A new feature selection and feature contrasting approach based on quality metric: application to efficient classification of complex textual data

Jean-Charles Lamirel[1], Pascal Cuxac[2], Kafil Hajlaoui[2] and Aneesh Sreevallabh Chivukula[3]

[1]Equipe SYNALP - LORIA, INRIA Nancy - Grand Est, Vandœuvre-lès-Nancy, France
[2]INIST-CNRS, Vandœuvre-lès-Nancy, France
[3]Center For Data Engineering, International Institute of Information Technology, Hyderabad, India

Abstract: Feature maximization is a cluster quality metric which favors clusters with maximum feature representation as regard to their associated data. This metric has already been successfully exploited, altogether, for defining unbiased clustering quality indexes, for efficient cluster labeling, as well as for substituting to distance in the clustering process, like in the IGNGF incremental clustering method. In this paper we go one step further showing that a straightforward adaptation of such metric can provide a highly efficient feature selection and feature contrasting model in the context of supervised classification. We more especially show that this technique can enhance the performance of classification methods whilst very significantly outperforming (+80%) the state-of-the art variable selection techniques in the case of the classification of unbalanced, highly multidimensional and noisy textual data, with a high degree of similarity between the classes. Our experimental dataset is a reference dataset of 7000 publications related to patents classes issued from a reference classification in the domain of pharmacology.

Introduction

Since the 1990s, advances in computing and storage capacity allow the manipulation of very large data. Whether in bio-informatics or in text mining, it is not uncommon to have description space of several thousand or even tens of thousands of variables. One might think that classification algorithms are more efficient if there are a large number of variables. 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 variables 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 variables. The integration of a variable selection process in the framework of the classification of high dimensional data becomes thus a central challenge.

In the literature, three types of approaches for variable 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 et al. [LAD 11], Bolón-Canedo et al. [BOL 12] Guyon et al. [GUY 03] or Daviet [DAV 09]. We thus only provide hereafter a rapid overview of existing approaches and related methods.

The integrated (embedded) approaches incorporate the selection of the variables in the learning process [BRE 84]. 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) [GUY 02] is an integrated process that performs the selection of variables an iterative basis using a SVM classifier. The process starts with the complete feature set and remove the variables given as the least important by the SVM. The original version uses a linear kernel. However, some extension using non-linear kernels have been proposed to consider potential non-linear dependencies between variables. 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 feature that may be relevant and provide a ranking [MEJ 06].

On their own side, "wrapper" methods explicitly use a performance criterion for searching a subset of relevant predictors. 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 attribute sets using a learning approach. Cross-Validation is used to estimate the accuracy of the learning for a given set of attributes. The algorithm starts with the empty set of attributes and continues until adding attributes does not improve performance [WIT 05].

Forman presents a remarkable work of methods comparison in [FOR 03]. As other similar works, this comparison clearly highlights that, disregarding of their efficiency, one of the main drawbacks of embedded and of the wrapper methods is that they are very computationally intensive. This prohibits their use in the case of highly multidimensional data description space. A potential alternative is thus to exploit filter-based 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 tests' results.

The Chi-square method exploits a usual statistical test that measures the discrepancy to an expected distribution assuming that a variable is independent of a class label. Like any statistical test, he is known to have erratic behavior for very low expected frequencies (which is common case in text classification) [LAD 11].

The information gain is also one of the most common methods of evaluation of the attributes. This univariate filter provides an ordered classification of all variables. Based on to this approach, selected variables are those who obtain a positive value of information gain [HAL 99b].

In the MIFS (Mutual Information Feature Selection) method, a variable X is added to the subset M (of cardinality m) of already selected variables if it maximizes the quantity:

$$I(X,Y|M) = I(X,Y) - \beta * \sum\nolimits_{Z \in M} \frac{I(X,Y)}{m}$$

Thus, a variable is considered to be interesting if its link with the target Y surpasses his average connection with already selected predictors. The method takes into account both the relevance and redundancy. The search stops when the best variable is X such $\mathrm{I}(\mathrm{Y}, \mathrm{X}^{*}|\mathrm{M}) \leq 0$ [BAT 94].

The ThemRMR (minimum Redundancy Maximum Relevance) method selects variables that are most relevant for the target class and that also have low redundancy: it thus comes to select the characteristics that are maximally different from each other. The two optimization criteria (maximum relevance and minimum redundancy) are based on mutual information [PEN 05].

