ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Unsupervised Feature Selection for Sparse Data

Artur Ferreira[1,3] Mário Figueiredo[2,3]

1- Instituto Superior de Engenharia de Lisboa, Lisboa, PORTUGAL
2- Instituto Superior Técnico, Lisboa, PORTUGAL
3- Instituto de Telecomunicações, Lisboa, PORTUGAL

Abstract. Feature selection is a well-known problem in machine learning and pattern recognition. Many high-dimensional datasets are sparse, that is, many features have zero value. In some cases, we do not known the class label for some (or even all) patterns in the dataset, leading us to semi-supervised or unsupervised learning problems.

For instance, in text classification with the bag-of-words (BoW) representations, there is usually a large number of features, many of which may be irrelevant (or even detrimental) for categorization tasks. In this paper, we propose one efficient unsupervised feature selection technique for sparse data, suitable for both standard floating point and binary features.

The experimental results on standard datasets show that the proposed method yields efficient feature selection, reducing the number of features while simultaneously improving the classification accuracy.

1 Introduction

The need for *feature selection* (FS) is a well-known fact, since it arises in many machine learning and pattern recognition problems [1]. For instance, in text classification based on the *bag-of-words* (BoW) [2] each document is represented by high dimensional sparse vectors with the frequencies of a set of terms in each text; we usually have a large number of features, many of which are irrelevant, redundant or even harmful for classification performance. In this context, the need for FS techniques arises; these techniques may improve the accuracy of a classifier (avoiding the "curse of dimensionality") and speed up the training process [1]. There is a vast literature on FS techniques; see, for instance, [1, 3, 4] for detailed analysis of FS techniques and many references.

1.1 Feature Selection in Text Categorization

In text classification tasks, each document is typically represented by a BoW [2], which is a high-dimensional vector with the relative frequencies of a set of terms in each document. A collection of documents is represented by the *term-document* (TD) [5] matrix whose columns hold the BoW representation for each document whereas its rows correspond to the terms in the collection. An alternative representation for a collection of documents is provided by the (binary) *term-document incidence* (TDI) matrix [5]; this matrix holds the information, for each document, if a given term (word) is present or absent.

Both these matrices usually have a large number of rows (i.e., terms/features); consequently, representing large collections of documents is expensive in terms of memory. Feature selection contributes to alleviating this problem, in addition to often improving the performance of the learning algorithm being applied [3].

For supervised and unsupervised text categorization tasks, several techniques have been proposed for FS (see [6, 7, 8]). The majority of these techniques is applied directly on the TD matrix with floating point BoW representations.

1.2 Our Contribution

In this paper, we propose one efficient unsupervised method for FS on sparse numeric floating point and binary features. We use a filter approach, which makes the method independent of the type of classifier considered.

The remaining text is organized as follows. Section 2 reviews document representations and supervised and unsupervised FS techniques. Section 3 presents the proposed unsupervised method for FS. Section 4 provides the experimental evaluation of our method, compared against other supervised and unsupervised methods. Finally, Section 5 ends the paper with some concluding remarks.

2 Background

This section briefly reviews some background concepts regarding document representations and supervised and unsupervised FS techniques.

2.1 Document Representation

Let $D = \{(\mathbf{x}_1, c_1), ..., (\mathbf{x}_n, c_n)\}$ be a labeled dataset with training and test subsets, where $\mathbf{x}_i \in \mathbb{R}^p$ denotes the i-th (BoW-type) feature vector and c_i is its class label. The BoW representation contains some measure, like the *term-frequency* (TF) or the *term-frequency inverse-document-frequency* (TF-IDF) of a term (or word) [2]. Since each document only contains a small subset of terms, the BoW vector is usually sparse [2].

Let $\mathbf{X}$ be the $p \times n$ *term-document* (TD) matrix representing D; each column of $\mathbf{X}$ corresponds to a document, whereas each row corresponds to a term (e.g., a word); each column is the BoW representation of a document [2, 5]. Let $\mathbf{X_b}$ be the corresponding $p \times n$ *term-document-incidence* (TDI) matrix, with binary entries, such that a 1 at line i, column j means that term i occurs in document j, whereas a 0 means that it does not occur [5]. TDI matrices are an adequate solution to represent very large collections of documents, due to the low memory requirements. The TDI matrix can be trivially obtained from the TD matrix.

2.2 Unsupervised and supervised feature selection

A comprehensive listing of FS techniques can be found in [1, 3, 4]. In this Subsection, we briefly describe three unsupervised FS techniques that have been proven effective for classification on sparse data. On text categorization problems, these techniques have been applied directly on the TD matrix, achieving considerable dimension reduction and improving classification accuracy.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

The unsupervised *term-variance* (TV) [9] method selects features (terms) X_i by their variance, given by

$$\mathrm{TV}_i = \mathrm{var}_i = \frac{1}{n}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2 = \mathbf{E}[X_i^2] - \bar{X_i}^2, \tag{1}$$

where $\bar{X}_i$ is the average value of feature X_i, and n is the number of patterns.

