focuses on the areas of automatic extraction of information, analysis, classification and usage of digital multimedia content for professionals and consumers. One specific subtask of the project is to develop automatic patents' validation tools.

[4]http://www.ncbi.nlm.nih.gov/pubmed/

Five other reference datasets which are described hereafter are considered in our experiments on textual data:

- R8 and R52 datasets[5] are respective adaptations achieved by Cardoso Cachopo of the R10 and R90 datasets issued from the Reuters 21578 collection.[6] The goal of these adaptations is to consider only single labeled datasets. For that purpose the documents with more than one topic are eliminated. Considering only the documents with a single topic and the classes which still have at least one train and one test example and following Sebastiani's convention, R8 is a reduction to 8 classes of the R10 (10 most frequent classes) dataset and R52 is a reduction to 52 classes of the R90 dataset (90 classes). The R8 and R52 have respective size of 7674 and 9100 and associated bag of words description spaces of 1187 and 2618 words.
- The 20Newsgroups dataset (Ken Lang 1995) is a collection of approximately 20,000 newsgroup documents partitioned (nearly) evenly across 20 different newsgroups. Although already cleaned-up, this dataset still had several attachments, many PGP keys and some duplicates. We consider two bag of words versions of the dataset. In the all-terms version (20N-AT), all words are kept and only non-alphabetic characters are turned into spaces. It resulted in an 11153 words description space. In the stemmed version (20N-ST), the words with less than 2 characters, as well as the stop words issued from the S24 SMART stop word list (Salton 1971), are eliminated. Moreover, the remaining words are stemmed using the Porter's stemmer (Porter 1980). The description space is thus reduced to 5473 words.
- Amazon^{TM} commerce reviews set (AMZ) is an UCI dataset (Bache and Lichman 2013) derived from the customers' reviews in Amazon commerce website and exploitable for authorship identification. To examine the robustness of classification algorithms to large number of target classes, 50 of the most active users (represented by a unique ID and username) who frequently posted reviews in these newsgroups are identified. The number of reviews collected for each author is 30. Each review includes authors' linguistic style such as usage of digit, punctuation, frequent words and sentences. For that reason, all words including above-mentioned signs are kept in this dataset and the resulting bag of words description space size of the collection holds 10000 words.
- The original WebKB dataset (WKB) contains 8282 webpages collected from computer science departments of various universities in January 1997 by the World Wide Knowledge Base (Web→KB) project of the CMU text learning group.[7] The pages have been manually classified into 7 classes: student, faculty, department, course, staff, project, other. We exploit the Cardoso Cachopo reduced version of the dataset in which department and staff classes have been discarded because of their low page count and the miscellaneous class has been discarded as well. Same cleaning and stemming processes on the rough indexing terms as the ones applied to 20Newsgroups dataset are performed. It resulted in a dataset of size 4158 described by a 1805 words description space.

[5] http://web.ist.utl.pt/~acardoso/datasets/

[6] http://www.research.att.com/~lewis/reuters21578.html

[7] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

5 Experiments and results

5.1 Experiments

To perform our experiments we firstly take into consideration different classification algorithms which are implemented in the Weka toolkit:[8] J48 Decision Tree algorithm (Quinlan 1993), Random Forest algorithm (Breiman 2001) (RF), KNN algorithm (Aha and Kibler 1991), Bayesian algorithms, like Multinomial Nave Bayes (MNB) and Bayes Net (BN), and finally, SMO-SVM algorithm (Platt 1998) (SMO). Most of these algorithms are general purpose classification algorithms, except from MNB which is a Discriminative Multinomial Naive Bayes classifier especially developed for text classification. Other general purpose algorithms whose accuracy has especially been reported for text classification are SMO and KNN (Zhang and Oles 2001). Default parameters are used when executing these algorithms, except for some methods where the parameters are optimized based on resulting accuracy. These parameters relate to:

- the number of neighbors for KNN.
- the number of trees and the number of features used in random selection for RF.
- the type of kernel, the kernel parameters and the complexity parameter (C) for SMO.