In a similar way, the CFS method (Correlation-based Feature Selection) uses a global measure of "merit" of a subset M of m variables taking into account both their relevance and their redundancy. Then, a relevant subset consists of variables highly correlated with the class, and lowly correlated one to another [HAL 99].

The CBF (Consistency-based Filter) method evaluates the relevance of a subset of variables by the resulting level of consistency of the classes when learning samples are projected onto that subset [DAS 03]. The FCBF method is based on the "symmetrical uncertainty" criterion. A variable is considered to be interesting if: (1) its correlation with the target is high enough (2) it does not exist in the base a variable that is more strongly correlated to that latter [YUL 03].

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 variable. For that, a pairwise correlation coefficient that represents the linear correlation between two variables is used. Once established the table pairwise correlations, the calculation of partial correlation coefficients allows performing a stepwise variable selection [LAL 00] [RAK 02].

The basic assumption of the Relief variable ordering method is to consider that a variable is even more relevant that it discriminates well an object from its nearest neighbor out of its own class, and conversely, that a variable will be irrelevant if it distinguishes between an object and its class nearest neighbor. ReliefF, an extension of Relief, adds the ability to address multiple-class problems. It is also more robust and capable of handling of incomplete and noisy data [KIR 92] [KON 94]. This technique is considered as one of the most efficient filter-based technique.

In this paper, we show that, despite of their diversity, all the existing filter-based approaches fail to successfully solve the variable selection task in the case they are faced with highly unbalanced, highly multidimensional and noisy textual data, with a high degree of similarity between the classes. We thus propose a new filter-based variable selection approach which relies on the exploitation of a class quality measure based on a specific feature maximization metric. Such metric already demonstrated high potential in the framework of unsupervised learning.

The paper is structured as follows. The first section presents the feature maximization principle along with the new proposed technique. The second section describes our dataset and our experiment which is performed experimental is a reference dataset of 7000 publications related to patents classes issued from a reference classification in the domain of pharmacology. The last section draws our conclusion and our perspectives.

Feature maximization for variable selection

Feature maximization metric principles in unsupervised learning

Feature maximization is an unbiased cluster quality metrics that exploits the properties of the data associated to each cluster without prior consideration of clusters profiles. This metrics has been initially proposed in [LAM 04]. Its main advantage is to be independent altogether of the clustering methods and of their operating mode. Whenever it is used during the clustering process, it can substitute to distance during that process [LAM 11b]. In a complementary way, whenever it is used after learning, it can be exploited to set up overall clustering quality indexes [LAM 10][GHR 10] or for cluster labeling [LAM 08].

Let us consider a set of clusters *C* resulting from a clustering method applied on a set of data *D* represented with a set of descriptive features *F*, feature maximization is a metric which favors clusters with maximum *Feature F-measure*. The *Feature F-measure* $FF_c(f)$ of a feature *f* associated to a cluster *c* is defined as the harmonic mean of *Feature Recall* $FR_c(f)$ and *Feature Precision* $FP_c(f)$ indexes which in turn are defined as:

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

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

where W_d^f represents the weight of the feature *f* for data *d* and F_c represent the set of features occurring in the data associated to the cluster *c*.

An important application of the feature maximization metric is related to the estimation of the overall clustering quality. For that purpose, averaged *Macro-Recall* (*MR*) and *Macro-Precision* (*MP*) indexes can be directly derived from the former indexes.

They are expressed as:

$$MR = \frac{1}{|C|}\sum_{c\epsilon C}\frac{1}{|F_c|}\sum_{f\epsilon F_c} FR_c(f), MP = \frac{1}{|C|}\sum_{c\epsilon C}\frac{1}{|F_c|}\sum_{f\epsilon F_c} FP_c(f) \qquad (3)$$

Macro-Recall and *Macro-Precision* indexes have opposite behaviors according to the number of clusters. Thus, these indexes permit to estimate in a global way an optimal number of clusters for a given method and a given dataset. The best data partition, or clustering result, is in this case the one which minimizes the difference between their values [LAM 04]. Conversely to classical distance-based indexes, they are independent of the clustering process. Moreover, it has been demonstrated in [LAM 11] that straightforward adaptations of these indexes permits to detect degenerated clustering results, whenever those jointly include a small number of heterogeneous or "garbage" clusters with large size and a big number of "chunk" clusters with very small size.

