New Metrics and Related Statistical Approaches for Efficient Mining in Very Large and Highly Multidimensional Databases

Jean-Charles Lamirel[1,2(✉)]

[1] Department of Computer Science, University of Tartu,
J. Liivi 2, 50409 Tartu, Estonia
jean-charles.lamirel@ut.ee
[2] LORIA, Equipe Synalp, Bâtiment B,
54506, Vandoeuvre Cedex, France
lamirel@loria.fr

Abstract. As regard to the evolution of the concept of text and to the continuous growth of textual information of multiple nature which is available online, one of the important issues for linguists and information analysts for building up assumptions and validating models is to exploit efficient tools for textual analysis, able to adapt to large volumes of heterogeneous data, often changing and of distributed nature. We propose in this communication to look at new statistical methods that fit into this framework but that can also extent their application range to the more general context of dynamic numerical data.

For that purpose, we have recently proposed an alternative metric based on feature maximization. The principle of this metric is to define a measure of compromise between generality and discrimination based altogether on the properties of the data which are specific to each group of a partition and on those which are shared between groups. One of the key advantages of this method is that it is operational in an incremental mode both on clustering (i.e. unsupervised classification) and on traditional categorization. We have shown that it allowed to very efficiently solve complex multidimensional problems related to unsupervised analysis of textual or linguistic data, like topic tracking with data changing over time or automatic classification in natural language processing (NLP) context. It can also adapt to the traditional discriminant analysis, often exploited in text mining, or to automatic text indexing or summarization, with performance that are far superior to conventional methods. In a more general way, this approach that freed from the exploitation of parameters can be exploited as an accurate feature selection and data resampling method in any numerical or non numerical context.

We will present the general principles of feature maximization and we will especially return to its successful applications in the supervised framework, comparing its performance with those of the state of the art methods on reference databases.

Keywords: Classification · Feature selection · Resampling · Clustering · Big data

S. Kozielski et al. (Eds.): BDAS 2015, CCIS 521, pp. 3–20, 2015.
DOI: 10.1007/978-3-319-18422-7_1

1 Introduction

Since the 1990s, progress in computing, and in storage capacities, has allowed the handling of extremely large volumes of data: it is not rare to deal with space for the description of several thousand, or even tens of thousands, features. It could be thought that the classification algorithms are more effective with a large number of features, but the situation is not so simple. The first problem is the increase in the calculation time. Additionally, the fact that a large number of features are redundant, or irrelevant, for the classification task, considerably disrupts the functioning of the classifiers. Furthermore, most training algorithms use probabilities whose distributions may be difficult to estimate in the presence of a very large number of features. The integration of a process of feature selection in the frame of large dimension data classification has thus become a central issue. In the literature, essentially three types of approach are proposed for the selection of features: approaches directly incorporated into the classification methods, known as "embedded", methods based on techniques of optimization, or "wrapper", and approaches based on statistical tests, also named filter-based methods. Thorough states-of-the art have been described by numerous authors, such as Ladha et al. [21,3,13] ou [8]. Therefore, below we will simply give a brief overview of the existing approaches.

"Embedded" approaches integrate the selection of features in the learning process [5]. The most popular methods in this category are those based on SVM and the methods founded on neural networks. For example, RFE-SVM (Recursive Feature Elimination for Support Vector Machines) [14] is an integrated process, where the selection of features is carried out in an iterative manner using an SVM classifier and suppressing features that are the most distant from the decision boundary.

For their part, the "wrapper" methods use a performance criterion to seek out a pertinent sub-group of predictors [20]. Most often it is the error rate (but that can be a prediction cost, or the area under the ROC curve). As an example, the WrapperSubsetEval method begins with an empty set of features, and continues until the addition of new features no longer improves performance. It uses cross-validation to estimate learning for a given group of features [39]. Comparisons between methods, such as that of Forman [10] clearly demonstrate that, without taking their effectiveness into account, one of the principal drawbacks of these two classes of methods is that they require long calculation times. This prohibits their use in the case of strongly multidimensional data. In this context, a possible alternative is to exploit filter-based methods.

Filter-based approaches are selection methods that are used upstream and independently of the learning algorithm. Based on statistical tests, they require less calculation time than do other approaches. Most classical examples of filter-based methods are chi-squared method [21], mutual information-based methods, like MIFS [16], information gain-based methods, like CBS [7], correlation-based methods, like MODTREE [22], or, nearest-neighbour-based methods, like Relief or ReliefF [19].

As for all statistical tests, filter-based approaches are known to behave erratically in the case of very low frequency features, which are common in text classification [21]. In this article we show that, despite their diversity, all existing approaches are inoperative, or even detrimental, in the case of extremely unstable, multidimensional and noisy data, with a high degree of similitude between classes. As an alternative, we propose a new method of feature selection and contrast, based on the recently developed feature maximization metric. Furthermore, we compare the performance of this method to that of classical techniques in the context of help with patent validation. Then we extend the range of our study to habitually used textual reference data. The rest of this manuscript is structured as follows: section 2 presents our new approach for feature selection; section 3 details the data used; section 4 compares the results for the different data corpora of the classification with and without the use of the proposed approach; section 5 outlines the use of the method in unsupervised context; section 6 presents our conclusions and perspectives.