We then more especially focus on the efficiency testing of the feature selection approaches including our new proposal that we abbreviated as Feature Maximization and Contrast (FMC). We include in our test a panel of filter approaches which are computationally tractable with high dimensional data, making again use of their Weka toolkit implementation. The panel of tested methods includes: Chi-square selector (Ladha and Deepa 2011), Information gain selector (Hall and Smith 1999), CBF subset selector (Dash and Liu 2003) (CBF), Symmetrical Uncertainty selector (Yu and Liu 2003), ReliefF selector (Kononenko 1994) (RLF), Principal Component Analysis selector (Pearson 1901) (PCA). Defaults parameters are also used for most of these methods, except for PCA for which the percentage of explained variance is tuned based on resulting accuracy.

Finally, the SMOTE approach (Chawla et al. 2002) is included in our experimental process to figure out the efficiency of resampling techniques.

We first experiment the methods separately. In a second phase we combine the feature selection provided by the methods with the feature contrasting technique we have proposed. 10-fold cross validation is used on all our experiments.

5.2 Results

The different results are reported in Tables 1, 2, 3, 4, 5, 6, 7, 8 and 9 and in Figs. 1, 2, 3, 4, 5, 6, 7, and 8. Tables and figures present standard performance measures (True Positive Rate (TP) or Recall (R), False Positive Rate (FP), Precision (R), F-measure (F) and ROC) weighted by class sizes and averaged over all classes. For each table, and each combination of selection and classification methods, a performance increase indicator is computed using the SMO True Positive results on the original data as the reference. Finally, as soon as the results are identical for Chi-square, Information Gain and Symmetrical Uncertainty, they are thus reported only once in the tables as Chi-square results (and noted CHI+).

[8] http://www.cs.waikato.ac.nz/ml/weka/

Table 1 Classification results on initial data (PAT-QUAERO dataset)

	TP R	FP	P	F	ROC	TP Incr
J48	0.42	0.16	0.40	0.40	0.63	−23 %
RandomForest	0.47	0.22	0.51	0.40	0.76	−13 %
SMO	**0.54**	**0.14**	**0.53**	**0.52**	**0.80**	**(Reference)**
BN	0.48	0.14	0.47	0.47	0.78	−10 %
MNB	0.53	0.18	0.54	0.47	0.85	−10 %
KNN (k = 3)	0.53	0.16	0.53	0.51	0.77	−2 %

Bold data highlights the best results

Table 2 Classification results after feature selection on PAT-QUAERO dataset (BN classification, All feature selection methods)

	TP R	FP	P	F	ROC	Nbr. Feat	TP Incr
CHI+	0.52	0.17	0.51	0.47	0.80	282	−4 %
CBF	0.47	0.21	0.44	0.41	0.75	37	−13 %
PCA (50 % vr.)	0.47	0.18	0.47	0.44	0.77	483	−13 %
RLF	0.52	0.16	0.53	0.48	0.81	937	−4 %
FMC	**0.99**	**0**	**0.99**	**0.99**	**1**	**262/cl**	**+90 %**

Bold data highlights the best results

Table 3 Classification results after FMC feature selection on PAT-QUAERO dataset (All classification methods)

	TP R	FP	P	F	ROC	TP Incr
J48	0.80	0.05	0.79	0.79	0.92	+48 %
RandomForest	0.76	0.09	0.79	0.73	0.96	+40 %
SMO	0.92	0.03	0.92	0.91	0.98	+70 %
BN	**0.99**	**0**	**0.99**	**0.99**	**1**	**+90 %**
MNB	0.92	0.03	0.92	0.92	0.99	+71 %
KNN (k = 3)	0.66	0.14	0.71	0.63	0.85	+22 %

Bold data highlights the best results

Table 1 highlights that performance of all classification methods are low on the PAT-QUAERO dataset if no feature selection process is performed. It also confirms the superiority of the SMO, KNN and Bayes methods on the two other tree-based methods in that context. Additionally, SMO provides the best overall performance in terms of discrimination as it is illustrated by its highest ROC value. However, as it is also shown by confusion matrix of Fig. 1, the method is clearly not exploitable in an operational patent evaluation

Table 4 Classification results after feature selection by all methods + F-max contrasting on PAT-QUAERO dataset (BN classification)