Another 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 [ATT 06]. It can rely on endogenous data properties or on exogenous ones. Endogenous data properties represent the ones being used during the clustering process. Exogenous data properties represent either complementary properties or specific validation properties. Exploiting feature maximization metric for cluster labeling results in a parameter-free labeling technique [LAM 08]. 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_c of prevalent features of a cluster *c* can be defined as:

$$L_c = \{f \in F_C \mid FF_c(f) = Max_{c' \in C}(FF_{c'}(f))\} \qquad (4)$$

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 clustering output issued from the management of highly multidimensional data [LAM 08]. 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 [LAM 10b]. 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 [LAM 12].

Last but not least, a central application of feature maximization metric is related to incremental clustering. The IGNGF (Incremental Neural Gas with Feature Maximization) clustering method is a neural-based parameter-free incremental clustering algorithm that substitutes feature maximization to usual distance in the clustering process. Thanks to this approach, the IGNGF clustering process is roughly the following. During learning, an incoming data point *d* is temporary added to every existing cluster, its feature profile is updated (i.e. each cluster is associated with its maximal features) and its average *Feature F-measure* is computed. Then the winning cluster is the cluster which maximizes the *Kappa* criteria given by:

$$K(c) = \Delta((FF_c) * |F_c \cap F_d| - \frac{EucDist(\vec{c},d)}{weight} \qquad (5)$$

where $\Delta(FF_c)$ represents the gain in *Feature F-measure* for the new cluster and $F_c \cap F_d$ are the features shared by cluster *c* and the data point *d*. 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 can be optionally adjusted by the Euclidean distance of the new data point *d* to the cluster centroid vector $\vec{c}$. Clusters with negative *Kappa* score are ignored. The data point is then added to the cluster *c* with maximal Kappa and Hebbian connections between winner and its neighbors are updated. If not such cluster is found, a new cluster is created.

The IGNGF method was shown to outperform other usual neural and non neural methods for clustering tasks on relatively clean data, and especially if said data are sparse and/or highly multidimensional [Lam 11]. The first applications of the IGNF method for clustering of textual data revealed very promising results. Especially, this method was exploited for the automatic classification of the French verbs using syntactic and semantic features issued from several reference lexicons. The method showed significantly better performance (+20%) than the best state-of-the-art methods of the field, including the reference methods based on spectral clustering [FAL 12]. In the context of the websites' classification, it has been also shown that the IGNGF method allowed, in an unattended way, to obtain better results (in terms of sensibility and purity) than those provided by the supervised methods this by automatically isolating latent, not originally labeled, classes [LAM 12b].

Adaptation of feature maximization metric for feature selection in supervised learning

Taking into consideration the basic definition of feature maximization metric presented in the former section, its exploitation for the task of feature selection in the context of supervised learning becomes a straightforward 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 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 ($\mathrm{FR}_{\mathrm{c}}(\mathrm{f})$ index).

The set S_c of features that are characteristic of a given class *c* belonging to an overall class set *C* results in:

$$S_c = \{f \in F_C \mid FF_c(f) > \overline{FF}(f) \; and \; FF_c(f) > \overline{FF}_D \} \tag{6}$$

where $\overline{FF}(f) = \sum_{c' \in C} FF_{c'}(f) \,/\, |C_{/f}|$ and $\overline{FF}_D = \sum_{f \in F} \overline{FF}(f) / \, |F|$.

and $C_{/f}$ represent 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:

$$S_C = \cup_{c \in C} S_c \tag{7}$$

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 fit the description of each data to the specific characteristic 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. This step more precisely operates as described in Algorithm 1.

Thanks to the former strategy, the information gain provided by a feature in a given class is proportional to the ratio between the value of the *Feature F-measure* of this feature in the class and the average value of the *Feature F-measure* of the said feature on all the partition. For a given data and a given feature describing this data, the resulting gain acts a contrast weight factorizing with any existing feature weight that can be issued from data preprocessing. Moreover, each data description can be optionally reduced to the features which are characteristic of its associated class. If it is present, normalization of the data description is discarded by those operations. Optional renormalization can also be performed in the curse of the algorithm.

