Discriminative Features for Text Document Classification

Kari Torkkola

Motorola Labs, 2900 South Diablo Way,

MD DW286, Tempe AZ 85282, USA

tel: (602) 6596620, fax: (602) 659-6662

email: Kari.Torkkola@motorola.com

Discriminative Features for Text Document Classification

Abstract

The bag-of-words approach to text document representation typically results in vectors of the order of 5000 to 20000 components as the representation of documents. In order to make effective use of various statistical classifiers, it may be necessary to reduce the dimensionality of this representation. We point out deficiencies in class discrimination of two popular such methods, Latent Semantic Indexing (LSI), and sequential feature selection according to some relevant criterion. As a remedy, we suggest feature transforms based on Linear Discriminant Analysis (LDA). Since LDA requires operating both with large and dense matrices, we propose an efficient intermediate dimension reduction step using either a random transform or LSI. We report good classification results with the combined feature transform on a subset of the Reuters-21578 database. Drastic reduction of the feature vector dimensionality from 5000 to 12 actually improves the classification performance.

Keywords

Text classification, Dimension reduction, Linear discriminant analysis, Random transforms

Originality and Contribution

Linear Discriminant Analysis has its origins in Fisher's work dating back to 1936. However, application of LDA to very high dimensional data, such as to the popular bag-of-words text document representation is difficult (and thus absent in the literature), because the LDA criterion matrix becomes a large and dense matrix, even if the original representation may be sparse. This paper shows that LDA is still applicable to high dimensional data by introducing an intermediate transform, either a random transform, or LSI, which can be effectively computed from large sparse matrices. This enables a feature vector dimension reduction of 2-3 orders of magnitude without sacrificing accuracy. Faster and more accurate on-line text classification/categorization systems become now possible, which may be especially useful in information retrieval and search tasks.

1 Introduction

Classification of textual documents denotes assigning an unknown document to one of predefined classes. This is a straightforward concept from pattern recognition or from supervised machine learning. It implies the existence of a labeled training data set, a way to represent the documents, and a statistical classifier trained using the chosen representation of the training set.

Some classifiers are very sensitive to the representation, for example, failing to generalize to unseen data (over-fitting) if the representation contains too much irrelevant information [2]. It would thus be advantageous to be able to extract only that information which is pertinent to the classification task. However, some classifiers, such as Support Vector Machines [17], are able to tolerate better the existence of irrelevant information. Either case, in general, it is computationally cheaper to operate the classifier in low dimensional spaces. If this can be done without sacrificing the accuracy of the classifier, the better.

We present a fairly straightforward application of Linear Discriminant Analysis (LDA) to document classification, when vector space document representations are employed. LDA is a well known method in statistical pattern recognition literature. Its aim is to learn a *discriminative* transformation matrix from the original high-dimensional space to a desired dimensionality [11]. The idea is to project the documents into a low dimensional space in which the classes are well separated. This can also be viewed as extracting features that only carry information pertinent to the classification task. A subsequent classification task should then become easier.

This paper proceeds as follows. We discuss document representation methods and approaches to reduce their dimensionality, especially Latent Semantic Indexing (LSI) [7]. We discuss why LSI cannot result in optimal representation for document classification. Methods to select a number of relevant features, as well as their shortcomings are then discussed. We introduce LDA,

and we present two avenues to solve the difficulties related to processing large and dense matrices that arise in applying LDA to very high dimensional data. Classification experiments using derived LDA-features with a support vector machine classifier on the Reuters-21578 database are presented. We discuss the computational complexity of the approach, and possibilities to enhance the interpretation of the derived basis vectors.

The point of view of this paper is that of statistical pattern recognition. This means that we approach the problem as a classification task with exclusive classes aiming to maximize the classification accuracy. Performance is evaluated by assigning a single label to each unknown document. This label can be either correct or incorrect, and the accuracy is defined as the number of correctly assigned labels divided by the total number of documents in the test set. This is not the usual case in information retrieval where documents may carry several labels (or cover several topics). Given a document collection, the aim is to retrieve all documents relevant to a particular topic or class. Performance is measured as precision and recall [15], that cannot be related to the mere classification accuracy. Thus, in the last section of the paper we also discuss extensions to information retrieval tasks.

