**Enhancing feature selection with feature maximization metric**

**Jean-Charles Lamirel**
Synalp Team, LORIA, Nancy, France
Email: lamirel@loria.fr

## Abstract

This paper deals with a new feature selection and feature contrasting approach for classification of highly unbalanced textual data with a high degree of similarity between associated classes. The efficiency of the approach is illustrated by its capacity to enhance the classification of bibliographic references into a patent classification scheme. A complementary experiment is performed on a non textual dataset issued form the UCI repository.

Keywords: feature selection, feature maximization, classification, imbalanced data, similar classes,

## 1. Introduction

Text categorization is a machine learning task which aims at automatically assigning predefined category labels to new upcoming free text documents with related characteristics [COH 05]. Because of its numerous applications, like mail or news filtering [COR 07], emotion detection [PAN 08], text genre analysis [BHA 93], text classification has been one of the most studied branches within the field of machine learning [HIL 07]. However, several classification problems raise new challenges in the domain, especially those ones which implies to deal with imbalanced data and highly similar classes. In the context of text categorization, patents validation assistance takes part in that class. It consists in generating help to experts in their task of evaluation of the novelty of a patent based on the automatic assignation of the most relevant scientific papers related with the classification codes of the said patent. As soon as learning is based on citations extracted from the patents which are usually associated with a hierarchy of classification codes having different levels of generality, first, there is no guaranty of a homogeneous distribution of the citations (i.e. learning samples) among the codes, second, there is a high chance to have similar citations in different classes.

We illustrate in that paper that the exploitation of standard strategies for classification or preprocessing, like feature selection, would not produce any exploitable results in the above mentioned context. We thus propose a new feature selection approach. The remaining of the paper is structured as follows. Section 2 presents the process we used to generate our experimental dataset. Section 3 presents a new feature selection approach suitable to deal with the unsolved class imbalance and class similarities problems. Section 4 compares the results with and without the use of the proposed approach. Section 5 draws our conclusion and perspectives.

## 2. Patents and references data

Our main experimental resource is issued from the QUAREO[1] project. It is a collection of patents related to the pharmacology domain completed with bibliographic references issued from the Medline[2] database. The source data contains 6387 patents in XML format, grouped into 15 subclasses of the A61K class (medical preparation). For obtaining the bibliographical references, 25887 citations are firstly extracted from the patents. Then the Medline database is queried with extracted citations for related ref-

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

erences. The querying results in 7501 bibliographical references[3] which are then labeled by the class code of their citing patent.

The set of labeled references represents the final document set on which the training is performed. It is converted to a bag of words model [SAL 71] using the TreeTagger syntactic analyzer [SCH 94]. In our case, the full text of the references is 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). Every reference is finally represented as a term vector filled with term frequencies. The description space generated by the tagger has dimensionality 31214. To reduce the generated noise, 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. Finally, TF-IDF weighting scheme [SAL 88] is exploited on the thresholded space to obtain a sparse representation of the data.

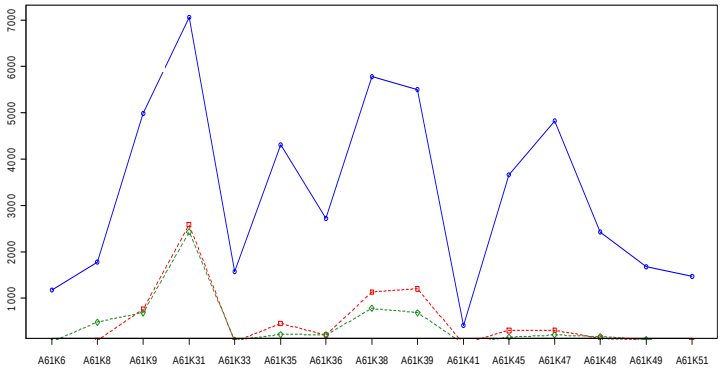



**Fig. 1.** Distribution of data in patents classes
(patents (Green) - cited references (Red) – descriptors (Blue)).

Figure 1 highlights the highly imbalanced distribution of both, patents, extracted references and keywords attached with references relatively to the different class codes. As an example, smallest class contains only 22 extracted references (A61K41 class) whilst the biggest one has more than 2500 (A61K31 class).

The exploitation of resampling techniques [GOO 06] as well as the one of standard feature selection techniques [FOR 03] could be envisaged to compensate influence of the biggest classes. However, in our context, the ability of such techniques to precisely detect the right class is curtailed by the high class to class similarity due to the association of the initial patents to a specialized branch of the patent classification: inter-class similarity computed using cosine correlation between class profiles generated by the descriptors issued from the extracted bibliographical references indicates that more than 70% of classes' couples have a similarity between 0.5 and 0.9.

As an alternative, we thus propose a new filter approach which relies on the exploitation of a class-based quality measure grounded on the feature maximization metric (F-max). Such metric has been formerly exploited by Falk et al. in the unsupervised context for clustering French verbs relying on syntactic and semantic features [FAL 12] and said authors demonstrated both its intrinsic efficiency for the clustering task and its generic advantages for cluster labeling.