Algorithm 1: feature maximization-based data descriptions contrasting

Data:
C: set of data classes
F: set of descriptive features (variables)
D: set of learning data (vectors on F)

Output:
D': set of updated learning data

```
$Foreach\ c \in C$
    $Foreach\ d \in c$
        $Foreach\ f \in F$
            $If\ (f \in F_c)$
                $then\ W_d^f \leftarrow (FF_c(f)/\overline{FF}(f)) * W_d^f$
                $else\ W_d^f \leftarrow 0\ (optional)$
            $Enfif$
        $Endfor$
        $Normalize\ (W_d)\ (optional)$
    $Endfor$
$Endfor$
```

Experimental data and results

Data extraction and preprocessing

The data is a collection of patent documents related to pharmacology domain. The bibliographic citations in the patents are extracted from the Medline database[1]. 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 [HAL 12]. Then the Medline database is queried with extracted citations for related scientific articles. The querying gives 7501 articles with 90% recall. Each article is then labeled by the 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 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 sampling that introduces redundancy in dataset does not improve the performance in this dataset, as it has been shown in [HAL 12]. So that bootstrapping of train and test data may solve problems of classification sensibility, stability, scalability and dimensionality but does not improve accuracy computation over the sampled correlations. Conversely, pruning irrelevant features and contrasting the relevant ones has we propose hereafter seems thus to be a good alternative.

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

The document set is converted to a bag of words model [SAL 71] using the TreeTagger tool [SCH 94] developed by the Institute for Computational Linguistics of the University of Stuttgart. This tool is both a lemmatizer and a tagger. A lemmatizer associates a lemma, or a syntactic root, to each word in the text and a tagger automatically annotates text with morpho-syntactic information. In our case, the documents are firstly lemmatized and the tagging process is performed on lemmatized items (in the case when a word is unknown to the lemmatizer, its original form is conserved). The punctuation signs and the numbers identified by the tagger are deleted. The feature selection according to grammatical categories allows identifying salient features for the documents classification according to document types or opinions.

Every document is represented as a term vector filled with keyword frequencies. 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. The whole text collection is then represented as a (N+1) x J matrix where J is number of articles in the collection in a N-dimensional space. Each line j of this matrix is an N-dimensional bag of words vector for article j, plus its class label. The Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme [SAL 88] gives a sparse matrix representation of the text collection.

Testing process

To perform our experiments we firstly take into consideration different classification algorithms which are implemented in the Weka toolkit:

- Weka’s Decision Tree algorithm: weka.classifiers.trees.J48 [QUI 93] ;
- Weka’s Random Forest algorithm: weka.classifiers.trees.RandomForest [BRE 01] ;
- Weka’s KNN algorithm: weka.classifiers.lazy.IBk [AHA 91] ;
- Weka’s Bayesian Network algorithm: weka.classifiers.bayes.DMNBtext [SU 08] ;
- Weka’s SVM algorithm: weka.classifiers.functions.SMO [SCH 98], [KEE 01].

Most of these algorithms are general purpose classification algorithms, except from DMNBtext which is a Discriminative Multinomial Naïve Bayes classifier especially developed for text classification. As compared to classical Multinomial Naïve Bayes classifier this algorithm cumulate the computational efficiency of Naïve Bayes approaches and the accuracy of Discriminating approaches by taking into account both the likelihood and the classification objectives during the frequency counting. Other general purpose algorithms whose accuracy has especially been reported for text classification are SMO and KNN [ZHA 02]. Default parameters are used when executing these algorithms, except for KNN for which the number of neighbors is optimized.

To more especially focus on the efficiency testing of the variable (i.e. feature) selection approaches including our new proposal. We include in our test a panel of filter-based approaches which are computationally tractable with high dimensional data, making again use of their Weka toolkit implementation. The panel of tested methods includes:

- Weka’s Chi-square method: weka.attributeSelection.ChiSquaredAttributeEval [LAD 11] ;
- Weka’s Information gain method: weka.attributeSelection.InfoGainAttributeEval [HAL 99b] ;
- Weka’s CBF method: weka.attributeSelection.ConsistencySubsetEval [DAS 03] ;
- Weka’s SU method: weka.attributeSelection.SymmetricalUncertaintyAttributeEval [YUL 03] ;
- Weka’s ReliefF algorithm: weka.attributeSelection.ReliefFAttributeEval [KIR 92] ;
- Weka’s Principal Component Analysis: weka.attributeSelection.PrincipalComponents [PER 01] ;
- Feature maximization based method including contrasting (our current proposal).