2 Feature Maximization for Feature Selection

Feature maximization (F-max) is an unbiased metric with whichto estimate the quality of an unsupervised classification, which uses the properties (i.e. the features) of data associated with each cluster without prior examination of the cluster profiles [24]. Its principal advantage is that it is totally independent of the classification method and of its operating mode. When it is used after learning, it can be exploited to establish global indices of clustering quality [26] or for cluster labelling [28].

Consider a group of clusters C which results from a method of clustering applied to a dataset D represented by a group of features F. The feature maximization metric favours clusters with a maximal feature F-measure. The feature F-measure $FF_c(f)$ of a feature f associated with a cluster c c is defined as the harmonic mean of the feature recall $FR_c(f)$ and of the feature precision $FP_c(f)$, themselves defined as follows:

$$FR_c(f) = \frac{\Sigma_{d \in c'} W_d^f}{\Sigma_{c' \in C} \Sigma_{d \in c'} W_d^f} \quad FP_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{f' \in F_c, d \in c} W_d^{f'}} \tag{1}$$

with

$$FF_c(f) = 2 \left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)} \right) \tag{2}$$

where W_d^f represents the weight of the feature f for the data d and F_c represents all the features present in the dataset associated with the cluster C.

Taking into account the basic definition of the feature maximization metric, its use for the task of feature selection in the context of supervised learning becomes a simple process. Therefore, this generic metric can be applied to data

associated with a class, as well as those associated with a cluster. The selection process can thus be defined as non-parametered, based on classes in which a class feature is characterised using both its capacity to discriminate between classes ($FP_c(f)$ index) and its ability to faithfully represent the class data ($FR_c(f)$ index). The Sc set of features that are characteristic of a given class c belonging to the group of classes C is translated by:

$$S_c = \left\{ f \in F_c \mid FF_c(f) > \overline{FF}(f) \text{ and } FF_c(f) > \overline{FF}_D \right\} \text{ where} \tag{3}$$

$$\overline{FF}(f) = \Sigma_{c' \in C} \frac{FF_{c'(f)}}{|C_{/f}|} \text{ and } \overline{FF}_D = \Sigma_{f \in F} \frac{\overline{FF}(f)}{|F|} \tag{4}$$

where $C_{/f}$ represents the subset of C in which the f feature is represented.

Finally, the set of all selected features S_C is the subset of F defined by:

$$S_C = \cup_{c \in C} S_C. \tag{5}$$

In other words, the features that are judged relevant for a given class are those whose representations are better than average in this class, and better than the average representation of all the features in terms of feature F-measure.

In the specific context of the process of feature maximization, an improvement by contrast step can be exploited as a complement to the first step of selection. The role of this is to adapt the description of each single data to the specific characteristics of its associated class. This consists of modifying the data weighting schema in a distinct way for each class, taking into account the information gain supplied by feature F-measure of the features locally in this class.

The information gain is proportional to the relation between the F-measure value of a feature in the $FF_c(f)$ class and the average F-measure value of this feature for the whole partition. Given one single data and one single feature describing this data, the resulting gain acts as a contrast factor that adjusts the weight of this feature in the data profile, optionally taking into account its prior establishment. For a feature f belonging to the group of selected features S_c from a class C, the gain $G_c(f)$ is expressed as:

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

where k is a magnification factor that can be optimized according to the resulting accuracy.

The active features of a class are those for which the information gain is higher than 1. Given that the proposed method is one of selection and of contrast based on the classes, the average number of active features per class is comparable to the total number of features selected in the case of habitual selection methods.

3 Validating the Approach on Real-World Data

One of the goals of the QUAERO project is to use bibliographic information to help experts to judge patent precedence. Thus, initially it was necessary to prove that it is possible to associate such information with the patent classes in a pertinent manner; or in other words, to classify it correctly within these classes. Main experimental data source comprised 6387 patents from the pharmacological domain in an XML format, grouped into 15 sub-classes of the A61K class (medical preparation). The bibliographic references in the patents were extracted from the Medline database[1]. 25887 citations were extracted from the 6387 patents. Interrogation of the Medline database with the extracted citations allowed bibliographic notices of 7501 references to be recovered. Each notice was then labelled with the first classment code of the citing patent [15]. Each notice's abstract was treated and transformed into a bag of words [36] using the TreeTagger tool [37]. To reduce the noise generated by this tool, a frequency threshold of 45 (i.e. an average threshold of 3 per class) was applied to the extracted descriptors. The result was a description space limited to the 1804 dimension. A last TF-IDF weighting step was applied [36]. The series of labelled notices, which were thus pre-treated, represented the final corpus on which training was carried out. This last corpus was highly unbalanced. The smallest class (A61K41) contained 22 articles, whereas the largest contained 2500 (A61K31 class). The inter-class similarity was calculated using a cosine correlation. This indicated that more than 70% of pairs of classes had a similarity of between 0.5 and 0.9. Thus, the ability of a classification model to precisely detect the correct class is strongly reduced. A solution commonly used to contend with an imbalance in classes' data is sub-sampling of the larger classes [12] and/or over-sampling of the smaller ones [6]. However, re-sampling, which introduced redundancy into the data, does not improve the performance of this dataset, as was shown by Hajlaoui et al. (2012). Therefore, we have proposed an alternative solution detailed below, namely to edit out the features that are judged irrelevant and to contrast those considered reliable [25].