2 Vector Space Document Representations

The dominant vectorial text document representation is based on the so called bag-of-words approach, in which each document is essentially represented as a histogram of terms, or as a function of the histogram. One straightforward function is normalization: the histograms are divided by the number of terms of the document to account for different document lengths.

What are the terms that are counted in the histograms? Terms (words) that occur in every document obviously do not convey much useful information for classification. Same applies to

rare terms that are found only in a few documents. These, as well as common stop words, are usually filtered out of the corpus. Furthermore the words may be stemmed. These operations leave a term dictionary that can range in size from thousands to tens of thousands. Correspondingly, this is the dimension of the space in which documents now are represented as vectors. Although the dimension may be high, a characteristic of this representation is that the vectors are sparse.

For many statistical pattern classification methods this dimensionality may be too high. Thus dimension reduction methods are called for. Two possibilities exist, either selecting a subset of the original features, or transforming the features into new ones, that is, computing new features as some functions of the old ones. We examine both in turn.

3 Feature Selection

Optimal feature selection coupled with a pattern recognition system leads to a combinatorial problem since all combinations of available features need to be evaluated, by actually training and evaluating a classifier. This is called the *wrapper* configuration [21, 24]. Obviously the wrapper strategy does not allow to learn parametric feature transforms, such as linear projections, because all possible transforms cannot be enumerated.

Another approach is to evaluate some criterion related to the final classification error that would reflect the "importance" of a feature or a number of features jointly. This is called the *filter* configuration in feature selection [21, 24]. What would be an optimal criterion for this purpose? Such a criterion would naturally reflect the classification error rate. Approximations to the Bayes error rate can be used, based on Bhattacharyya bound or an interclass divergence criterion. However, these joint criteria are usually accompanied by a parametric, such as Gaussian, estimation of the multivariate densities at hand [13, 29], and are characterized by heavy computational demands.

In document classification problems, the dominant approach has been sequential greedy selection using various different criteria [33, 4, 25]. This is dictated by the sheer dimensionality of the document-term representation. However, greedy algorithms based on sequential feature selection using any criterion are suboptimal because they fail to find a feature set that would *jointly* optimize the criterion. For example, two features might both be very highly ranked by the criterion, but they may carry the same exact information about class discrimination, and are thus redundant.

Thus, feature selection through any joint criteria such as the actual classification error, leads to a combinatorial explosion in computation. For this very reason finding a *transform* to lower dimensions might be easier than selecting features, given an appropriate objective function.

4 Latent Semantic Indexing

One well known dimension reducing transform is the principal component analysis (PCA), also called Karhunen-Loeve transform. PCA seeks to optimally *represent* the data in a lower dimensional space in the mean squared error sense. The transform is derived from the eigenvectors corresponding to the largest eigenvalues of the covariance matrix of training data.

In the information retrieval community this method has been named Latent Semantic Indexing (or LSI) [7]. The covariance matrix of data in PCA corresponds now to the document-term matrix multiplied by its transpose[1]. Entries in the covariance matrix represent co-occurring terms in the documents. Eigenvectors of this matrix corresponding to the dominant eigenvalues are now directions related to dominant combinations of terms occurring in the corpus. These dominant combinations can be called “topics” or “semantic concepts”. A transform matrix constructed from these eigenvectors projects a document onto these “latent semantic concepts”, and the new low

[1]The only difference to PCA is that in PCA the mean of the data is removed in computing the covariance matrix.

dimensional representation consists of the magnitudes of these projections. The eigenanalysis can be computed efficiently by a sparse variant of singular value decomposition of the document-term matrix [7, 1].

Let $\mathbf{D}$ denote the $t \times d$ document-term matrix with rank r where each of the d columns represents a document vector of dimension t. The singular value decomposition results in

$$\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{1}$$

where $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix of singular values of $\mathbf{D}$, $\mathbf{U}$ is a $t \times r$ matrix of left singular column vectors, and $\mathbf{V}$ is a $d \times r$ matrix of right singular vectors of $\mathbf{D}$. Dropping all but k largest singular values and corresponding singular vectors gives the truncated approximation of $\mathbf{D}$

$$\mathbf{D}_k = \mathbf{U}_k\mathbf{\Sigma}_k{\mathbf{V}_k}^T. \tag{2}$$

The approximation is the best rank k approximation of $\mathbf{D}$ in the sense of minimizing the sum of squared differences between the entries of $\mathbf{D}$ and $\mathbf{D}_k$ (the Frobenius 2-norm).