	TP R	FP	P	F	ROC	Nbr. Feat	TP Incr
CHI+	0.79	0.08	0.82	0.78	0.98	282	+46 %
CBF	0.63	0.15	0.69	0.59	0.90	37	+16 %
PCA (50 % vr.)	0.71	0.11	0.73	0.67	0.53	483	+31 %
RLF	0.79	0.08	0.81	0.78	0.98	937	+46 %
FMC	**0.99**	**0**	**0.99**	**0.99**	**1**	**262/cl**	**+90 %**

Bold data highlights the best results

Table 5 Class data and FMC selected features/class on PAT-QUAERO dataset (BN Classification)

Class label	Class size	Selected features	TP Rate FMC	TP Rate before
a61k31	2533	223	**1**	0.79
a61k33	60	276	**0.95**	0.02
a61k35	459	262	**0.99**	0.31
a61k36	212	278	**0.95**	0.23
a61k38	1110	237	**1**	0.44
a61k39	1141	240	**0.99**	0.65
a61k41	22	225	**0.24**	0
a61k45	304	275	**0.98**	0.09
a61k47	304	278	**0.99**	0.21
a61k48	140	265	**0.98**	0.12
a61k49	90	302	**0.93**	0.26
a61k51	78	251	**0.98**	0.26
a61k6	47	270	**0.82**	0.04
a61k8	87	292	**0.98**	0.02
a61k9	759	250	**1**	0.45

Bold data highlights the best results

context because of the high resulting confusion between classes. It highlights its intrinsic incapacity to cope with the very important data attraction effect of the biggest classes.

Whenever a usual feature selection process is performed in combination with the best methods, its exploitation slightly alters the quality of the results, instead of bringing up an added value, as it is shown in Table 2. The same remarks can be done concerning the use of resampling technique, like SMOTE, which does not produce any performance increase on the PAT-QUAERO dataset. Alternatively, Table 2 highlights that the feature reduction of F-max selection method is similar to Chi square[9] but its combination with F-max data description contrasting boosts the performances of the classification methods, and especially the ones of the Bayes methods (Table 3), leading to awesome classification results

[9] In terms of active variables (see Section 3 for details).

Table 6 Classification results after FMC feature selection (5 reference datasets and MNB or BN classification)

		TP (R)	FP	P	F	ROC	TP Incr.
Reuters8 (R8)	–	0.93	0.02	0.92	0.93	0.98	
	FMC	**0.99**	**0**	**0.99**	**0.99**	**1**	**+6 %**
Reuters52 (R52)	–	0.91	0.01	0.90	0.90	0.98	
	FMC	**0.99**	**0**	**0.99**	**0.99**	**1**	**+10 %**
Amazon	–	0.74	0.05	0.78	0.74	0.98	
	FMC	**0.99**	**0**	**0.99**	**0.99**	**1**	**+33 %**
20Newsgroups	–	0.88	0.01	0.88	0.81	0.99	
(All-terms)	**FMC**	**0.99**	**0**	**0.99**	**0.99**	**1**	**+13 %**
20Newsgroups	–	0.86	0.01	0.87	0.86	0.99	
(Stemmed)	**FMC**	**0.99**	**0**	**0.99**	**0.99**	**1**	**+15 %**
WebKB	–	0.84	0.07	0.84	0.84	0.95	
	FMC	**0.99**	**0**	**0.99**	**0.99**	**1**	**+18 %**

Bold data highlights the best results

(Accuracy of 0.99 %, i.e. 94 misclassified instances among a total of 7252 with BN method) in a very complex classification context.

The results presented in Table 4 more specifically illustrates the efficiency of the FMC contrasting procedure that acts on the data descriptions. In the experiments related to that table, FMC contrasting is performed individually on the features extracted by each selection method and, in a second step, BN classifier is applied on the resulting contrasted data. The results show that, whatever is the kind of feature selection technique that is used, resulting classification performance is enhanced whenever is a former step of F-max data description contrasting is performed. The average performance increase is 44 %.

