i-ETC: ISEL Academic Journal of Electronics,
Telecommunications and Computers
CETC2011 Issue, Vol. 2, n. 1 (2013) ID-1

COMPRESSED LEARNING FOR TEXT CATEGORIZATION

Artur Ferreira[1,2] Mário Figueiredo[2,3]
[1]*Instituto Superior de Engenharia de Lisboa*
[2]*Instituto de Telecomunicações*
[3]*Instituto Superior Técnico, Lisboa, PORTUGAL*
arturj@isel.pt *mario.figueiredo@lx.it.pt*

Keywords: random projections, random subspaces, compressed learning, text classification, support vector machines.

Abstract: In text classification based on the *bag-of-words* (BoW) or similar representations, we usually have a large number of features, many of which are irrelevant (or even detrimental) for classification tasks. Recent results show that *compressed learning* (CL), i.e., learning in a domain of reduced dimensionality obtained by *random projections* (RP), is possible, and theoretical bounds on the test set error rate have been shown. In this work, we assess the performance of CL, based on RP of BoW representations for text classification. Our experimental results show that CL significantly reduces the number of features and the training time, while simultaneously improving the classification accuracy. Rather than the mild decrease in accuracy upper bounded by the theory, we actually find an increase of accuracy. Our approach is further compared against two techniques, namely the unsupervised random subspaces method and the supervised Fisher index. The CL approach is suited for unsupervised or semi-supervised learning, without any modification, since it does not use the class labels.

1 INTRODUCTION

The need for *feature selection* (FS) and/or *feature reduction* (FR) arises in many machine learning and pattern recognition problems [1]. On large datasets (in terms of dimension and/or number of samples), the use of search-based or wrapper techniques can be computationally prohibitive.

For instance, in text classification based on the *bag-of-words* (BoW) or similar representations (where texts are represented by high dimensional 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 (or even harmful) for the classification task in hand. In this context, FS and FR play important roles in reducing the number of features. The use of FS or FR techniques may improve the accuracy of a classifier (avoiding the "curse of dimensionality") and speeds up the training process [1]. The literature on FS and FR is too vast to be reviewed here. Comprehensive coverage of these techniques, and pointers to a vast literature, can be found in several books, namely [1], [2], [3], and [4].

1.1 Compressed Learning

In the past decade, there has been some interest in random projections (RP, see [5, 6, 7] and references therein) for FR. Recently, theoretic support for RP-based FR (termed *compressed learning* – CL) was provided in [8]. CL is inspired by the *compressed sensing* (CS) [9, 10] framework, in which an RP matrix is used to map from the data domain to the measurement domain. The theory of CS provides conditions (on the projection matrix and the level of sparseness of the data vectors) under which this (non-injective) mapping can be inverted. Some CS-based techniques for classification, based on RP, have been proposed [11]. Recently, it was shown that *compressed learning* (CL) is possible [8]; specifically, it was proved that learning in the compressed domain is guaranteed to be, in the worst case, only slightly worse than learning on the original data domain, if the RP matrix satisfies some conditions and the feature vectors are sparse (possibly on some unknown basis). Since BoW representations are usually very sparse, text classification using this type of representation seems like an obvious candidate for CL.

1.2 Our Contribution

In this work, we assess CL for BoW-based text classification using linear *support vector machines* (SVM) and several types of RP matrices. SVM have been found very effective for BoW-based text classification [12]. As shown in our experimental results, the classifiers obtained via CL, on significantly reduced dimensions, exhibit improved classification accuracy with respect to the classifiers trained on the original features. These results suggest that (for this type of data) the bound for CL given in [8] is pessimistic: instead of the mild decrease in accuracy upper bounded by the theory in [8], we actually find an increase of accuracy.

The remaining text is organized as follows. Section 2 briefly reviews the main results of CL theory and RP-based FR techniques. Some FR and FS techniques, used for benchmark purposes are described in Section 3. The experimental setup and the experimental results are described in Section 4. Section 5 ends the paper with some concluding remarks.