The $t \times k$ matrix $\mathbf{U}_k$ can now be used as a projection matrix to map each subsequent $t \times 1$ document column vector $\mathbf{d}$ into a k-dimensional representation

$$\mathbf{d}_k = {\mathbf{U}_k}^T\mathbf{d}. \tag{3}$$

LSI was introduced to improve precision/recall, and it has been proven to be extremely useful in various information retrieval tasks. However, it is not an optimal representation for classification. LSI/PCA is completely unsupervised, that is, it pays no attention to the class labels of the existing training data. LSI aims at optimal *representation* of the original data in the lower dimensional space in the mean squared error sense. This representation has nothing to do with the optimal *discrimination* of the document classes.

Independent component analysis (ICA) has also been proposed as a tool to find "interesting" projections of the data [12, 32, 20]. Girolami et al. maximize negentropy to find a subspace on which the data has the least Gaussian projection [12]. The criterion corresponds to finding a clustered structure in the data, and bears a close relationship to projection pursuit methods [10]. This appears to be a very useful tool revealing non-Gaussian structures in the data. However, as PCA, the method is completely unsupervised with regard to the class labels of the data, and it is not able to enhance class separability.

Sammon mapping [28], multidimensional scaling (MDS) [22], and FastMap [9] are examples of further methods that map data points into a lower dimensional space trying to preserve the mutual distances in the original high-dimensional space. These methods pay no attention to class labels either, only to the goodness of the representation. Thus, *supervised* feature extraction schemes are called for. We describe one such method, Linear Discriminant Analysis.

5 Linear Discriminant Analysis

The term linear discriminant analysis (LDA) refers to two distinct but related methods. The first is classifier design. Given a number of variables as the data representation, each class is modeled as Gaussian (with a covariance matrix and a mean vector). Observations are now classified to the class of the nearest mean vector according to Mahalanobis distance. The decision surfaces between classes become linear if the classes have a shared covariance matrix. In this case the decision surfaces are called Fisher discriminants, and the procedure of constructing them is called Linear Discriminant Analysis [11, 2].

The second use of the term LDA refers to a discriminative feature transform that is optimal for certain cases [11]. This is what we denote by LDA throughout this paper. In the basic formulation,

LDA finds eigenvectors of matrix

$$\mathbf{T} = \mathbf{S}_w^{-1}\mathbf{S}_b. \tag{4}$$

Here $\mathbf{S}_b$ is the between-class covariance matrix, that is, the covariance matrix of class means. $\mathbf{S}_w$ denotes the within-class covariance matrix, that is equal to the weighted sum of covariance matrices computed for each class separately. $\mathbf{S}_w^{-1}$ captures the compactness of each class, and $\mathbf{S}_b$ represents the separation of the class means. Thus $\mathbf{T}$ captures both. The eigenvectors corresponding to largest k eigenvalues of $\mathbf{T}$ form the rows of the transform matrix $\mathbf{W}$, and new discriminative features $\mathbf{d}_k$ are derived from the original ones $\mathbf{d}$ simply by

$$\mathbf{d}_k = \mathbf{W}\mathbf{d}. \tag{5}$$

The relation to LDA as a classifier design is that these eigenvectors span the same space as directions orthogonal to the decision surfaces of Fisher discriminants.

The straightforward algebraic way of deriving the LDA transform matrix is both a strength and a weakness of the method. LDA makes use of only second-order statistical information, the covariances. Furthermore, LDA assumes that all classes have the same covariance, through using a shared within-class covariance matrix. Thus it is optimal for data in which each class has a unimodal Gaussian density with well separated means and similar covariances. Large deviations from these assumptions may result in sub-optimal features. Also the maximum rank of $\mathbf{S}_b$ in this formulation is $N_c - 1$, where N_c is the number of different classes. Thus basic LDA cannot produce more than $N_c - 1$ features. This is, however, simple to remedy by projecting the data onto a subspace orthogonal to the computed eigenvectors, and repeating the LDA analysis in this space [26].

Further extensions to LDA exist. For example, Heteroscedastic Discriminant Analysis (HDA) allows the classes have different covariances [23]. However, simple linear algebra is no longer

sufficient to compute the solution. One must resort to iterative optimization methods. Same applies to methods that further relax the Gaussianity assumptions of the classes [29, 31].