As a complement, 4 other well-known reference text datasets have been exploited for validation of the method:

- The R8 and R52 corpora were obtained by Cardoso Cachopo[2] from the R10 and R90 datasets, which are derived from the Reuters 21578 collection[3]. The aim of these adjustments was to only retain data that had a single label. Considering only monothematic documents and classes that still had at least one example of training and one of test, R8 is a reduction of the R10 corpus (the 10 most frequent classes) to 8 classes and R52 is a reduction of the R90 corpus (90 classes) to 52 classes.
- The Amazon^{tm} corpus (AMZ) is a UCI dataset [2] derived from the recommendations of clients of the Amazon web site that are usable for author

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://web.ist.utl.pt/~acardoso/datasets/
[3] http://www.research.att.com/~lewis/reuters21578.html

identification. To evaluate the robustness of the classification algorithms with respect to a large number of target classes, 50 of the most active users who have frequently posted comments in these newsgroups were identified. Thirty messages were collected for each of them. Each message included the authors' linguistic style, such as the use of figures, punctuation, frequent words and sentences.

- The 20Newsgroups dataset [19] is a collection of approximately 20,000 documents (almost) uniformly distributed among 20 different discussion groups. We consider two "bag of words" versions of this dataset in our experiments. In the (20N - AT) version, all words are preserved and non-alphabetic characters are converted into spaces. It resulted in a 11153 words description space. The (20N - ST) version is obtained after a additionnal step of stemming.The words of less than 2 characters, as well as stopwords (S24 SMART list [36]), are eliminated. The stemming is performed using Porter's algorithm [33]. The description space is thus reduced to 5473 words.
- The WebKB dataset (WKB) contains 8282 pages collected from the departments of computer science of various universities in January 1997 by the World WideKnowledge Base, a project of the CMU text learning group[4] (Carnegie Mellon University, Pittsburgh). The pages have been manually divided into 7 classes: student, faculty, department, course, personal, project, other. We operate on the Cardoso Cachopo's reduced version in which classes "department" and "staff" were rejected due to their low number of pages, and the class "other" has been deleted. Cleaning and stemming methods used for the 20Newsgroups dataset are then applied on the reduced dataset. It resulted in a 4158 items dataset described by a 1805 words description space.

4 Experiments and Results

4.1 Experiments

To carry out our experiments, we first took into consideration different classification algorithms that are implemented in the Weka tool box[5]: decision trees (J48) [35], random forests (RF)[4], KNN [1], habitual Bayesian algorithms, i.e. the Multinomial Nave Bayes (MNB) and Bayesian Network (BN), and finally, the SMO-SVM algorithm (SMO) [32]. The default parameters were used during the implementation of these algorithms, apart from KNN for which the number of neighbours was optimized based on the resulting precision. Secondly, we placed the accent more particularly on tests of the efficacy of feature selection approaches, including our new proposition (FMC). In our test, we included a panel of filter-based approaches applicable on large dimension data, using once again the Weka platform. The methods tested include: chi-squared [21], information gain [16], CBF [7], symmetric incertitude [40], ReliefF [19] (RLF), Principal

[4] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/
[5] http://www.cs.waikato.ac.nz/ml/weka/

Component Analysis [31] (PCA). Default parameters were used for most of these methods except for PCA, where the explained variance percentage is tuned with respect to the resulting accuracy. Initially we tested the methods separately. In a second phase, we combined the feature selection supplied by the different methods with the F-max contrast method that we have proposed (eq. 6). We used a 10-fold cross-validation in all our experiments.

4.2 Results

The different results are presented in tables 1 to 8. They are based on measurements of standard performance (level of true positives [TP] or recall [R], level of false positives [FP], Precision [P], F-measure [F] and ROC) weighted by class size, then averaged for all the classes. For each table and each combination of selection and classification methods, an indicator of performance gain/loss (TP Incr) is calculated using the TP of SMO level on original data as a reference. Finally, as the results for chi-squared, information gain and symmetric incertitude

Table 1. Classification results on initial data

	TP(R)	FP	P	F	ROC	TP Incr
J48	0.42	0.16	0.40	0.40	0.63	-23%
RandomForest	0.45	0.23	0.46	0.38	0.72	-17%
SMO	**0.54**	**0.14**	**0.53**	**0.52**	**0.80**	**0% (Ref)**
BN	0.48	0.14	0.47	0.47	0.78	-10%
MNB	0.53	0.18	0.54	0.47	0.85	-2%
KNN (k=3)	0.53	0.16	0.53	0.51	0.77	-2%

Table 2. Results of classification after the selection of features (BN classifier)

	TP(R)	FP	P	F	ROC	Nbr. var.	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**	**0.003**	**0.99**	**0.99**	**1**	**262/cl**	**+90%**

Table 3. Results of classification after the selection of FMC features

	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**	**0.003**	**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%