Table 5 and Fig. 2 illustrate the capabilities of the FMC approach to efficiently cope with the class imbalance and class similarity problems. Hence, the joint examination of TP rate changes (especially in small classes) in Table 5 and confusion matrix of Fig. 2 shows that

Table 7 Dataset information an complementary results after FMC feature selection (5 reference datasets and MNB or BN classification)

	R8	R52	AMZ	20N-AT	20N-ST	WKB
Nb. class	8	52	50	20	20	4
Nb. data	7674	9100	1500	18820	18820	4158
Nb feat.	3497	7369	10000	11153	5473	1805
Nb. sel. feat.	1186	2617	3318	3768	4372	725
Act. feat./class (av.)	268.5	156.05	761.32	616.15	525.95	261
Magnification factor	4	2	1	4	4	4
Misclassed (Std)	373	816	378	2230	2544	660
Misclassed (FMC)	**19**	**91**	**3**	**157**	**184**	**17**
Comp. time (s)	1	3	1.6	10.2	4.6	0.8

Bold data highlights the best results

Table 8 List of highest contrasted features (stemmed forms) for the 8 classes of the REUTER8 dataset

Trade	Grain	Ship	Acq
6.35 tariff	5.60 agricultur	6.59 ship	5.11 common
5.49 trade	5.44 farmer	6.51 strike	4.97 complet
5.04 practic	5.33 winter	6.41 worker	4.83 file
4.86 impos	5.15 certif	5.79 handl	4.65 subject
4.78 sanction	4.99 land	5.16 flag	4.61 tender
4.77 japanes	4.94 soviet	5.06 bulk	4.53 share
4.76 bilater	4.90 grain	5.04 wind	4.45 merger
4.73 washington	4.87 spark	5.03 gulf	4.36 transact
4.52 semiconductor	4.84 provinc	4.89 brazilian	4.35 subsidiari
4.42 surplu	4.77 bad	4.87 contain	4.312 acquir
Learn	**Money-fx**	**Interest**	**Crude**
7.57 net	6.13 currenc	5.95 rate	6.99 oil
7.24 loss	5.55 dollar	5.85 prime	5.20 ceil
6.78 profit	5.52 germani	5.12 point	4.94 post
6.19 prior	5.49 shortag	5.10 percentag	4.86 quota
5.97 split	5.16 stabil	4.95 surpris	4.83 crude
5.74 earn	4.87 assist	4.70 lend	4.48 offshor
5.09 gain	4.79 pari	4.41 yield	4.46 output
4.88 jan	4.70 underli	4.39 barclai	4.15 light
4.87 mln	4.65 governor	4.26 borrow	4.12 intermedi
4.60 oper	4.51 accord	4.25 cut	4.07 price

the data attraction effect of the biggest classes that occurs at a high level in the case of the exploitation of the original data (Fig. 1) is quite completely overcome whenever the FMC approach is exploited (Table 5 and Fig. 2). In the same table, the capability of the approach to correct class imbalance is also clearly highlighted by the homogeneous distribution of the active features (see Section 3 for details) in the classes it provides, despite of the very heterogeneous sizes of these latter.

The summary of the results on the 5 complementary reference datasets described in Section 4 are presented in Tables 6, 7. They highlight that the FMC method can very significantly enhance the performance of the classifiers in various cases. As in the former PAT-QUAERO context, the best performance upgrade is obtained by the use of the FMC

Table 9 Classification results on UCI Wine dataset

	TP R	FP	P	F	ROC	TP Incr
J48	0.94	0.04	0.94	0.94	0.95	(Reference)
FMC + BN	**1**	**0**	**1**	**1**	**1**	**+6 %**