2 COMPRESSED LEARNING

2.1 Projection-Based Dimension Reduction

Let $D = \{(\mathbf{x}_1, c_1), ..., (\mathbf{x}_p, c_p)\}$ be a labeled dataset, where $\mathbf{x}_i \in \mathbb{R}^n$ denotes the i-th feature vector and $c_i \in \{-1, +1\}$ is its class label. Letting $\mathbf{A}$ be an $m \times n$ matrix, with $m < n$, we obtain a *reduced/compressed* training dataset $D_{\mathbf{A}} = \{(\mathbf{y}_1, c_1), ..., (\mathbf{y}_p, c_p)\}$ via

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i. \tag{1}$$

Each new feature (component of $\mathbf{y} = \mathbf{A}\mathbf{x}$) is a linear combination of the original features. Many techniques have been proposed to obtain "good" (in some sense) projection matrices.

In the case of *random projections* (RP), the entries of $\mathbf{A}$ are randomly generated. For reasons explained below, the following distributions yield good RP matrices:

- (i) Gaussian $\mathcal{N}(0, 1/\sqrt{m})$;
- (ii) Bernoulli over $\pm 1/\sqrt{m}$ with equal probability;
- (iii) probability mass function $\{1/6,\ 2/3,\ 1/6\}$ over $\{-\sqrt{3/m}, 0, \sqrt{3/m}\}$, proposed by Achlioptas [5];
- (iv) probability mass function $\{1/(2s), 1 - 1/s, 1/(2s)\}$ over $\{-\sqrt{s/m}, 0, \sqrt{s/m}\}$, proposed by Li et al [7].

Notice that the Bernoulli and Achlioptas matrices are particular cases of (iv), with $s = 1$ and $s = 3$, respectively. Choices (iii) and (iv) lead to sparse $\mathbf{A}$, which may be interesting from a computational point of view.

2.2 Restricted Isometry Properties

The use of RP is inspired by the Johnson-Lindenstrauss lemma [13, 14], which states that, under some conditions, a set of points in a high-dimensional space can be mapped down to a much lower dimensional space, such that the Euclidean distances between these points are approximately preserved. A closely related concept is that of *restricted isometry property* (RIP) [9, 10, 13]: a $m \times n$ matrix is said to satisfy the (k, ε)-RIP if for any k-sparse vector $\mathbf{x}$ (up to k non-zeros),

$$(1-\varepsilon)\|\mathbf{x}\|^2 \leq \|\mathbf{A}\mathbf{x}\|^2 \leq (1+\varepsilon)\|\mathbf{x}\|^2, \tag{2}$$

where $\|\cdot\|$ is the ℓ_2 norm. The random generation procedures described in Subsection 2.1 are known to yield matrices satisfying the RIP with small ε, with overwhelming probability, if

$$m = \Omega(k \log(n/k)), \tag{3}$$

that is, m is a (small) factor of Ω and needs to grow only logarithmically with the dimensionality of the input patterns n.

As is well-known, linear SVM classifiers use only inner products between training patterns. It is thus clear that a good RP matrix should preserve these inner products, a stronger requirement than the RIP. The *generalized RIP* (GRIP) gives conditions under which the inner products are approximately preserved [8] (see also [15]).

Lemma 1 ([8]) *: Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a matrix satisfying the $(2k, \varepsilon)$-RIP and $\mathbf{x}$ and $\mathbf{x}'$ be two k-sparse vectors such that $\|\mathbf{x}\|, \|\mathbf{x}'\| \leq R$. Then, letting $\mathbf{y} = \mathbf{A}\mathbf{x}$ and $\mathbf{y}' = \mathbf{A}\mathbf{x}'$,*

$$(1+\varepsilon)\mathbf{x}^T\mathbf{x}' - 2R^2\varepsilon \leq \mathbf{y}^T\mathbf{y}' \leq (1-\varepsilon)\mathbf{x}^T\mathbf{x}' + 2R^2\varepsilon$$

This lemma suggests that, if the training patterns are k-sparse and $\mathbf{A}$ satisfies the $(2k, \varepsilon)$-RIP, a linear SVM learnt from the compressed patterns $\mathbf{Y}$ will be very similar to one obtained from the original patterns $\mathbf{X}$, as formalized in the compressed learning bound shown in [8]. The following subsection details the compressed learning bound which is the main motivation for this work.