The following section discusses what difficulties arise applying LDA to high-dimensional document-term data, how those can be solved, and why LDA and the assumptions behind it are thereafter well suited to document-term data.

6 Linear Discriminant Analysis for Document-Term Data

To our knowledge, LDA feature transforms have not been applied earlier to document classification tasks, although LDA has been used before in the sense of designing a linear classifier [16, 30]. In contrast, we suggest LDA as a means of deriving efficient features, which can be classified by any, possibly nonlinear, classifier.

If LDA is such a well known and well-behaving method why is it not in wider use in the document analysis community? LDA is, of course, only applicable when labeled training data exists, and this represents only a part of possible document analysis tasks. Unsupervised tasks, such as clustering are thus excluded. However, nothing prevents applying LDA *after* clustering to find features that best separate the clusters, then repeating the clustering using new features, and iterating this a few times.

One high barrier in applying LDA directly to document-term data of tens of thousands dimensions is the following. Unlike the document-term matrix, the criterion matrix $\mathbf{T}$ in LDA is no longer sparse. Thus efficient methods for inversion, singular value decomposition, or eigenanalysis, of sparse matrices cannot be used. For example, if the number of terms $t = 20.000$, $\mathbf{T}$ requires memory of 3.2GB. Since inversion of a matrix requires computation $O(t^3)$, only the inversion of $\mathbf{S}_w$ would take approximately 12 hours of CPU time on a 700MHz machine (on which the inversion

of a 1000x1000 matrix takes 5.6 seconds).

However, a simple remedy exists. Random projections have been shown to be very useful in various dimension reduction tasks where the source data has had extremely high dimensionality [3, 27, 6]. Random projections tend to Gaussianize the data. This is because each resulting component is a sum of a large number of original components each multiplied by a random weight. In this respect, data after a random projection conforms well to the assumptions behind LDA. A straightforward method is thus to generate a random matrix with, say, normally distributed entries, to transform the original dimension down by an order of magnitude, thereafter followed by a conventional LDA transformation with one or two more orders of magnitude further dimension reduction.

Another convenient option is to perform a conventional LSI-based dimension reduction by an order of magnitude starting from the original document-term matrix. This is a feasible task because the document-term matrix is a sparse matrix, and efficient SVD methods can be used [1]. This initial transform can again be followed by LDA. This is illustrated in Fig. 1.

Since both types of suggested initial transforms, as well as the LDA transform, are linear, the two sequential transforms can be combined into a single matrix multiplication. The following section reports experiments using these approaches.

7 Document Classification Experiments

Experiments were performed using the Reuters-21578 database[2]. We used the ”ModLewis” split of the database into training and testing parts. Since the aim is classification, in which each document has an exclusive category, we discarded documents with no label or with multiple labels.

[2]http://www.research.att.com/ lewis/reuters21578.html

Furthermore, those rare classes were discarded that did not occur at least once in both training and testing set. The resulting training set has 6535 documents, and the test set 2570 documents with 52 document classes.

Granted, from the information retrieval standpoint this is an artificial task, which makes it difficult to compare our results to those of others. However, this facilitated a straightforward application of statistical pattern classification tools to the task. Possible improvements in class discrimination within this classification task are very likely to carry over to a retrieval task.

The term dictionary was constructed by discarding stop words, too frequent words, too infrequent words, and words shorter than three letters. Porter stemmer was used. This resulted in a term dictionary of 5718 terms. Normalized term histograms of this dimension were thus produced from the corpus.

As the classifier, we used a Support Vector Machine implementation called SVMTorch[3], which is well suited for large scale and/or sparse problems [5]. In all experiments we used a Gaussian kernel ($width = 10$), and $C = 100$ as the trade-off between training error and margin. These are the default values of SVMTorch. A binary classifier was trained for each of the 52 classes by using each class at a time as positive examples with the rest of the data as negative examples.

Each unknown vector from the test set is presented to all 52 SVM-classifiers. The output of an SVM is positive when it decides that an unknown test vector belongs to the class it was trained for. Since there may be several simultaneous such claims, and since the setting is exclusive classification, a simple maximum selector is applied to the SVMs to make the final decision for an exclusive class.

The LSI analysis was done using SVDPAKC/las2[4] [1].