## 3. New feature selection approach

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 intro-

[3] Medline recall is 90%, relatively to unique references.

duced by Falk et al. in [FAL 12] 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:

$$FR_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{c' \in C} \sum_{d \in c'} W_d^f}, \qquad FP_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{f' \in F_c, d \in c} W_d^{f'}}$$

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*.

Taking into consideration the basic definition of feature maximization metric presented above, the feature maximization-based feature selection process can thus be defined as a parameter-free and class-based process in which a class feature is characterized using both its capacity to discriminate a given class from the others ($FR_c(f)$ index) and its capacity to accurately represent the class data ($FP_c(f)$ index). The set $S_c$ of features that are characteristic of a given class *c* belonging to an overall class set *C* results in:

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

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

with $C_{/f}$ representing the restriction of the set *C* to the classes in which the feature *f* is present.

In other words, 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 F-max metric.

In a complementary way, a class-based feature contrast factor can be introduced by taking into consideration the "information gain" provided by the Feature F-measures of the features, locally to that class. For a feature *f* belonging to the set of selected features $S_C$ of a class *c*, the gain $G_c(f)$ results in:

$$G_c(f) = (FF_c(f) / \overline{FF}(f))$$

## 4. Experiments and results

To perform our experiments we firstly exploit different classification algorithms which are implemented in the Weka toolkit[4]: J48 Decision Tree algorithm [QUI 93], Random Forest algorithm [BRE 01] (RF), KNN algorithm [AHA 91], DMNBtext Bayesian Network algorithm [SU 08] (DMT) and SMO-SVM algorithm [PLA 98] (SMO).

Most of these algorithms are general purpose classification algorithms, except from DMNBtext which is a Discriminative Multinomial Naive Bayes classifier especially developed for text classification. Default parameters are used when executing these algorithms, except for KNN for which the number of neighbors is optimized based on resulting accuracy.

We then focus on testing the efficiency of the feature selection approaches including our new proposal (FMC). We include in our test a panel of filter approaches which are

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

computationally tractable with high dimensional data[5], making again use of their Weka toolkit implementation: Chi-square selector [LAD 11], Information gain selector |HAL 99], CBF subset selector [DAS 03] (CBF), Symmetrical Uncertainty selector [YUL 03], ReliefF selector [KON 94] (RLF) and Principal Component Analysis selector [PER 01] (PCA). Defaults parameters are also used for most this methods, except for PCA for which the percentage of explained variance is tuned based on resulting accuracy. 10-fold cross validation is used on all our experiments.

The different results are reported in tables 1 to 3 and in figures 2 to 3. Tables and figures present standard performance measures weighted by class sizes and averaged over all classes. For each table, and each combination of feature selection and classification methods, a performance increase indicator is computed using the DMT True Positive results on the original data as the reference. Finally, as soon as the results are identical for Chi-square, Information Gain and Symmetrical Uncertainty, they are thus reported only once in the tables as Chi-square results (and noted CHI+).

Table 1 highlights that performance of all classification methods are low on the considered dataset if no feature selection process is performed. DMNBtext provides the best overall performance in terms of discrimination as it is illustrated by its highest ROC value. However, as it is also shown by confusion matrix of figure 2, the method is clearly inefficient in an operational patent evaluation context because of its high resulting confusion between classes.

Whenever a usual feature selection process is performed in combination with the best method, that is DMT 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 1. Same table highlights that, conversely, FMC feature selection and contrast boosts the performance of the DMT method: Accuracy of 0.96 (+81%) and ROC of 0.999 (+21%).

| | TP | FP | P | R | F | ROC | TP Incr. |
|---|---|---|---|---|---|---|---|
| J48 | 0.42 | 0.16 | 0.40 | 0.42 | 0.40 | 0.63 | -23% |
| RF | 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% |
| **DMT** | **0.54** | **0.15** | **0.53** | **0.54** | **0.50** | **0.82** | **0% (Ref)** |
| KNN | 0.53 | 0.16 | 0.53 | 0.53 | 0.51 | 0.77 | -2% |

**Table 1**. Classification results on initial data.

| | TP | FP | P | F | ROC | Nbr Feat. | TP Incr. |
|---|---|---|---|---|---|---|---|
| CHI+ | 0.52 | 0.17 | 0.51 | 0.47 | 0.80 | 282 | -4% |
| CBF | 0.47 | 0.21 | 0.44 | 0.41 | 0.75 | 37 | -13% |
| PCA | 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** | **0.96** | **0.01** | **0.96** | **0.96** | **0.999** | **262/cl** | **+81%** |

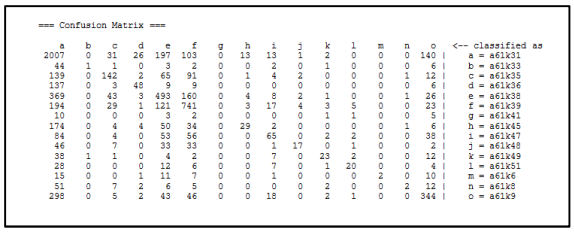
**Table 2**. Classification results after feature selection (DMT classification, all feature selection methods).