Defaults parameters are also used for most this methods, except for PCA for which the percentage of explained variance is tuned for optimization.

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

Results

The different results are reported in tables 1 to 5 and in figure 1. Tables present standard performance measures (True Positive, False Positive, Precision, Recall, F-measure and ROC) weighted and averaged over all classes. For each table, and each combination of selection and classification methods, a performance increase indicator is computed using the DMNBtext 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.

Table 1 highlights that performance of all classification methods are low on the considered dataset if no feature selection process is performed. They also confirm the superiority of the DMNBtext, SMO and KNN methods on the two other tree-based methods in that context. Additionally, DMNBtext provides the best overall performance in terms of discrimination as it is illustrated by its highest ROC value.

Whenever a usual feature selection process is performed in combination with the best method, that is DMNBtext method, the exploitation of the usual feature selection strategies slightly alters the quality of the results, instead of bringing up an added value, as it is shown in table 2. Alternatively, same table highlights that even if the feature reduction effect is less with the F-max selection method, its combination with F-max data description contrasting boosts the performance of the method (+81%), leading to excellent classification results (Accuracy of 0.96) in a very complex classification context.

Even if the benefit of the former use of F-max selection and contrasting approach is very high with the DMNBtext method, table 3 shows that the added value provided by this preprocessing approach also concerns, to a lesser extent, all the other classifiers, leading to an average increase of their performance of 45% as compared to the reference result. Another interesting phenomenon that can be observed is that, with such help, tree-based classification methods significantly, and unusually, outperform the KNN method on text.

The results presented in table 4 more specifically illustrates the efficiency of the F-max contrasting procedure that acts on the data descriptions. In the experiments related to that table, F-max contrasting is performed individually on the features extracted by each selection method and, in a second step, DMNBtext classifier is applied on the resulting contrasted data (see algo 1). 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%.

	TP Rate	FP Rate	Precision	Recall	F-measure	ROC	TP Incr. /Ref
J48	0.42	0.16	0.40	0.42	0.40	0.63	-23%
Random Forest	0.45	0.23	0.46	0.45	0.38	0.72	-17%
SMO	0.54	0.14	0.53	0.54	0.52	0.80	0%
DMNBtext	**0.54**	**0.15**	**0.53**	**0.54**	**0.50**	**0.82**	**0% (Ref)**
KNN (k=3)	0.53	0.16	0.53	0.53	0.51	0.77	-2%

Table 1: classification results on initial data.

Table 5 and figure 1 illustrate the capabilities of the F-max approach to efficiently cope with the class imbalance problem. Hence, examination of the confusion matrices of figure 1 shows that the data attraction effect of the majority class that occurs at a high level in the case of the exploitation of the original data (figure 1(a)) is quite completely overcome whenever the F-max approach is exploited (figure 1(b)). The capability of the approach to correct class imbalance is also clearly highlighted by the homogeneous distribution of the selected variables in the classes it provides, despite of their very different sizes (table 5).

	TP Rate	FP Rate	Precision	Recall	F-measure	ROC	Nbr. of select. features	TP Incr./Ref
ChiSquare (+..)	0.52	0.17	0.51	0.52	0.47	0.80	282	-4%
CBF	0.47	0.21	0.44	0.47	0.41	0.75	37	-13%
PCA (50% vr.)	0.47	0.18	0.47	0.47	0.44	0.77	483	-13%
Relief	0.52	0.16	0.53	0.52	0.48	0.81	937	-4%
F-max sel. + contrast	**0.96**	**0.01**	**0.96**	**0.96**	**0.96**	**0.999**	**1419**	**+81%**

Table 2: classification results after feature selection (DMNBtext classification).