[3] http://www.idiap.ch/learning/SVMTorch.html

[4] http://www.netlib.org/svdpack/

Results are depicted in a single chart in Fig. 2. The horizontal axis denotes the dimension of the document representation. Instead of precision/recall we are reporting just a single number, the classification accuracy, which is common in pattern recognition literature. This is defined as the number of correctly classified documents divided by the total number of documents. The classification accuracy is represented by the vertical axis of the chart. Naturally, the classifier is trained using only the training set, and the accuracy is evaluated using only the testing set.

The single diamond represents the accuracy using the original 5718-dimensional normalized term histograms as the document representation (88.8%). Results on features generated using completely random transform matrices to various output dimensions are plotted as black triangles. Each point is an average of results from five random trials (standard deviation plotted). Variance is high at low dimensions depending whether and how pertinent information happened to be included in the features. Results with LSI are plotted as squares ranging from an order of magnitude dimension reduction (513) down to one. LSI with dimension 513 is very close to the original accuracy (88.6%), but deteriorates approximately logarithmically. Halving the dimension decreases the accuracy by 4-5 percentage points, which resembles the behavior of a random projection [3], and shows that LSI does not really provide good features for discrimination.

In contrast, LDA exhibits higher accuracies (open triangles). The starting point of LDA was the 513-dimensional representation produced by LSI, which was transformed down to 1-64 dimensions. It is remarkable that a document representation of only twelve features achieves an accuracy as high as 91.1%. The optimal dimension appears to be somewhat higher. With dimension 64 we achieved an accuracy of 92.2%. LDA computed from a 513-dimensional random projection behaved very similarly (diamonds). These figures are again averages of five random trials. Standard deviations are also plotted but they are so tight to be almost invisible. LSI appears to be able to provide somewhat more pertinent information to the 513-dimensional data for classification purposes,

than a mere random transformation matrix, which LDA takes advantage of.

We also tested the signifigance of the difference between results of full dimensional data, and 12-dimensional LDA-from-LSI data. The χ^2 statistic of McNemar's test is 18.9, which indicates a significant difference beyond all confidence limits.

A transform to a lower dimension can only retain or reduce relevant information of the original data. If an SVM was completely immune to irrelevant information, the accuracy should only decrease as the dimension is reduced. Since higher accuracies were achieved, this is not obviously the case. Furthermore, this difference should be much more pronounced using classifiers, such as neural networks or decision trees that are more susceptible to overfitting than SVMs.

As opposed to feature transform methods that produce new features, each being a function of a large number of original features, we made a comparison to a scheme that selects a number of original features by evaluating each of them individually. We implemented average mutual information (also called as information gain) between a feature and the labels as the feature selection criterion [33]. Results are plotted as circles in Fig. 2. Selection fails almost completely at very low dimensions, and appears to behave slightly better than random projections, and slightly worse than LSI at medium dimensions. This is in line with experiments reported in [33], where selection of as many 1000 features was necessary to produce good results using Reuters database.

The top part of Figure 3 depicts an example projection of the training data onto a discriminative subspace of dimension two (which resulted in an accuracy of 65.1%). This projection uses the two best basis vectors as determined by LDA-criterion. In the bottom part of the figure the data is projected onto basis vectors ranked first and fourth by the criterion. Figure also lists all the categories present in this experiment. Comparing the two projections highlights an issue with the LDA-criterion. Since it attempts to combine both average class compactness and average class separation, either one could be increased to improve the criterion. When projected onto the first

and second best ranked basis vectors, the data appears to have more compact clusters, but with the expense of class separation. For example, the two largest categories, *acq* and *earn* both have quite compact clusters but unfortunately the clusters are partly on top of another. The projection at the lower half of the figure produces class clusters not quite as compact, but separation actually appears to be better.

Figure 4 depicts the same exact projections but displaying only seven of the classes in order not to clutter the picture.

In these projections, class distributions appear to be Gaussian-like, with unequal covariances, though. Thus there might be hope that projections based on HDA [23], or on joint maximum mutual information [31], might provide even better results in discrimination. This is also suggested by the compromise in compactness and separation of LDA illustrated in Figure 3.