The supervised *minimum redundancy maximum relevancy* (mrMR) method [10] computes both the redundancy and the relevance of each feature. Redundancy is the *mutual information* (MI) [11] between pairs of features, whereas relevance is measured by the MI between features and class label.

The supervised *Fisher index* (FI) of each feature, on binary classification problems, is given by

$$\mathrm{FI}_i = \left|\mu_i^{(-1)} - \mu_i^{(+1)}\right|^2 / (\mathrm{var}_i^{(-1)} + \mathrm{var}_i^{(+1)}), \tag{2}$$

where $\mu_i^{(\pm1)}$ and $\mathrm{var}_i^{(\pm1)}$, are the mean and variance of feature i, for the patterns of each class. The FI measures how well each feature separates the two (or more, since it can be generalized) classes. Since these three methods are filter approaches, to perform FS we select the m $(\leq p)$ features with the largest rank.

3 Proposed Unsupervised Method

In this section we present the proposed method for unsupervised FS, which can be applied to floating point and binary representations (equivalently, the TD or TDI matrices). It relies on the idea that for sparse data, a feature has an importance/relevance proportional to its dispersion [9].

We use the generic dispersion of the values of each feature, instead of the dispersion around the mean, like in the TV method. We propose to measure this dispersion with the *arithmetic mean* (AM) and the *geometric mean* (GM). For a given feature X_i on n patterns, the AM and the GM are

$$\mathrm{AM}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij} \qquad \text{and} \qquad \mathrm{GM}_i = \left(\prod_{j=1}^{n} X_{ij}\right)^{\frac{1}{n}}, \tag{3}$$

respectively; it is well known that $\mathrm{AM}_i \geq \mathrm{GM}_i$, with equality holding if and only if $X_{i1} = X_{i2} = ... = X_{in}$. The relevance criterion for each feature i is $\mathrm{R}_i = \mathrm{AM}_i/\mathrm{GM}_i$, with $R_i \in [1, +\infty)$. However if a given feature has at least one zero occurrence, we have zero GM making this criterion useless. To overcome this problem, we apply the exponential function to each feature, yielding

$$\mathrm{R}'_i = \mathrm{AM'}_i/\mathrm{GM}'_i = \underbrace{\frac{1}{n}\sum_{j=1}^{n}\exp(X_{ij})}_{a} \, / \, \underbrace{\left(\exp\left(\sum_{j=1}^{n} X_{ij}\right)\right)^{\frac{1}{n}}}_{g}, \tag{4}$$

with $0 < g \leq a$. If we take the logarithm of (4), we keep the ranking of the set of features; after some algebraic manipulation and dropping constant terms, we obtain our *feature dispersion* (FD) criterion

$$\mathrm{FD}_i = \log(a/g) = \log\left(\sum_{j=1}^{n} \exp(X_{ij})\right) - \frac{1}{n}\sum_{j=1}^{n} X_{ij}. \tag{5}$$

In many cases each BoW feature has values close to zero; in this case if we use the linear approximation $\exp(x) \approx 1 + x$ and for convenience we define $S_i = \sum_{j=1}^{n} X_{ij}$, the expression in (5) becomes

$$\mathrm{FD}_i \approx \log(n + S_i) - \frac{1}{n}S_i, \tag{6}$$

being quite efficient to compute (we have the sum of each feature). In the special case of TDI matrices S_i becomes the ℓ_0 norm (the number of non-zero entries).

4 Experimental Evaluation

This section reports the experimental results of our technique for text classification based on both standard and binary BoW representations. Our methods are evaluated by the test set error rate obtained by linear *support vector machine* (SVM) classifier, provided by the PRTools [12][1] toolbox. We use the SpamBase and the Dexter datasets from the UCI Repository[2]; the SpamBase task is to classify email messages as SPAM or non-SPAM. For Dexter, the task is to classify Reuters articles as being about "corporate acquisitions" or not.

Table 1: Standard Datasets SpamBase and Dexter. p and n are the number of features and patterns, respectively. $\overline{\ell_{0p}}$ and $\overline{\ell_{0n}}$ are the average ℓ_0 norm of each feature and pattern, respectively. $(-1, +1)$ are the number of patterns per class.

Dataset	p	$\overline{\ell_{0p}}$	$\overline{\ell_{0n}}$	**Partition**	n	$(\mathbf{-1}, \mathbf{+1})$
SpamBase	54	841.2	9.8	—	4601	(1813,2788)
Dexter	20000	1.4	94.1	Train	300	(150,150)
				Test	2000	(1000,1000)
				Valid.	300	(150,150)

In the SpamBase dataset, we have used the first 54 features, which constitute a BoW. The Dexter dataset has 10053 additional distractor features (independent of the class), at random locations, and was created for the NIPS 2003 FS challenge[3]. We train with a random subset of 200 patterns and evaluate on the validation set, since the labels for the test set are not publicly available; the results on the validation set correlate well with the results on the test set [3].

[1] http://www.prtools.org/prtools.html

[2] http://archive.ics.uci.edu/ml/datasets.html

[3] http://www.nipsfsc.ecs.soton.ac.uk

We evaluate the test set error rate on the SpamBase and Dexter datasets, using TD and TDI representations, respectively. The reported results are averages over ten replications of training/testing partition.