	TP Rate	FP Rate	Precision	Recall	F-measure	ROC	TP Incr./Ref
J48	0.80	0.05	0.79	0.80	0.79	0.92	+48%
Random Forest	0.76	0.09	0.79	0.76	0.73	0.96	+40%
SMO	0.92	0.03	0.92	0.92	0.91	0.98	+70%
DMNBtext	**0.96**	**0.01**	**0.96**	**0.96**	**0.96**	**0.999**	**+81%**
KNN (k=3)	0.66	0.14	0.71	0.66	0.63	0.85	+22%

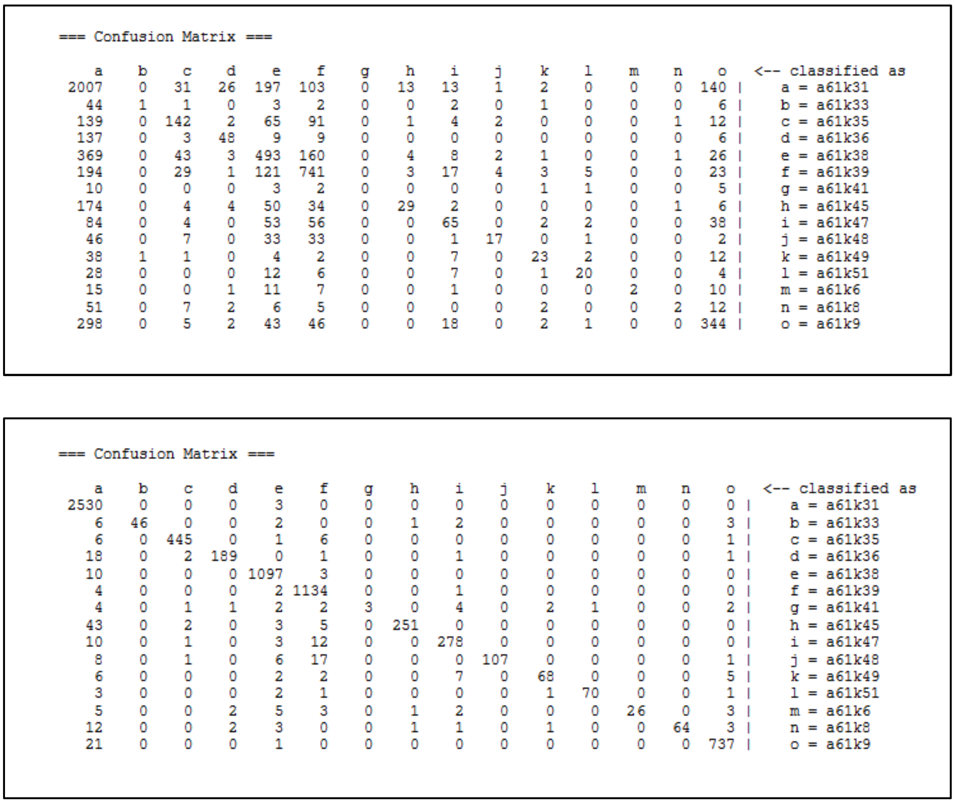
Table 3: classification results after F-max + contrast feature selection (all classification methods).

	TP Rate	FP Rate	Precision	Recall	F-measure	ROC	Nbr. select. features	TP Incr./Ref
ChiSquare (+..)	0.79	0.08	0.82	0.79	0.78	0.98	282	+46%
CBF	0.63	0.15	0.69	0.63	0.59	0.90	37	+16%
PCA (50% vr.)	0.71	0.11	0.73	0.71	0.67	0.53	483	+31%
Relief	0.79	0.08	0.81	0.79	0.78	0.98	937	+46%
F-max sel. + contrast	**0.96**	**0.01**	**0.96**	**0.96**	**0.96**	**0.999**	**1419**	**+81%**

Table 3: classification results after feature selection by all methods and F-max contrasting (DMNBtext classification).

Class label	Class size	Selected features	TP Rate
a61k31	2533	223	0.999
a61k33	60	276	0.77
a61k35	459	262	0.97
a61k36	212	278	0.89
a61k38	1110	237	0.99
a61k39	1141	240	0.99
a61k41	22	225	0.14
a61k45	304	275	0.83
a61k47	304	278	0.91
a61k48	140	265	0.76
a61k49	90	302	0.76
a61k51	78	251	0.90
a61k6	47	270	0.55
a61k8	87	292	0.74
a61k9	759	250	0.97
Distinct features		1419	

Table 5: class data and F-max selected features/class.



```
=== 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
```

Figure 1: confusion matrices of the optimal results - before (1) and after (2) feature selection (Classification: DMNBtext – Feature selection: F-max + Contrast).