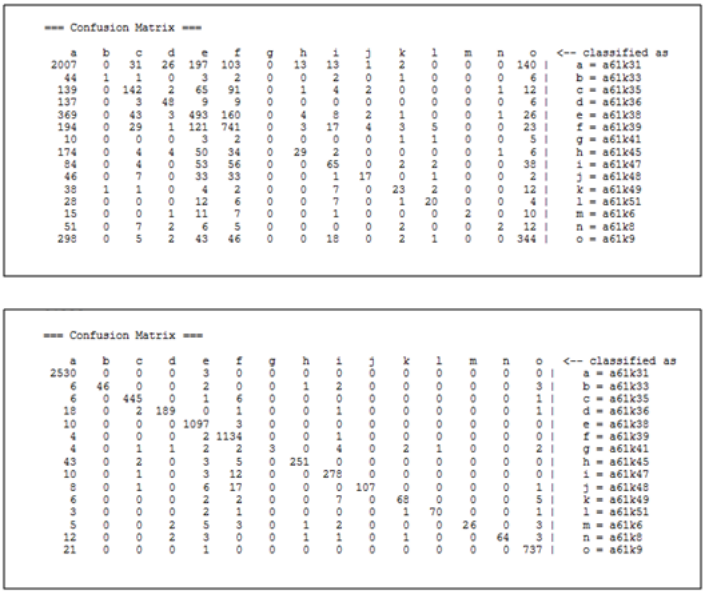
were identical, they only figure once in the tables, as results of the chi-squared type (and are noted CHI+).

For our main patent collection, table 1 shows that the performances of all classification methods are weak for the dataset considered, provided no feature selection process is carried out. In this context, this table also confirms the superiority of the SMO, KNN and Bayesian methods compared to the other two methods, based on decision trees. Additionally, SMO gave the best global performance in terms of discrimination, as demonstrated by its highest ROC value. However, this method is clearly not usable in an operational context of patent evaluation such as QUAERO, because of the major confusion between classes. This shows its intrinsic inability to cope with the attraction effect of the largest classes. Each time that a standard feature selection method is applied in our context, in association with the best classification methods, its use alters the quality of the results slightly, as indicated in table 2. Table 2 also underlines the fact that the reduction in the number of feature by the FMC method is similar to CHI+ (in terms of active features; see section 2 for more details), but that its use stimulates the performances of classification methods, particularly those of Bayesian methods (table 3), leading to impressive classification results in the context of highly complex classification: 0.987 accuracy i.e. only 94 misclassed data with the BN method, amongst a total of 7252.

The results presented in table 4 illustrate more precisely the efficiency of the F-max contrast method that acts on data description (eq. 6). In experiments relating to this table, the contrast is applied individually to the features extracted by each selection method, and in a second step a BN classifier is applied to the resulting contrasted data. The results show that, irrespective of the type of method used for feature selection, the performances of the resulting classification are re-enforced each time that the F-max contrast is applied downstream of the selection. The average performance increase is 44%. Finally, table 5 illustrates the ability of the FMC approach to efficiently confront the problems of imbalance and class similitude. The examination of TP level variations (especially in the small classes) seen in this Table shows that the attraction effect of data from the largest classes, produced at a high level in the case of the use of original data, is practically systematically overcome each time the FMC approach is exploited. The ability of this approach to correct class imbalance is equally clearly demonstrated by the homogeneous distribution of active features in the different classes, despite the extremely heterogeneous class size.

Table 4. Results of classification with different feature selection methods, and F-max contrast (BN classifier)

	TP(R)	FP	P	F	ROC	Nbr. var.	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**	**0.003**	**0.99**	**0.99**	**1**	**262/cl**	**+90%**



```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j    k    l    m    n    o   <-- classified as
 2007    0   31   26  197  103    0   13   13    1    2    0    0    0  140 |    a = a61k31
   44    1    1    0    3    2    0    0    2    0    1    0    0    0    6 |    b = a61k33
  139    0  142    2   65   91    0    1    4    2    0    0    0    1   12 |    c = a61k35
  137    0    3   48    9    9    0    0    0    0    0    0    0    0    6 |    d = a61k36
  369    0   43    3  493  160    0    4    8    2    1    0    0    1   26 |    e = a61k38
  194    0   29    1  121  741    0    3   17    4    3    5    0    0   23 |    f = a61k39
   10    0    0    0    3    2    0    0    0    0    1    1    0    0    5 |    g = a61k41
  174    0    4    4   50   34    0   29    2    0    0    0    0    1    6 |    h = a61k45
   84    0    4    0   53   56    0    0   65    0    2    2    0    0   38 |    i = a61k47
   46    0    7    0   33   33    0    0    1   17    0    1    0    0    2 |    j = a61k48
   38    1    1    0    4    2    0    0    7    0   23    2    0    0   12 |    k = a61k49
   28    0    0    0   12    6    0    0    7    0    1   20    0    0    4 |    l = a61k51
   15    0    0    1   11    7    0    0    1    0    0    0    2    0   10 |    m = a61k6
   51    0    7    2    6    5    0    0    0    0    2    0    0    2   12 |    n = a61k8
  298    0    5    2   43   46    0    0   18    0    2    1    0    0  344 |    o = a61k9
```