Bold data highlights the best results

```
=== Confusion Matrix ===

a    b    c    d    e    f    g    h    i    j    k    l    m    n    o    <-- classified as
2073   0   11    3  154  182    0    8    8    1    1    0    0    0   92 |    a = a61k31
  48   0    0    1    2    3    0    0    0    0    0    0    0    0    6 |    b = a61k33
 262   0   35    4   57   92    0    0    1    0    0    0    0    0    8 |    c = a61k35
 182   0    0    6    8   13    0    1    0    0    0    0    0    0    2 |    d = a61k36
 651   0   15    0  225  199    0    0    7    0    0    0    0    0   13 |    e = a61k38
 459   0    3    0  124  523    0    0    6    0    1    0    0    0   25 |    f = a61k39
  10   0    0    0    2    4    0    0    0    0    0    0    0    0    6 |    g = a61k41
 219   0    3    0   35   30    0   12    1    0    0    0    0    0    4 |    h = a61k45
 132   0    1    0   32   58    0    0   32    0    1    0    0    0   48 |    i = a61k47
  72   0    2    0   26   38    0    0    1    0    0    0    0    0    1 |    j = a61k48
  41   0    2    0    5   13    0    0    3    0    0    1    0    0   25 |    k = a61k49
  50   0    0    0   18    8    0    0    0    0    0    0    0    0    2 |    l = a61k51
  21   0    1    1   10    7    0    0    0    0    0    0    0    0    7 |    m = a61k6
  66   0    1    0    5    5    0    1    0    0    0    0    0    0    9 |    n = a61k8
 368   0    5    1   33   60    0    0   12    0    0    0    0    0  280 |    o = a61k9
```

Fig. 1 Confusion matrix of the optimal results before feature selection on PAT-QUAERO dataset (SMO classification)