8 Discussion

8.1 Computational Complexity.

The classification process with an SVM consists of evaluating the kernel between an unknown vector and all the support vectors. Applying a kernel that makes use of an inner product, the bulk of computations consists of evaluating those inner products. The document-term data is sparse by nature, and since support vectors are actual samples of the data, the inner products are computed between two sparse vectors. In the Reuters-21578 experiments in this paper the dimension of the sparse vectors is 5718, but on the average, each vector has only 36 non-zero components. The inner product between two such vectors requires on the average 72 indexing operations of which some may lead to actual multiply-accumulate operations. We count this as 72 operations/vector. Direct

application of the SVM to the document-term data and to the 52 binary classification problems produced 10496 support vectors altogether in the 52 classifiers. Roughly, the total number of operations in classifying one unknown vector is thus $10496 \times 72 = 755712$ operations. We compare now this to LDA.

Assuming an LDA transform to dimension 12, the transform matrix size is 12×5718. Computing the transform consists of evaluating the 12 inner products between the sparse input vector and each row of the dense transform matrix. This takes on the average 36 indexing operations and as many multiply/accumulates. We count this as 72 operations/vector, too, which results in only $12 \times 72 = 864$ operations. This amount is independent of whether LSI or a random projection is applied prior to LDA, because the final transform matrix is a product of the two matrices, and it is computed beforehand. Assuming the same number of support vectors, we have now $10496 \times 12 = 125952$ operations in the recognition phase of the SVM. Thus even with an efficient sparse implementation of the SVM, using LDA features cuts the computation down to one fifth of the original. Comparing to the direct application of an SVM to the data, the breakeven point appears to be at the LDA transform dimension of about twice the average number of nonzero entries in the original document vectors. The exact point depends, of course, on the implementation. For example, inner products with LDA vectors, as they are dense, can be executed using the vector instructions of the processor.

A Gaussian kernel is somewhat less favorable to the direct sparse SVM implementation without LDA, because computing the difference between two sparse vectors effectively doubles the number of nonzero elements in the result. The squared norm of the difference is then computed as an inner product with itself. This doubling of the size does not happen with dense and low-dimensional LDA vectors.

Computation in SVM training appears to retain the same proportions as the testing phase.

Computing the LSI and LDA are extra, but the LSI takes only a few minutes for the Reuters collection, and the LDA no more than 30 seconds (on a 700 MHz Pentium III), and these only need to be computed once. Of course, the training data needs to be transformed, but this is again an insignificant amount of computation and it only needs to be done once.

8.2 Application to Retrieval Tasks.

The largest difference between a classification and a retrieval task is the occurrence of multiple labels per document in retrieval tasks. There are now two options to incorporate them into the LDA. The first is what has been done in this paper: Assume that the documents possessing multiple labels do not carry information about class separation and ignore them in the computation of the LDA basis vectors.

This is not entirely true, of course, since a multi-labeled document carries information about the separation of the labeled classes from all other classes. The second option is thus to account for multiply labeled documents multiple times, once for each label, weighted by the reciprocal of the number of labels. For example, when computing the class mean vectors and covariance matrices required for $\mathbf{S}_w$ and $\mathbf{S}_b$, a document carrying three labels is added to the mean vectors of all those three classes, but only by a weight of one third each. This may have an effect of smearing class distinctions, and needs to be experimentally evaluated.

The actual SVM approach is straightforward to modify to make use of training data with multiple labels. Since a multi-class SVM can consist of a number of binary classifiers where one class is set against all others, all documents carrying a particular label, independent of whether they also have other labels or not, are counted as positive examples of that class, and the rest as negative. Setting the decision thresholds for desired precision/recall is then another matter that requires a

validation data set.

8.3 Interpretation of the LDA Basis Vectors.

It would be interesting to study to what original terms these discriminating features correspond. However, these 12 basis vectors are dense and have both positive and negative entries, which makes the interpretation hard. Now, as the transform, we can have any 12 vectors that span the same subspace as the original basis vectors. It is possible to find a rotation of this 12-dimensional subspace such that the basis vectors become positive and sparse. This has been suggested by Kabán and Girolami as a means to make LSI more amenable to interpretation [18]. They propose projection pursuit with a *skewness* projection index. Same method could now be applied after LDA, if interpretability by means of the original terms is desired.