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j    k    l    m    n    o   <-- classified as
 2530    0    0    0    3    0    0    0    0    0    0    0    0    0    0 |    a = a61k31
    6   46    0    0    2    0    0    1    2    0    0    0    0    0    3 |    b = a61k33
    6    0  445    0    1    6    0    0    0    0    0    0    0    0    1 |    c = a61k35
   18    0    2  189    0    1    0    0    1    0    0    0    0    0    1 |    d = a61k36
   10    0    0    0 1097    3    0    0    0    0    0    0    0    0    0 |    e = a61k38
    4    0    0    0    2 1134    0    0    1    0    0    0    0    0    0 |    f = a61k39
    4    0    1    1    2    2    3    0    4    0    2    1    0    0    2 |    g = a61k41
   43    0    2    0    3    5    0  251    0    0    0    0    0    0    0 |    h = a61k45
   10    0    1    0    3   12    0    0  278    0    0    0    0    0    0 |    i = a61k47
    8    0    1    0    6   17    0    0    0  107    0    0    0    0    1 |    j = a61k48
    6    0    0    0    2    2    0    0    7    0   68    0    0    0    5 |    k = a61k49
    3    0    0    0    2    1    0    0    0    0    1   70    0    0    1 |    l = a61k51
    5    0    0    2    5    3    0    1    2    0    0    0   26    0    3 |    m = a61k6
   12    0    0    2    3    0    0    1    1    0    1    0    0   64    3 |    n = a61k8
   21    0    0    0    1    0    0    0    0    0    0    0    0    0  737 |    o = a61k9
```

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

Table 5. Characteristics/class before and after FMC selection (BN classifier)

Class Label	Size	Feat. Select.	% TP FMC	% TP 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

The summary of the results of the four complementary datasets is presented in tables 6 to 8. These tables highlight the fact that the FMC method can very significantly improve the performance of the classifiers in different types of cases. As in the context of our previous experience (patents), the best performances are obtained with the use of the FMC method in combination with the MNB

Table 6. List of high contrast features (lemmes) for the 8 classes of the REUTERS8 corpus

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
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

Table 7. Results of classifications after FMC feature selection (MNB/BN classifier)

		TP (R)	FP	P	F	ROC	TP Incr.
	-	0.937	0.02	0.942	0.938	0.984	
Reuters8 (R8)	**FMC**	**0.998**	**0.001**	**0.998**	**0.998**	**1**	**+6%**
	-	0.91	0.01	0.909	0.903	0.985	
Reuters52 (R52)	**FMC**	**0.99**	**0.001**	**0.99**	**0.99**	**0.999**	**+10%**
	-	0.748	0.05	0.782	0.748	0.981	
Amazon	**FMC**	**0.998**	**0.001**	**0.998**	**0.998**	**1**	**+33%**
	-	0.882	0.006	0.884	0.881	0.988	
20NewsGroup-AT (all terms)	**FMC**	**0.992**	**0**	**0.992**	**0.1**	**1**	**+13%**
	-	0.865	0.007	0.866	0.864	0.987	
20NewsGroup-ST (stemmed)	**FMC**	**0.991**	**0.001**	**0.991**	**1**	**1**	**+15%**
	-	0.842	0.068	0.841	0.841	0.946	
WebKB	**FMC**	**0.996**	**0.002**	**0.996**	**0.996**	**0.996**	**+18%**

Table 8. 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

C1- 7(7) [315(315)]

Prevalent Label --- = Cause-Experiencer

0.273245 G-Cause-Experiencer
0.173498 C-SUJ:Ssub,OBJ:NP
0.138411 C-SUJ:NP,DEOBJ:PP
0.091732 C-SUJ:NP,DEOBJ:PP,DUMMY:REFL
...
****
****
0.013839 T-Asset
0.013200 C-SUJ:NP,DEOBJ:Ssub,POBJ:PP
0.009319 C-SUJ:Ssub,OBJ:NP,POBJ:PP
...
[flatter 0.907200 3(1)] [charmer 0.889490 3(0)] [exulter 0.889490 3(0)] [**frissonner 0.889490 3(0)] [mortifier 0.889490 3(0)] [époustoufler 0.889490 3(0)] [pâtir 0.889490 3(0)] [ravir 0.889490 3(0)] [**trembler 0.889490 3(0)] [**trembloter 0.889490 3(0)] [décourager 0.872350 2(2)]...

Fig. 2. Sample output for a French verb cluster produced with the IGNGF clustering method. The exploited features represent either verb subcategorization frames or semantic labels.

and BN Bayesian classifiers. Table 7 presents the comparative results of such a combination. It demonstrates that the FMC method is particularly effective in increasing the performance of the classifiers when the complexity of the classification task becomes higher because of an increasing number of classes (AMZ corpus). Table 8 supplies general information about the data and the behaviour of the FMC selection method. They illustrate the significant reduction in the classification complexity obtained with FMC because of the drop in the number of features to manage, as well as the concomitant decrease of badly classed data. It also stresses the calculation time, which is highly curbed for this method (the calculation is carried out on Linux using a laptop computer equipped with an Intel® Pentium® B970 2.3Ghz processor and 8Go of memory).