Fig. 1 shows the average test set error rates of the linear SVM classifier for the SpamBase dataset (TD and TDI matrices) using supervised and unsupervised FS methods, as functions of the number of features m. The horizontal dashed line corresponds to the classifier trained without FS (baseline error). We see that the proposed method, FD, is able to perform better than the supervised methods. Fig. 2 shows the average test set error rates of the linear SVM classifier for the Dexter dataset on TD and TDI matrices, respectively. On the TD matrix, our FD method has about the same test error rate as TV; on the TDI matrix we get lower test error rate than TV.

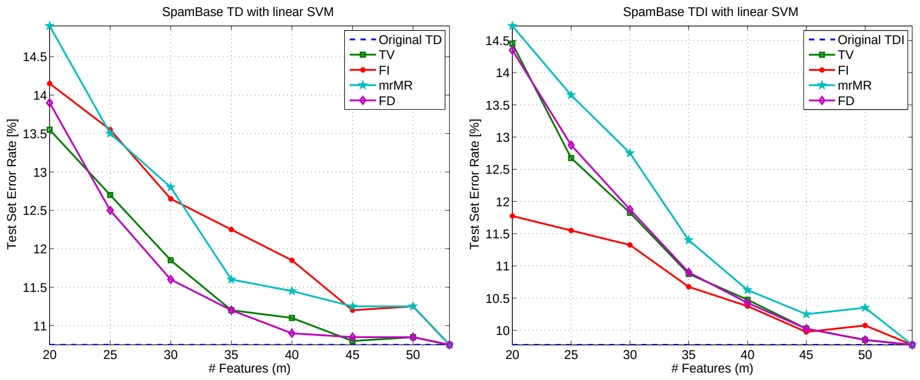

Fig. 1: Test set error rates for SpamBase dataset, with TD and TDI matrices, for $20 \leq m \leq 54$ (average of ten train/test replications).

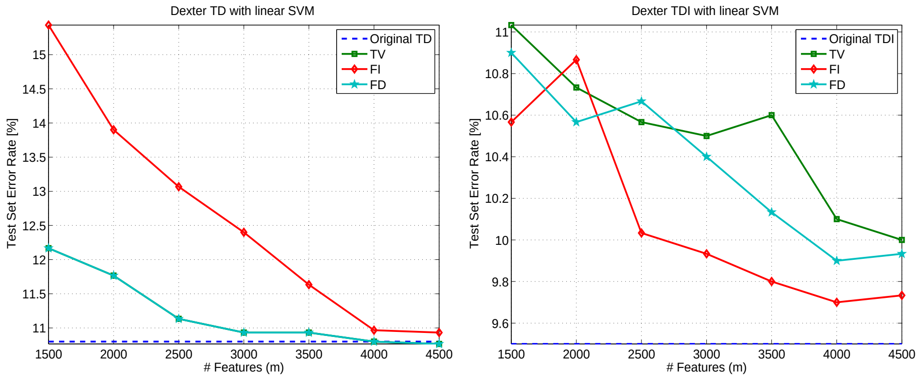

Fig. 2: Test set error rates for Dexter dataset, with TD and TDI matrices, for $1500 \leq m \leq 4500$ (average of ten train/test replications).

5 Conclusions

In this paper, we have proposed one efficient unsupervised method for feature selection on sparse data, based on simple analysis of the input training data, regardless of the label of each pattern. We compared our method with supervised and unsupervised techniques, on standard datasets with floating point and binary features. The experimental results show that the proposed method works equally well on both types of features reducing the number of features and improving classification accuracy, performing better than supervised feature selection methods in some cases. These methods can be applied to multi-class problems without modification, as we intend to do in future work.

Acknowledgements
This work was partially supported by the *Fundação para a Ciência e Tecnologia* (FCT), under grants PTDC/EEA-TEL/72572/2006 and SFRH/BD/45176/2008.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, 2nd edition, 2009.

[2] T. Joachims. *Learning to Classify Text Using Support Vector Machines.* Kluwer Academic Publishers, 2001.

[3] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh (Editors). *Feature Extraction, Foundations and Applications.* Springer, 2006.

[4] F. Escolano, P. Suau, and B. Bonev. *Information Theory in Computer Vision and Pattern Recognition.* Springer, 2009.

[5] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[6] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.

[7] K. Hyunsoo, P. Howland, and H. Park. Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 6:37–53, 2005.

[8] K. Torkkola. Discriminative features for text document classification. *Pattern Analysis and Applications*, 6(4):301–308, 2003.

[9] L. Liu, J. Kang, J. Yu, and Z. Wang. A comparative study on unsupervised feature selection methods for text clustering. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pages 597–601, 2005.

[10] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, August 2005.

[11] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley, 1991.

[12] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. Ridder, D. Tax, and S. Verzakov. PRTools4.1, a Matlab Toolbox for Pattern Recognition. Technical report, Delft University of Technology, 2007.