2.3 Compressed Learning Bound

Theorem 1 ([8]) *: Let $\mathbf{w}_0$ be the best linear classifier in the original data domain, with low expected hinge loss $H(\mathbf{w}_0) = \mathbb{E}_{\mathcal{D}}\left[1 - c\,\mathbf{w}_0^T\mathbf{x}\right]$, where the expectation is with respect to an unknown distribution $\mathcal{D}$ from which the training set D is assumed to have been generated. Let $\mathbf{z_A}$ be the soft-margin SVM trained on the compressed dataset $D_{\mathbf{A}}$. Then, with probability $1-2\eta$,*

$$H(\mathbf{z_A}) \leq H(\mathbf{w}_0) + O\left(\sqrt{\|\mathbf{w}_0\|^2\left(R^2\varepsilon + \frac{\log\frac{1}{\eta}}{p}\right)}\right), \tag{4}$$

where R is as defined in Lemma 1.

This theorem shows that the SVM obtained from the compressed data is never much worse than the best linear classifier in the original high dimensional space.

3 FEATURE REDUCTION AND SELECTION

This section presents some FR and FS techniques, used as benchmark for comparison purposes with the RP-based techniques.

3.1 Random Subspaces

The (unsupervised) *random subspaces method* (RSM) [16] acts by (pseudo) randomly selecting subsets of components of the feature vector, that is, it choosing axes-aligned random subspaces. A multivariate approach for FS based on RSM has been proposed [17]. The authors apply a multivariate search technique on a randomly selected subspace from the original feature space, to better handle the noise in the data on the reduced dimensionality domain. This procedure is repeated many times and the q chosen feature subsets are combined into a final list of selected features, used to train some classifier.

3.2 Fisher ratio

The well-known (supervised) *Fisher ratio* (FiR) for binary problems (*e.g.*, where $c_i \in \{0,1\}$) of each feature is defined as

$$\mathrm{FiR}_i = \frac{\left|\overline{X}_i^{(0)} - \overline{X}_i^{(1)}\right|}{\sqrt{\mathrm{var}(X_i)^{(0)} + \mathrm{var}(X_i)^{(1)}}}, \tag{5}$$

where $\overline{X}_i^{(0)}$, $\overline{X}_i^{(1)}$, $\mathrm{var}(X_i)^{(0)}$, and $\mathrm{var}(X_i)^{(1)}$, are the sample means and variances of feature *i*, for the patterns of each class. The FiR measures how well each feature alone separates the two classes [18].

4 EXPERIMENTS

In this paper, we report a set of experiments on CL for text classification based on BoW representations. As mentioned above, BoW representations are usually very sparse, thus being in favorable conditions for the applicability of CL via (1).

4.1 Experimental Setup

We consider the four RP matrices described in Subsection 2.1, which we will refer to as: Gaussian, Bernoulli, Achlioptas, and Li et al (with $s = n$). We use linear SVM classifiers, provided by the ENTOOL[1] toolbox, trained up to 20000 iterations. Each input pattern is normalized to unitary ℓ_2 norm (original domain).

We have used the following four (publicly available) BoW datasets: Example1[2], Example2[2], Dexter[3], and SpamBase[4].

These datasets have undergone the standard preprocessing (stop-word removal, stemming). Table 1 shows the main characteristics of these datasets,as discussed in Sub-sections 2.2 and 2.3 [8]:

- $\bar{k}$ is the average ℓ_0 norm of each pattern;
- $\hat{m}_R$ is an estimate of m to satisfy the (k,ε)-RIP condition, given by $\hat{m}_R = \Omega(\bar{k}\log(n/\bar{k}))$;
- $\hat{m}_G$ is an estimation of m to satisfy the $(2\underline{k},\varepsilon)$-RIP condition, given by $\hat{m}_G = \Omega(2\bar{k}\log(n/(2\bar{k})))$.

In the case of Example1, each pattern is a 9947-dimensional BoW vector. The classifier is trained on a random subset of 1000 patterns (500 per class) and tested on 600 patterns (300 per class). On Example2 dataset, we have 9930-dimensional BoW vectors, with only 10 training patterns.