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j    k    l    m    n    o   <-- classified as
2523    0    1    0    2    0    0    7    0    0    0    0    0    0    0 |    a = a61k31
   2   55    0    0    0    0    0    0    1    0    0    0    0    0    2 |    b = a61k33
   4    0  454    0    1    0    0    0    0    0    0    0    0    0    0 |    c = a61k35
   6    0    0  205    0    1    0    0    0    0    0    0    0    0    0 |    d = a61k36
   7    0    0    1 1100    0    0    1    0    0    0    0    1    0    0 |    e = a61k38
   1    0    0    0    2 1138    0    0    0    0    0    0    0    0    0 |    f = a61k39
   0    1    0    2    1    0    3    0    1    1   10    1    1    1    0 |    g = a61k41
   5    0    0    0    1    0    0  298    0    0    0    0    0    0    0 |    h = a61k45
   1    0    0    0    0    0    0    0  303    0    0    0    0    0    0 |    i = a61k47
   1    0    2    0    1    0    0    0    1  134    1    0    0    0    0 |    j = a61k48
   0    0    0    0    0    0    0    0    2    0   88    0    0    0    0 |    k = a61k49
   0    0    0    1    0    0    0    0    1    0    0   75    0    0    1 |    l = a61k51
   0    0    0    8    1    0    0    0    1    0    1    0   34    0    2 |    m = a61k6
   0    0    0    2    0    0    0    0    0    0    0    0    0   84    1 |    n = a61k8
   1    0    0    0    0    0    0    0    0    0    0    0    0    0  758 |    o = a61k9
```

Fig. 2 Confusion matrix of the optimal results after FMC feature selection on PAT-QUAERO dataset (BN classification)

Fig. 3 Comparative distribution of active FMC features and documents/class (20Newsgroups datasets)

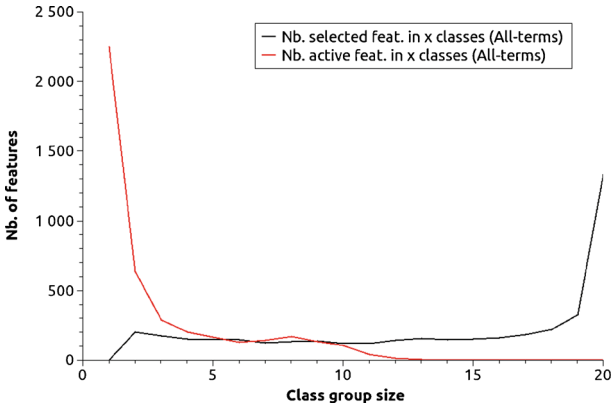

Fig. 4 Comparative trends of active and selected features shared by group of classes - by group size (20Newsgroups datasets)

method in combination with MNB and BN Bayes classifiers. Table 6 shows the comparative performance of such a combination with the direct use of the said classification methods. It also highlight the FMC method is especially efficient for increasing the performance of the classifiers whenever the complexity of the classification task becomes higher due to an increasing number of classes. Table 7 provides general information on the datasets and on the selection process. It illustrates the significant decrease in classification complexity obtained with FMC due to the reduction of the number of features to be managed, as well as the parallel decrease of misclassified data. It also highlights the very moderate computation time of the FMC method.[10]

On these datasets, similar remarks as the one mentioned for the PAT-QUAREO dataset can be done concerning the poor efficiency of usual feature selection and resampling methods. Another interesting observation is provided by the comparison between the results obtained with and without the exploitation of stemming on the 20Newgroups dataset. Indeed, the FMC method is able to maintain the performance of the classifiers even if no stemming is used.

Similarly to the observations achieved in the PAT-QUAERO context, Fig. 3 confirms that an almost uniform distribution of the active variables between classes can be obtained with FMC, whatever the sizes of the said classes are. It also highlights that the number and the distribution of active features per class remain almost stable when the number of initial features decreases by a factor 2 (i.e. from 11153 features to 5473 features between the two 20Newsgroups datasets). Addtionally, Fig. 4 shows that the number of active features that are common to large groups of classes remains limited although a large number of selected features can be shared by such groups. It also clearly points out that the FMC method put the focus on the activity of the features which are discriminant for the classes as soon as the distribution of active feature among groups follow an opposite trend to the one of selected features among groups.

Table 7 illustrates that the value of the contrast magnification factor (4) that is exploited to get best performance can vary over the experiments (i.e. from 1 to 4). However, it can be observed that setting this factor at a fixed value, like the highest one used (here 4), is not

[10]The computation is performed on Linux with a laptop equipped with Intel®Pentium® cpu B970 2.3Ghz and with 8Go standard memory.

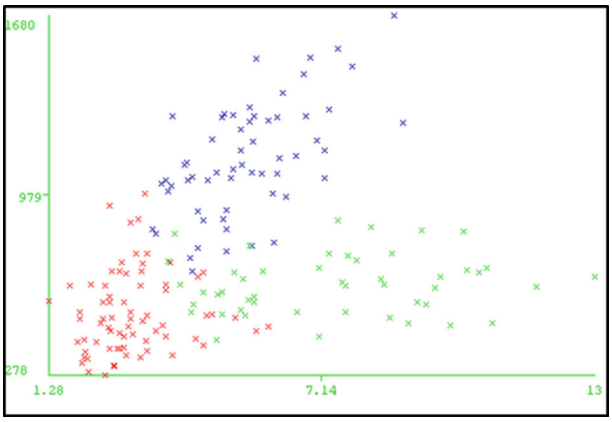