Conclusion

Feature maximization is a cluster quality metric which favors clusters with maximum feature representation as regard to their associated data. In this paper, we have proposed a straightforward adaptation of such metric, which has already demonstrated several generic advantages in the framework of unsupervised learning, to the context of supervised classification. 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 data, and more especially in that of large full text data. These problems relate to classes' imbalance, high dimensionality, noise, and high degree of similarity between classes. Through our experiments on a large dataset constituted of bibliographical records extracted from a patents' classification, we more especially showed that our approach can naturally cope with the said handicaps. Hence, in such context, whereas the state-of-the-art variable selection techniques remain inoperative, feature maximization-based variable selection and contrasting can very significantly enhance the performance of classification methods (+80%). Another important advantage of this technique is that it is a parameter-free approach and it can thus be used in a larger scope, like in the one of semi-supervised learning.

References

[AHA 91] Aha, D. & D. Kibler (1991). Instance-based learning algorithms. Machine Learning. 6:37-66.

[ATT 06] Attik, M., J.-C. Lamirel & S. Al Shehabi (2006). Clustering analysis for data with multiple labels, Proceedings of the IASTED International Conference on Databases and Applications (DBA), Innsbruck, Austria.

[BAT 94] Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4): 537-550, 1994.

[BRE 01] Breiman, L. (October 2001). Random forests. Machine Learning 45(1), 5–32.

[BOL 12] Bolón-Canedo, V., N. Sánchez-Maroño & A. Alonso-Betanzos (2012). A Review of Feature Selection Methods on Synthetic Data. Knowledge and Information Systems (mars 1, 2012): 1-37.

[BRE 84] Breiman, L., J.H. Friedman, R.A. Olshen, & C.J. Stone (1984). Classification and Regression Trees, Wadsworth International Group, Wadsworth International Group, Belmont, CA.

[COH 02] Cohen, I., T.X. Qi, Z. Sean, S. Xiang, T. Zhou & T.S. Huang (2002). Feature Selection Using Principal Feature Analysis. Rochester, New York, USA, S, 2002.

[CRI 00] Nello, C., H. Lodhi, & J. Shawe-taylor (2000). Latent Semantic Kernels for Feature Selection, 2000.

[DAS 03] Dash, M. & H. Liu (2003). Consistency-based search in feature selection. Artificial Intelligence 151, n° 1 (2003): 155-176.

[DAV 09] 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, 2009.

[FAL 12] Falk, I., C. Gardent & J.-C. Lamirel (2012). Classifying French Verbs using French and English Lexical Resources Proceedings of ACL 2012. Jeju Island Korea, July 2012.

[FOR 03] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research 3 (2003): 1289–1305.

[GHR 10] Ghribi, M., P. Cuxac, J.-C. Lamirel & A. Lelu (2010). Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots-clés, Proceedings of the 10th International Francophone Conference on Knowledge Extraction and Management (EGC 2010), Hammamet, Tunisia, January 2010.

[GUY 02] Guyon, I., J. Weston, S. Barnhill & V. Vapnik (2002). Gene selection for cancer classification using support vector machines. Machine learning 46, n° 1 (2002), 389–422.

[GUY 03] Guyon, I. & A. Elisseeff (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research 3 (2003): 1157–1182.

[HAL 99] Hall, M.A. (1999). Correlation-based Feature Selection for Machine Learnin. PhD, University of Waikato, 1999.

[HAL 99b] Hall, M.A. & L.A. Smith (1999). 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, 235–239. AAAI Press, 1999.

[HAL 12] Hajlaoui, K., P. Cuxac, J.-C Lamirel & C. François (2012). Enhancing patent expertise through automatic matching with scientific papers. Discovery Science, LNCS 7569, 299–312.

[KEE 01] Keerthi, S., S. Shevade, C. Bhattacharyya & K. Murthy (2001). Improvements to platt's smo algorithm for svm classifier design. Neural Computation 13(3), 637–649.

[KIR 92] Kira, K. & L.A. Rendell (1992). A Practical Approach to Feature Selection. In Ninth International Workshop on Machine Learning, 249-256, 1992.

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

[KON 94] Kononenko, I. (1994). Estimating Attributes: Analysis and Extensions of RELIEF. In: European Conference on Machine Learning, 171-182, 1994.

[LAD 11] Ladha, L. & T. Deepa (2011). Feature selection methods and algorithms. International Journal on Computer Science and Engineering, 3, n° 5 (2011): 1787–1797.