For these datasets, similar remarks to those mentioned for the patent dataset can be made on the subject of the low efficiency of common feature selection methods and the re-sampling methods. Table 8 also shows that the value of the contrast magnification factor utilised to obtain the best performances can vary throughout the experiments (from 1 to 4 in this last context). However, it can be observed that by taking a fixed value for this factor, for example the highest (here 4), the results are not down-graded. This choice thus represents a good alternative to confront the problem of configuration.

The 5 most contrasted feature (lemmes) of the 8 classes issued from the Reuter8 corpus are shown in table 6. The fact that the main lines of the themes covered by the classes can be clearly demonstrated in this way illustrates the

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	0% (Ref)
BN + FMC	**1**	**0**	**1**	**1**	**1**	**+6%**

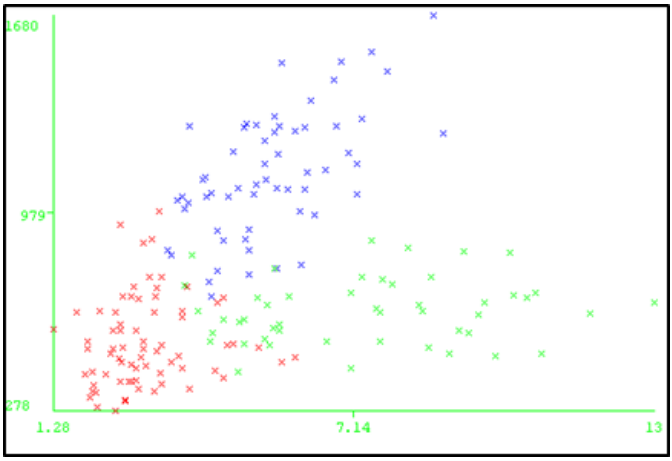


Fig. 3. WINE dataset: "Proline-Color intensity" decision plan generated by J48 - Proline is on Y axis on this and next figures

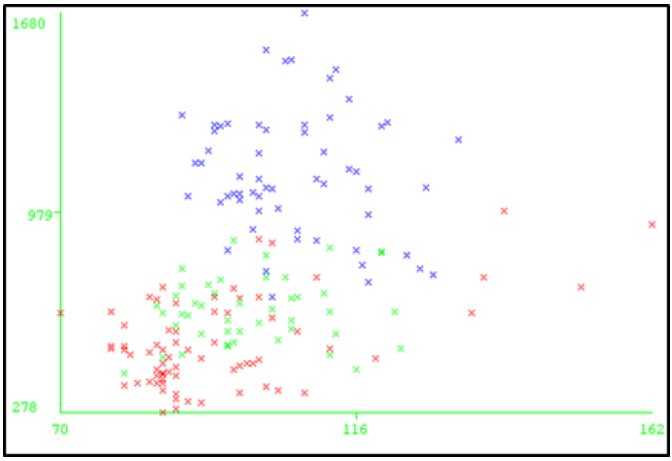


Fig. 4. WINE dataset: "Proline-Magnesium" decision plan generated by FMC (before data contrasting)

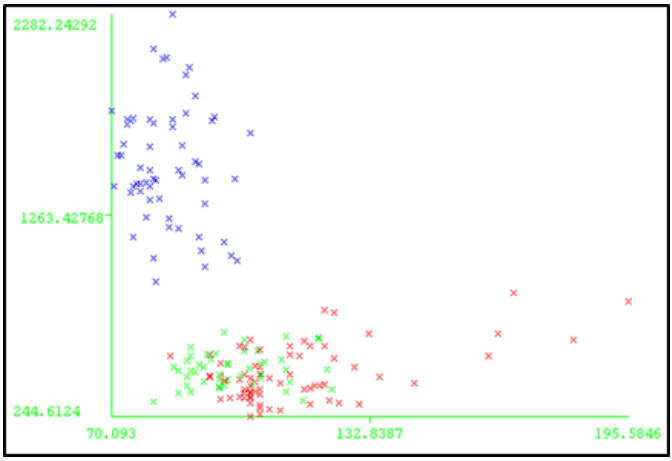


Fig. 5. WINE dataset: "Proline-Magnesium" decision plan generated by FMC (after data contrasting with a magnification factor k=1)

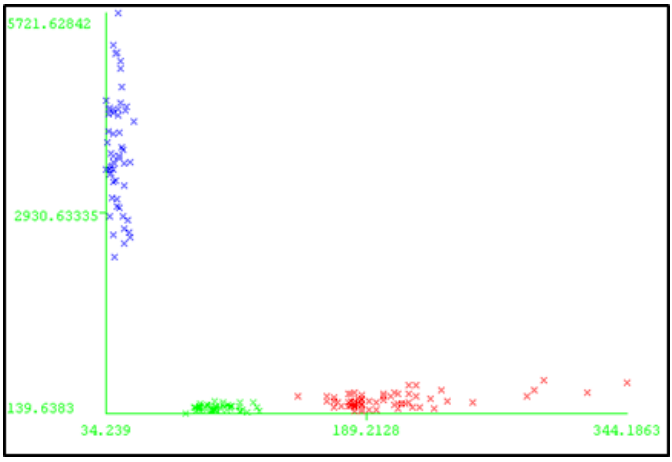


Fig. 6. WINE dataset: "Proline-Magnesium" decision plan generated by FMC (after data contrasting with a magnification factor k=4)

topic extraction capacities by the FMC method. Finally, the acquisition of very good performances by combining the FMC feature selection and constrast with a classification method such as MNB is a real advantage for large-scale usage, given that the MNB method has incremental abilities and that the two methods have low calculation times.

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 [27]) 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. 3). On its own side, the FMC method "apparently" generates an even more complex "Proline-Magnesium" decision plan, if contrast is not considered (Fig. 4). However, as shown in Fig. 5- 6, 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. 6).

5 Feature Maximization for Clustering

Like other neural free topology learning methods such as Neural Gas (NG) [30], Growing Neural Gas (GNG) [11], or Incremental Growing Neural Gas (IGNG) [34], the IGNGF method makes use of Hebbian learning [17] for dynamically structuring the learning space. Hebbian learning is inspired by a theory from neurosciences which explains how neurons connect to build neural networks. Whereas for NG the number of output prototypes is fixed, GNG adapts this number during the learning phase, guided by the characteristics of the data to be classified. Prototypes and connections between them can be created or removed depending on evolving characteristics of learning (as for example the "age" or "maturity" of connections and the cumulated error rate of each prototype). A drawback of this approach is that prototypes are created or removed after a fixed number of iterations yielding results which might not appropriately represent complex or sparse multidimensional data. With the IGNG clustering method this issue is addressed by allowing more flexibility when creating new prototypes: a prototype is added whenever the distance of a new data point to an existing prototype is above a predefined global threshold, the average distance of all the data points to the centre of the data set. The learning process thus becomes incremental: each incoming data point is considered as a potential prototype. For all the above-mentioned methods, at each iteration over all the data points, a data point is connected with the "closest" prototypes and at the same time interacts with the existing model by strengthening the connections between these "closest" prototypes and weakening those to other, less related prototypes. Because of these dynamically changing interactions between prototypes, these methods are "winner take most" methods in contrast to K-means (for example), which represents a "winner-take-all" method. The notion of "closeness" is based on a distance function computed from the features associated to the data points.

IGNGF uses the Hebbian learning process as IGNG, but the use of a standard distance measure as adopted in IGNG for determining the "closest" prototype is replaced in IGNGF by feature maximization.