The Dexter dataset has the same data as Example1 with 10053 additional distractor features with no predictive power at random locations, and was created for the NIPS 2003 feature selection challenge[5]. We train with a random subset of 200 patterns (100 per class) and evaluate on the validation set, since the

[1] zti.if.uj.edu.pl/~merkwirth/entool.htm
[2] download.joachims.org/svm_light/examples
[3] archive.ics.uci.edu/ml/datasets/dexter
[4] archive.ics.uci.edu/ml/datasets/SpamBasebase
[5] www.nipsfsc.ecs.soton.ac.uk

Table 1: Example1, Example2, Dexter, and SpamBase datasets main characteristics. $\bar{k}$ is the average ℓ_0 norm of each pattern. $\hat{m}_R$ and $\hat{m}_G$ are the estimates of m to satisfy the (k,ε)-RIP and the $(2k,\varepsilon)$-RIP conditions, respectively.

Dataset, n	**Subset**	**Patterns (+1,-1)**	$\bar{k}, \hat{m}_R, \hat{m}_G$
Example1, 9947	Train	2000 (1000,1000)	47.1 , 253, 436
	Test	600 (300,300)	39.5 , 219, 383
Example2, 9930	Train	10 (5,5)	46.6 , 247, 429
	Test	600 (300,300)	39.4 , 216, 376
Dexter, 20000	Train	300 (150,150)	94.1 , 505, 878
	Test	2000 (1000,1000)	96.2 , 514, 894
	Valid.	300 (150,150)	93.1 , 501, 872
SpamBase, 54	–	4601 (1813,2788)	9.8 , 17, 20

labels for the test set are not publicly available; the results on the validation set correlate well with the results on the test set. The task of both Example1 and Dexter is learn to classify Reuters articles as being about "corporate acquisitions" or not.

In the SpamBase dataset, we have used the first 54 features, which constitute a BoW. We have randomly selected 1000 patterns for training (500 per class) and 1000 (500 per class) for testing. The SpamBase task is to classify email messages as SPAM or non-SPAM.

A collection of BoW documents is usually represented by the *term-document* (TD) [19] matrix whose columns hold the BoW representation for each document whereas its rows correspond to the terms in the collection.

The reported results are averages over 10 replications of different training/testing partition and random matrices, except on the Example2 dataset in which we make no partition (the dataset has only 10 patterns).

To serve as a benchmark, we compare with both the unsupervised RSM and supervised Fisher Index procedures, described in Sub-section 3.

4.2 Test Set Error Rate

Figure 1 and Figure 2 show the average test set error rates (over 10 replications) for the Example1 and Example2 datasets, as functions of the number of features m. The horizontal dashed blue line corresponds to the classifiers trained on the original data. Figure 3 shows the error rate on the validation set, for the Dexter dataset. Finally, in order to assess the performance on lower dimensional sparse datasets, we compute the test set error rate on the SpamBase dataset; Figure 4 plots these results.

On the first three datasets we have an improvement on the error rate, after FR with any of the four probability distributions except for the Li et al. distribution on Example2. Typically Achlioptas distribution leads to lower test set error rate than Gaussian and Rademacher matrices, with about 1/3 non-zero

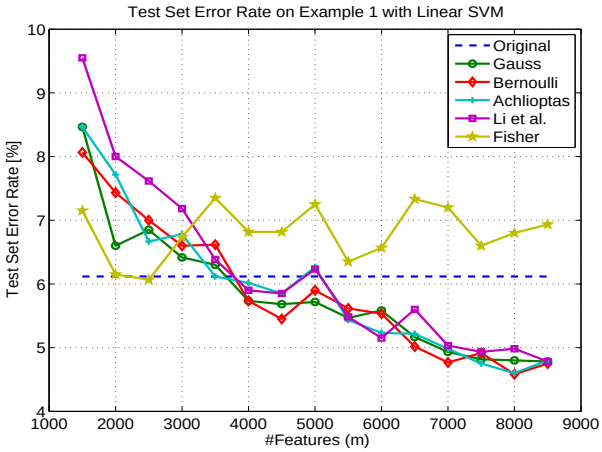

Figure 1: Average test set error rates (ten runs with different train/test partitions) for the Example1 dataset of the linear SVM classifier for FR based on RP and FS with Fisher Index.