[LAD 11] Lallich, S. & R. Rakotomalala (2000). Fast Feature Selection Using Partial Correlation for Multi-valued Attributes. In Principles of Data Mining and Knowledge Discovery, édité par Djamel A. Zighed, Jan Komorowski, et Jan Żytkow, 221-231. Lecture Notes in Computer Science 1910. Springer Berlin Heidelberg, 2000.

[LAM 04] Lamirel, J.-C., S. Al Shehabi, C. Francois & M. Hoffmann (2004). New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping, Scientometrics, 60(3) 2004.

[LAM 08] Lamirel, J.-C. & A.P. Ta (2008). Combination of hyperbolic visualization and graph-based approach for organizing data analysis results: an application to social network analysis, Proceedings of the 4th International Conference on Webometrics, Informetrics and Scientometrics and 9th COLLNET Meeting, Berlin, Germany, August 2008.

[LAM 10] Lamirel, J.-C., M. Ghribi, & P. Cuxac (2010). Unsupervised recall and precision measures: a step towards new efficient clustering quality indexes, Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010), Paris, France, August 2010.

[LAM 10b] Lamirel, J.-C, N. Priyankar, P. Cuxac & G. Safi (2010). Mining research topics evolving over time using a diachronic multi-source approach, Proceedings of ICDM 2010 International Workshop on Mining Multiple Information Sources, Sydney, Australia, December 2010.

[LAM 11] Lamirel, J.-C, R. Mall, P. Cuxac & G. Safi (2011). A new efficient and unbiased approach for clustering quality evaluation, Proceedings of PAKDD 2010 2nd International Workshop on Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models (QIMIE), Shenzen, China, May 2011.

[LAM 11b] Lamirel, J.-C, R. Mall, P. Cuxac & G. Safi (2011). Variations to incremental growing neural gas algorithm based on label maximization, Proceedings of IJCNN 2011, San Jose, CA, USA, August 2011.

[LAM 12] Lamirel, J.-C. (2012). A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research Scientometrics (2012) 93: 151-166 , October 01, 2012

[LAM 12b] Lamirel, J.-C, & D. Reymond (2012). Automatic websites classification and retrieval using websites communication signatures, 8th International Conference on WIS and 13th Collnet Meeting, Korea, October 2012.

[MEJ 06] Mejía-Lavalle, M., E. Sucar, & G. Arroyo (2006). Feature selection with a perceptron neural net. Feature Selection for Data Mining: Interfacing Machine Learning and Statistics (2006), 131.

[NOV 08] Novakovic, J. (2008). Using Information Gain Attribute Evaluation to Classify Sonar Targets. International Journal of Image Processing (2008).

[PEN 05] Peng, H., F. Long, & C. Ding (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27, nº 8 (2005): 1226–1238.

[PER 01] Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space". Philosophical Magazine 2 (11): 559–572.

[QUI 93] Quinlan, R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

[RAK 02] Rakotomalala, R. & S. Lallich (2002). Construction d'arbres de décision par optimisation", Revue Extraction des Connaissances et Apprentissage, vol. 16, n°6/2002, pp.685-703.

[SAL 71] Salton, G. (1971). Automatic processing of foreign language documents. Prentice-Hill: Englewood, Cliffs, NJ.

[SAL 88] Salton, G. & C. Buckley (1988). Term weighting approaches in automatic text retrieval. Information Processing and Management 24(5), 513–523.

[SCH 94] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. Proceedings of International Conference on New Methods in Language Processing.

[SCH 98] Schoelkopf, B., C. Burges & A. S. editors (1998). Advances in Kernel Methods - Support Vector Learning. MIT Press.

[SU 08] Su, J., H. Zhang, C. Ling, & S. Matwin (2008). Discriminative parameter learning for bayesian networks. ICML.

[WIT 05] Witten, I.H. & E. Frank (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[YUL 03] Yu, L. & H. Liu, (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. ICML 2003, 856-863, August 21-24, 2003, Washington DC, USA.

[ZHA 01] T. Zhang and F. J. Oles. Text categorization based on regularized linear classification methods. Inf. Retr., 4(1):5–31, 2001.

[ZHO 10] Zhong, J., D. Xiongbing, L. Jie, L. Xue & L. Chuanwei (2010). A Novel Chinese Text Feature Selection Method Based on Probability Latent Semantic Analysis. In Advances in Neural Network , 6064:276-281. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.