With feature maximization, the clustering process is roughly the following. During learning, an incoming data point x is temporary added to every existing cluster, its feature profile constituted by its maximal features and its average feature F-measure are computed. Then the winning prototype is the prototype whose associated cluster maximises the κ criterion given in Equation (7),

$$\kappa(c) = \Delta(FF_c) * |F_c \cap F_x| - \frac{\|p_c, x\|}{weight} \tag{7}$$

where $\Delta(FF_c)$ represents the gain in feature F-measure for the new cluster, $|F_c \cap F_x|$ represents the number of features shared by cluster c and the data point x and p_c is the codebook vector of the prototype associated to cluster c. This way, those clusters are preferred which share more features with the new data point and clusters which don't have any common feature with the data point are ignored. The gain in feature F-measure multiplied by the number of shared features is adjusted by the euclidean distance of the new data point x to the cluster's prototype codebook vector p_c. Thus, the smaller the euclidean distance to the cluster's prototype, less the κ value decreases. The influence of the euclidean distance can be parametrised with a *weight* factor ($\sqrt{2}$ for usual application). Clusters with negative κ score are ignored. The data point is then added to the cluster c with maximal $\kappa(c)$ and the connections between its associated prototype and the neighbour prototypes are updated. If κ value is negative for all clusters, a new prototype is created and an associated cluster is formed with the currently considered data point.

The IGNGF method was shown to outperform other usual neural and non neural methods for clustering tasks on sparse and/or highly multidimensionnal and/or noisy data [27]. Moreover, it can be fruitfully combined with unsupervised Bayesian reasoning for setting up the first parameter-free method capable of automatically tracking research topics evolving over time in a realistic multidimensionnal context [23]. It was also recently shown to outperform supervised classification methods in the context of websites classification task thanks to its capacity to highlight "latent classes" not initially planed by the analyst [29].

Another main advantage of the method is that maximized features used by IGNGF during learning can also be exploited in a final step for accurately labeling the resulting clusters. An example of such results is given in the case of French verb clustering [9,18]. In this specific context, the IGNGF clustering method does not only provides accurate verb clusters, outperforming state-of-the-art methods of the domain, like spectral clustering [38]. As a complementary result, it associates each verb cluster c with a profile containing syntactic and semantic features characteristic of that cluster. Features are displayed in decreasing order of feature F-measure given by Equation (2) and features whose feature F-measure is under the average feature F-measure of the overall clustering are clearly separated from others. In the sample cluster shown in Fig. 2 these are listed above the two star lines. In addition, for each verb in a cluster, a confidence score can be easily computed [9].

6 Conclusion

Our main aim was to develop an efficient method of feature selection and contrast, which would allow routine problems linked to the supervised classification of large volumes of textual data to be overcome. These problems are linked to class imbalance, with a high degree of similarity between them, as they house highly multidimensional and noisy data. To achieve our aim, we adapted a recently developed metric in the unsupervised framework to the context of supervised classification. By means of different experiments on large textual datasets, we illustrated numerous advantages of our approach, including its effectiveness to improve the performance of classifiers in such a context. Notably, this approach places the accent on the most flexible classifiers, and the least demanding in terms of calculation times, such as the Bayesian classifiers. Another advantage of this method is that it concerns an approach without parameters that depends on a simple feature extraction schema. The method can thus be used in numerous contexts, such as those of incremental or semi-supervised learning, and more generally, in large scale digital learning.

References

1. Aha, D., Kibler, D., Albert, M.: Instance-based learning algorithms. Machine Learning 6, 37–66 (1991)
2. Bache, K., Lichman, M.: Uci machine learning repository (http://archive.ics.uci.edu/ml): University of California, school of information and computer science, Irvine, CA, USA (2013)
3. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: A review of feature selection methods on synthetic data. Knowledge and Information Systems 34(3), 483–519 (2013)
4. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
5. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Tech. rep., Wadsworth International Group, Belmont, CA, USA (1984)
6. Chawla, N.V., Bowyer, K.V., Hall, L.O., Kegelmeyer, W.P.: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
7. Dash, M., Liu, H.: Consistency-based search in feature selection. Artificial Intelligence 151(1), 155–176 (2003)
8. Daviet, H.: 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. Thèse de doctorat, Université de Nantes (2009)
9. Falk, I., Gardent, C., Lamirel, J.-C.: Classifying French verbs using French and English lexical resources. In: Proceedings of ACL, Jeju Island, Korea (2012)
10. Forman, G.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305 (2003)
11. Fritzke, B.: A growing neural gas network learns topologies. Advances in Neural Information Processing Systems 7, 625–632 (1995)
12. Good, P.: Resampling methods. Ed. Birkhauser (2006)

13. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
14. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46(1), 389–422 (2002)
15. Hajlaoui, K., Cuxac, P., Lamirel, J.-C., François, C.: Enhancing patent expertise through automatic matching with scientific papers. In: Ganascia, J.-G., Lenca, P., Petit, J.-M. (eds.) DS 2012. LNCS, vol. 7569, pp. 299–312. Springer, Heidelberg (2012)
16. 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 (1999)
17. Hebb, D.O.: The organization of behavior: a neuropsychological theory. John Wiley & Sons, New York (1949)
18. Lamirel, J.-C., Falk, I., Gardent, C.: Federating clustering and cluster labeling capabilities with a single approach based on feature maximization: French verb classes identification with igngf neural clustering. Neurocomputing, Special Issue on 9th Workshop on Self-Organizing Maps (WSOM 2012) 147, 136–146 (2014)
19. Lang, K.: Learning to filter netnews. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 331–339 (1995)
20. Kohavi, R., John, G.R.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)
21. Ladha, L., Deepa, T.: Feature selection methods and algorithms. International Journal on Computer Science and Engineering 3(5), 1787–1797 (2011)
22. Lallich, S., Rakotomalala, R.: Fast feature selection using partial correlation for multi-valued attributes. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 221–231. Springer, Heidelberg (2000)
23. Lamirel, J.C.: A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research. Scientometrics 93, 151–166 (2012)
24. Lamirel, J.C., Al Shehabi, S., François, C., Hoffmann, M.: New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping. Scientometrics 60(3) (2004)
25. Lamirel, J.C., Cuxac, P., Chivukula, A.S., Hajlaoui, K.: Optimizing text classification through efficient feature selection based on quality metric. Journal of Intelligent Information Systems, Special Issue on PAKDD-QIMIE 2013, 1–18 (2014)
26. Lamirel, J.C., Ghribi, M., Cuxac, P.: 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, France (2010)
27. Lamirel, J.C., Mall, R., Cuxac, P., Safi, G.: Variations to incremental growing neural gas algorithm based on label maximization. In: Proceedings of IJCNN 2011, San Jose, CA, USA (2011)
28. Lamirel, J.C., Ta, A.P.: 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 Meetings, Berlin, Germany (2008)

29. Lamirel, J.-C., Reymond, D.: Automatic websites classification and retrieval using websites communication signatures. Journal of Scientometrics and Information Management: Special Issue on 8th International Conference on Webometrics, Informetrics and Scientometrics 8(2), 293–310 (2014)
30. Martinetz, T., Schulten, K.: A "neural-gas" network learns topologies. In: Artificial Neural Networks, pp. 397–402 (1991)
31. Pearson, K.: On lines an planes of closetst fit to systems of points in space. Philosophical Magazine 2(11), 559–572 (1901)
32. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods, pp. 185–208. MIT Press, Cambridge (1999)
33. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
34. Prudent, Y., Ennaji, A.: An incremental growing neural gas learns topologies. In: Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, vol. 2, pp. 1211–1216 (2005)
35. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
36. Salton, G.: Automatic processing of foreign language documents. Prentice-Hall, Englewood Cliffs (1971)
37. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing (1994)
38. Sun, L., Korhonen, A., Poibeau, T., Messiant, C.: Investigating the cross-linguistic potential of verbnet-style classification. In: Proceedings of ACL, Beijing, China, pp. 1056–1064 (2010)
39. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
40. Yu, L., Liu, H.: Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of ICML 2003, Washington DC, USA, pp. 856–863 (2003)