Fig. 5 WINE dataset: "Proline-Color intensity" decision plan generated by J48 - Proline is on Y axis on this and next figures

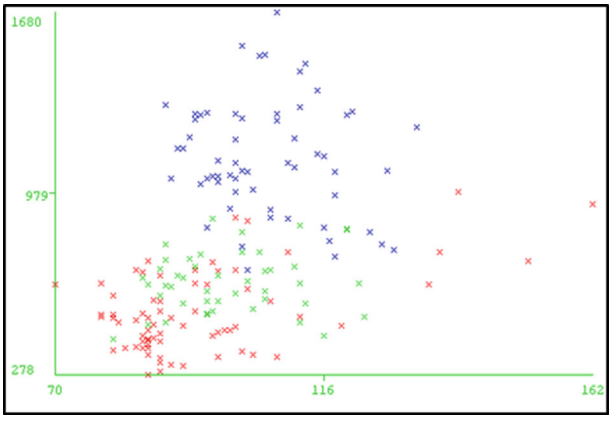

Fig. 6 WINE dataset: "Proline-Magnesium" decision plan generated by FMC (before data contrasting)

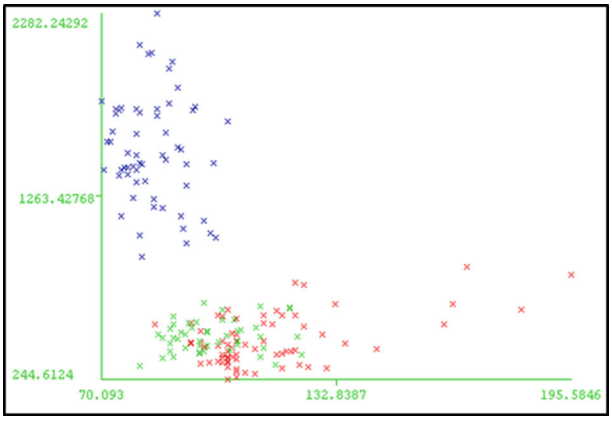

Fig. 7 WINE dataset: "Proline-Magnesium" decision plan generated by FMC (after data contrasting with a magnification factor k = 1)

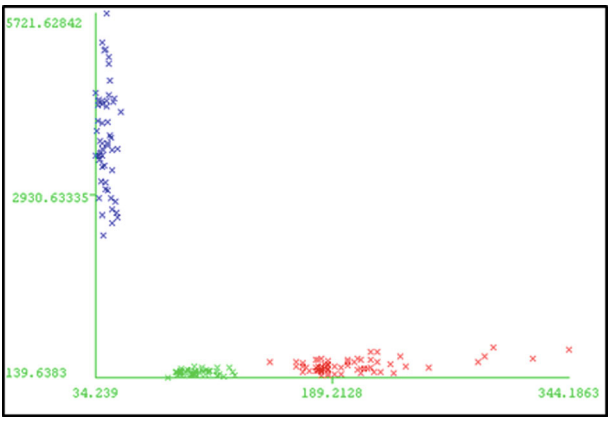

Fig. 8 WINE dataset: "Proline-Magnesium" decision plan generated by FMC (after data contrasting with a magnification factor k = 4)

degrading the results. It thus represents a good alternative to cope with the parameter setting problem.

The 10-highest contrasted features (stemmed forms) of the 8 classes issued from the Reuter8 dataset are presented in Table 8. The fact that mainlines of each topic can be clearly highlighted in such a way illustrate the complementary topic extraction capabilities of the FMC method.

Finally, obtaining very good performance by combining the FMC feature selection approach with a classification method like MNB is a real advantage for large scale exploitation, knowing that MNB method has natural incremental capabilities and that both methods have low computation time.

Complementary results obtained with the numerical UCI Wine dataset interestingly show that, with the help of FMC, NB/BN methods are able to exploit only two features (among 13) for classification as a decision tree classifier like J48 (i.e. C4.5 Meja-Lavalle et al. 2006) would do on standard data. The difference is that a perfect result is obtained with NB/BN and FMC whereas it is not the case with J48 (Table 9). Some explanations are provided by looking up at the distribution of the class samples on the alternative decision plans of the two methods. In the "Proline-Color intensity" decision plan exploited by J48, the different classes are not clearly discriminable (Fig. 5). On its own side, the FMC method "apparently" generates an even more complex "Proline-magnesium" decision plan, if contrast is not considered (Fig. 6). However, as shown in Figs. 7, 8, with the combined effect of contrast and magnification factor (4) on data features, the different classes become very clearly discriminable on that decision plan, especially when the magnification factor is increased sufficiently (Fig. 8).

6 Conclusion

Our main goal was to build up an efficient feature selection and feature contrasting model that could overcome the usual problems arising in the supervised classification of large volume of textual data. These problems relate to classes imbalance, high dimensionality, noise, and high degree of similarity between classes. For that purpose we have proposed to adapt

a recent metric to the context of supervised classification. Through various experiments on large textual datasets we have illustrated many advantages of our approach, and especially is high efficiency for enhancing the performance of the classifiers in such context, whilst putting the focus on the more flexible and the less computationally intensive classifiers, like Bayes classifiers.

Another main advantage of this technique is that it is a parameter-free approach which can rely on basic feature extraction scheme and it can thus be used in larger scopes, like in the ones of incremental and semi-supervised learning. Another interesting perspective would be to adapt this technique in text mining context for enriching ontologies and lexicons through the large scale exploitation of existing corpora.

Feature maximization metric is easily adaptable to a fuzzy classification context in which data can belong to several classes. Hence, such kind of extension only implyies a simple renormalization of the feature recall (FR) measure. Moreover, successull exploitation of the feature maximization principle to hierarchical clustering (Lamirel and Ta 2008) lead to consider that its adaptation to the context of hierarchical classification is also a straightforward process.

Additionally, our feature selection and contrasting technique can easily extent its application range to the broader context of numerical data.

Acknowledgments This work was done under the program QUAERO[11] supported by OSEO[12] French national agency of research development.

References

Aha, D., & Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning*, *6*, 37–66.