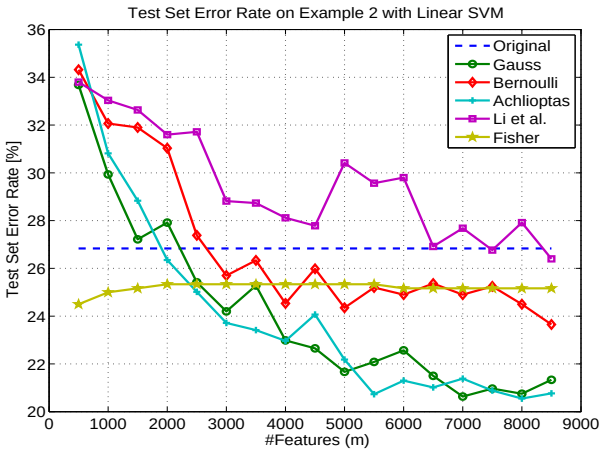

Figure 2: Average test set error rates (ten runs with different train/test partitions) for the Example2 dataset of the linear SVM classifier for FR based on RP and FS with Fisher Index.

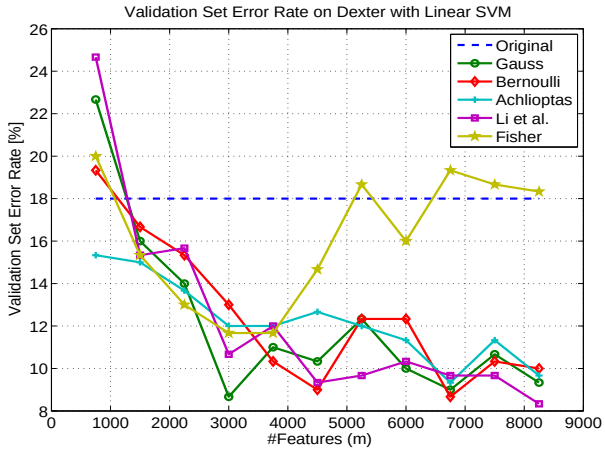

Figure 3: Average validation set error rates (ten runs with different train/test partitions) for the Dexter dataset of the linear SVM classifier for FR based on RP and FS with Fisher Index.

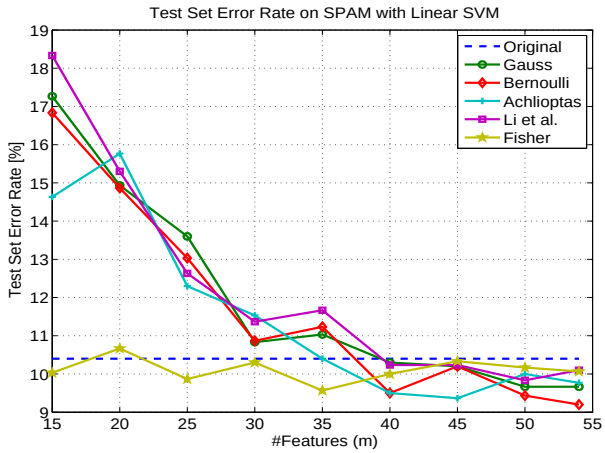

Figure 4: Average test set error rates (ten runs with different train/test partitions) for the SpamBase dataset of the linear SVM classifier for FR based on RP and FS with Fisher Index.

entries; the use of this distribution is efficient because the input patterns are also sparse and most of the point by point products can be avoided.

The upper bound for CL defined in (4) is conservative when applied to text classification problems. The test set error rate on the reduced domain is below the test set error rate on the original data domain. Moreover, an adequate value of m to achieve lower test set error than on the original domain is about $2\hat{m}_G$ to $5\hat{m}_G$, as shown in Table 1. On the SpamBase dataset, we get improvement with about $m \geq 40$.