| | TP | FP | P | F | ROC | TP Incr. |
|---|---|---|---|---|---|---|
| J48 | 0.80 | 0.05 | 0.79 | 0.79 | 0.92 | +48% |
| RF | 0.76 | 0.09 | 0.79 | 0.73 | 0.96 | +40% |
| SMO | 0.92 | 0.03 | 0.92 | 0.91 | 0.98 | +70% |
| **DMT** | **0.96** | **0.01** | **0.96** | **0.96** | **0.999** | **+81%** |
| KNN | 0.66 | 0.14 | 0.71 | 0.63 | 0.85 | +22% |

**Table 3**. Classification results after FMC feature selection (all classification methods).

[5] Exhaustive presentation and comparison of usual feature selection methods can be found in [FOR 03].

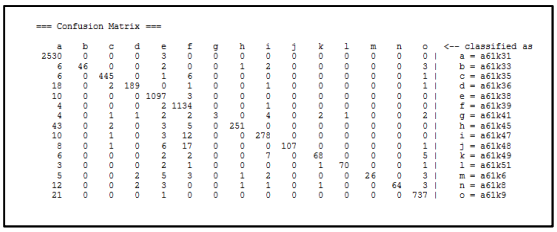
Table 2 and figures 2-3 illustrate the capabilities of the FMC approach to efficiently cope with the class imbalance and class similarity problems. Hence, examination of confusion matrices of figures 2-3 shows that the data attraction effect of the biggest classes that occurs at a high level in the case of the exploitation of the original data (figure 2) is quite completely overcome whenever the FMC approach is exploited (figure 3).



```
=== 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
```

**Figure 2.** Confusion matrix of the optimal results before feature selection (DMT classification).



```
=== 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 3**. Confusion matrix of the optimal results after FMC feature selection (DMT classification)

Table 4 presents the results of a complementary experiment performed on non textual data. The considered dataset is the UCI's Lung cancer dataset in its binary form: 32 samples are described by 144 binary features and are split into to 3 different classes. Even, if complementary test must be done, the obtained results figure out that the FMC method has interesting potential to be exploited in a broader context than the one of textual data.

| | TP | FP | P | F | ROC | TP Incr. | Classif. method |
|---|---|---|---|---|---|---|---|
| No selection | 0,63 | 0,21 | 0,68 | 0,64 | 0,71 | **0% (Ref)** | NB |
| CHI+ | 0,63 | 0,21 | 0,68 | 0,64 | 0,71 | 0% | NB |
| CBF | 0.69 | 0.16 | 0.70 | 0.69 | 0.87 | +8% | NB |
| PCA | 0.53 | 0.28 | 0.53 | 0.49 | 0.73 | -26% | NB |
| RLF | 0,63 | 0,21 | 0,68 | 0,64 | 0,71 | 0% | NB |
| **FMC** | **0.81** | **0.11** | **0.82** | **0.81** | **0.86** | **+26%** | BN |

**Table 4**. Best results on UCI Lung cancer dataset (mixed classification methods: NB = Naïve Bayes, BN = Bayesian Network).

## 5. Conclusion

Feature maximization is an efficient cluster quality metric which favors clusters with maximum feature representation as regard to their associated data. Using this metric we build up an efficient feature selection and feature contrasting model that proved to overcome the usual problems arising in the supervised classification of large volume of full text data. These problems relate to classes imbalance, high dimensionality, noise, and high degree of similarity between classes.

## References


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

[BHA 93] Bhatia, V. K. (1993) "*Analyzing Genre - Language Use in Professional Settings*", London, Longman, Applied Linguistics and Language Study Series.

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

[COH 05] Cohen, A.M. & Hersh, W.R. (2005) "*A survey of current work in biomedical text mining*", Briefings in Bioinformatics 6, pp. 57-71.

[COR 07] Cormack, G.V. & Lynam, T.R. (2007) "*Online supervised spam filter evaluation*", ACM Transactions on Information Systems, 25(3):11.

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

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

[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.

[GOO 06] Good, P. (2006) "*Resampling Methods*", 3rd Ed. Birkhauser.

[HAL 99] Hall, M.A. & Smith, L.A. (1999) "*Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper*", In Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, pp. 235–239. AAAI Press.

[HIL 07] Hillard, D. Purpura, S., Wilkerson, J. (2007) "*An active learning framework for classifying political text*", In Annual Meeting of the Midwest Political Science Association. Chicago.

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

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

[PAN 08] Pang, B. & Lee, L. (2008) "*Opinion mining and sentiment analysis*", Foundations and Trends in Information Retrieval, 2(1-2):1–135.

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

[PLA 98] Platt, J. (1998) "*Fast Training of Support Vector Machines using Sequential Minimal Optimization*", In: Advances in Kernel Methods - Support Vector Learning, Schoelkopf, B., Burges C. & Smola A. editors. MIT Press.

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

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

[SAL 88] Salton, G. & Buckley, C. (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*", In: Proceedings of International Conference on New Methods in Language Processing.

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

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

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