Attik, M., Lamirel, J.-C., Al Shehabi, S. (2006). Clustering analysis for data with multiple labels. In *Proceedings of the IASTED international conference on databases and applications (DBA)*. Innsbruck.

Bache, K., & Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A. (2012). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 1–37.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and regression trees*. Belmont: Wadsworth International Group.

Chawla, N.V., Bowyer, K.V., Hall, L.O., Kegelmeyer, W.P. (2002). Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. *Artificial Intelligence*, *151*(1), 155–176.

Daviet, H. (2009). Class-Add, une procédure de sélection de variables basée sur une troncature k-additive de l' information mutuelle et sur une classification ascendante hiérarchique en pré-traitement. PhD, Université de Nantes, France.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, *39*(1), 1–38.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, *3*, 1289–1305.

Good, P. (2006). *Resampling methods*, 3rd edn. Birkhauser.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1), 389–422.

[11] http://www.quaero.org

[12] http://www.oseo.fr/

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157–1182.

Hall, M.A., & Smith, L.A. (1999). Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In *Proceedings of the 12th international florida artificial intelligence research society conference* (pp. 235–239). AAAI Press.

Hajlaoui, K., Cuxac, P., Lamirel, J.C., Francois, C. (2012). Enhancing patent expertise through automatic matching with scientific papers. *Discovery Science LNCS*, *7569*, 299–312.

Ken Lang, K. (1995). Learning to filter netnews. In *Proceedings of the 12th international conference on machine learning* (pp. 331–339).

Kohavi, R., & John, G.H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273–324.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. In *European conference on machine learning* (pp. 171–182).

Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering*, *3*(5), 1787–1797.

Lallich, S., & Rakotomalala, R. (2000). Fast feature selection using partial correlation for multi-valued attributes. In D.A. Zighed, J. Komorowski, J. Zytkow (Eds.), *Principles of data mining and knowledge discovery. Lecture notes in computer science, 1910* (pp. 221–231). Berlin-Heidelberg: Springer.

Lamirel, J.-C., Al Shehabi, S., Francois, C., Hoffmann, M. (2004). New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping. *Scientometrics*, *60*(3).

Lamirel, J.-C., & Ta, A.P. (2008). Combination of hyperbolic visualization and graph-based approach for organizing data analysis results: an application to social network analysis. In *Proceedings of the 4th international conference on webometrics, informetrics and scientometrics and 9th COLLNET meeting*. Berlin.

Lamirel, J.-C., Ghribi, M., Cuxac, P. (2010). Unsupervised recall and precision measures: a step towards new efficient clustering quality indexes. In *Proceedings of the 19th international conference on computational statistics (COMPSTAT'2010)*. Paris.

Lamirel, J.-C., Mall, R., Cuxac, P., Safi, G. (2011). Variations to incremental growing neural gas algorithm based on label maximization. In *Proceedings of IJCNN 2011*. San Jose.

Lamirel, J.-C. (2012). A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research. *Scientometrics*, *93*, 151–166.

Mejía-Lavalle, M., Sucar, E., Arroyo, G. (2006). Feature selection with a perceptron neural net. Feature selection for data mining: interfacing machine learning and statistics.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(11), 559–572.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, A. Smola (Eds.), *Advances in kernel methods - support vector learning*. MIT Press.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Quinlan, R. (1993). *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann.

Salton, G. (1971). *Automatic processing of foreign language documents*. Englewood Cliffs: Prentice-Hill.

Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*(5), 513–523.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*.

Witten, I.H., & Frank, E. (2005). Data mining: practical machine learning tools and techniques. Morgan Kaufmann.

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: a fast correlation-based filter solution. ICML 2003, 856–863. Washington.

Zhang, T., & Oles, F.J. (2001). Text categorization based on regularized linear classification methods. *Information Retrieval*, *4*(1), 5–31.