On Figure 5 we compare the performance of RP-based methods with the RSM method, combining $q = 20$ different subspaces of features. We have the average test set error rates for the Dexter dataset, as functions of the number of features m. The horizontal dashed blue line corresponds to the linear SVM classifier trained on the original data with n features, which we call the baseline error. The vertical line corresponds to the m_G estimate, that is, the smallest value of m that satisfies the GRIP condition. The RP methods lead to features that are slightly better than those obtained by RSM. As compared to the RSM method, the RP method has the advantage to be faster, since it involves solely a matrix multiplication.

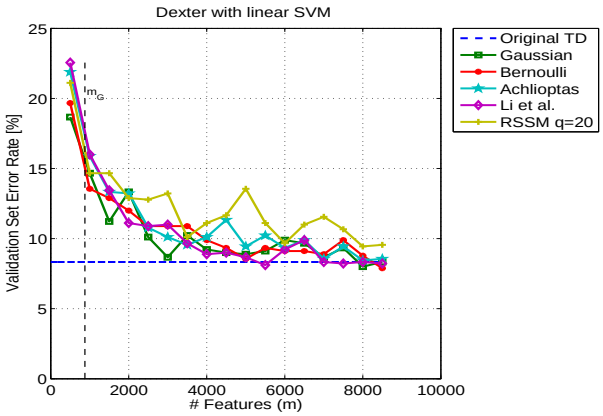

Figure 5: Validation set error rates (average over 10 random training/test partitions) for the Dexter dataset, with a linear SVM classifier, as functions of the number of features, using random projections and random subspaces (with q=20 subspaces).

4.3 Training Time Analysis

The reduction in the number of features leads to a reduction of the training time. Table 2 shows how the training time of the linear SVM varies with the number of features for the high-dimensional Dexter dataset. The decrease in the dimensionality of the data leads to reasonable improvements on the training time.

Table 2: Analysis of the training time (in seconds) for the Dexter dataset as a function of the number of features m, using RP (average of ten runs with different training/test partitions).

m	**Time [sec]**
20000	3.16
5000	2.41
4000	1.99
3000	1.85
2000	1.81

5 CONCLUSIONS

In this paper, we have reported a set of experiments on compressed learning for text classification based on (sparse) bag-of-words representations. The compressed features are obtained by (sparse) random projections. Our experimental results show that four probability distributions (two of which are sparse) for the random projection matrix significantly reduce the number of features as well as the training time, while also improving the classification accuracy.

We have found that the recently proved theoretical bound for the test error of compressed learning is conservative; the test set error on the compressed domain is below the test set error on the original data domain. The number of reduced dimensions can be computed by a simple sparsity analysis of the training data, regardless of the label of each pattern. The random projection technique performs slightly better than the random subspaces method and the Fisher index technique, on high-dimensional datasets.

In future work, we will apply this technique to semi-supervised text classification.

REFERENCES

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

[2] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[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] D. Achlioptas. Database-friendly random projections. In *ACM Symposium on Principles of Database Systems*, pages 274–281, Santa Barbara, California, 2001.

[6] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proc. 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining – KDD'01*, pages 245–250, New York, 2001.

[7] P. Li, T. Hastie, and K. Church. Very sparse random projections. In *KDD '06: Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 287–296, New York, 2006.

[8] R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. *[Online] http://dsp.rice.edu/files/cs/cl.pdf*, 2009.

[9] E. Candes, J. Romberg, and T. Tao. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

[10] D. Donoho. Compressed Sensing. *IEEE Transactions on Information theory*, 52(4):1289–1306, 2006.

[11] M. Duarte, M. Davenport, M. Wakin, J. Laska, D. Takhar, K. Kelly, and R. Baraniuk. Multiscale random projections for compressive classification. In *IEEE Int. Conf. on Image Processing (ICIP)*, 2007.

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

[13] R. Baraniuk, M. Davenport, R. Devore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 2007.

[14] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conf. in Modern Analysis and Probability*, pages 189–206, 1984.

[15] J. Haupt and R. Nowak. A generalized restricted isometry property. Technical report, University of Wisconsin, Madison, 2007.

[16] T. Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Analysis Machine Intelligence*, 20(8):832–844, 1998.

[17] Carmen Lai, Marcel Reinders, and Lodewyk Wessels. Random subspace method for multivariate feature selection. *Pattern Recognition Letters*, 27(10):1067–1076, 2006.

[18] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 2000.

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