8.4 Other Work on Feature Selection and Feature Transforms.

Comparing this work to others on both feature selection and feature transforms using the same database, Joachims reports SVM experiments with 86.4% precision using 9962 features [17]. Yang and Pedersen report an average precision of about 90% with 2000 selected features on a same task [33]. They used a k-NN classifier, but kept multiply labeled documents in the train and test sets. A few papers report classification error rate (or accuracy which is one minus error rate) on a uni-labeled subset of Reuters-21578. Han et al report a classification accuracy of 90% also by using 2000 selected features, and a weighted k-NN approach [14]. Karypis et al describe a method that determines the columns of the projection matrix as the differences of the means between clusters or classes in the data [19]. This is a similar but slightly more heuristic criterion than using only the between-class covariance matrix in LDA and paying no attention to class compactness. This

criterion has been used earlier in data visualization [8]. Their results are 82-85% classification accuracy on small subsets of Reuters using 50 transformed features.

9 Conclusion

This paper shows how Linear Discriminant Analysis can be used to reduce drastically the dimension of document representation in classification tasks without sacrificing the accuracy. In fact, the classification error rate decreased from 11.2% to 8.9% when reducing the original 5718 dimensional document representation into mere 12 features.

Although these results can not be directly compared to previous work on the same database due to different methods of selecting the train/test data and different scoring methods, the trend is visible: Previous work shows that a high accuracy is possible with a large subset of features. This work points out that a high accuracy in document classification is possible with a small number of discriminative features. This offers some computational advantages even with Support Vector Machines that take advantage of the sparse nature of the data.

References

[1] Michael W. Berry. Large scale singular value computations. *International Journal of Supercomputer Applications*, 6(1), 1992.

[2] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, New York, 1995.

[3] William Campbell, Kari Torkkola, and Sree Balakrishnan. Dimension reduction techniques for training polynomial networks. In *Proceedings of the 17th International Conference on Machine Learning*, pages 119–126, Stanford, CA, USA, June 29 - July 2 2000.

[4] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *VLDB Journal: Very Large Data Bases*, 7(3):163–178, 1998.

[5] Ronan Collobert and Samy Bengio. SVMTorch: Support Vector Machines for Large-Scale Regression Problems. *Journal of Machine Learning Research*, 1:143–160, 2001.

[6] Sanjoy Dasgupta. Experiments with random projection. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 143–151, Stanford, CA, June30 - July 3 2000.

[7] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407, 1990.

[8] I.S. Dhillon, D.S. Modha, and W.S. Spangler. Visualizing class structure of multidimensional data. In *Proceedings of the 30th Symposium of on the Interface, Computing Science, And Statistics*, pages 488–493, Minneapolis, MN, USA, May 1998. Interface Foundation of North America.

[9] C. Faloutsos and K.-I. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 163–174, San Jose, CA, 1995.

[10] J.H. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. on Computers*, C-23:881, 1974.

[11] K. Fukunaga. *Introduction to statistical pattern recognition (2nd edition)*. Academic Press, New York, 1990.

[12] Mark Girolami, Andrzej Cichocki, and Shun-Ichi Amari. A common neural network model for unsupervised exploratory data analysis and independent component analysis. *IEEE Transactions on Neural Networks*, 9(6):1495 – 1501, November 1998.

[13] Xuan Guorong, Chai Peiqi, and Wu Minhui. Bhattacharyya distance feature selection. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 2, pages 195 – 199. IEEE, 25-29 Aug. 1996.

[14] Eui-Hong (Sam) Han, George Karypis, and Vipin Kumar. Text categorization using weight adjusted k-nearest neighbor classification. In *Proc. PAKDD*, 2001.

[15] David Hull. Using statistical testing in the evaluation of retrieval performance. In *Proc. of the 16th ACM/SIGIR Conference*, pages 329–338, 1993.

[16] David Hull. Improving text retrieval for the routing problem using latent semantic indexing. In *Proc. SIGIR'94*, pages 282–291, Dublin, Ireland, July 3-6 1994.

[17] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[18] Ata Kabán and Mark Girolami. Fast extraction of semantic features from a latent semantic indexed corpus. *Neural Processing Letters*, 15(1), 2002.

[19] G. Karypis and E. Sam. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. Technical Report TR-00-0016, University of Minnesota, Department of Computer Science and Engineering, 2000.

[20] Thomas Kolenda, Lars Kai Hansen, and Sigurdur Sigurdsson. Indepedent components in text. In M. Girolami, editor, *Advances in Independent Component Analysis*. Springer-Verlag, 2000.

[21] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *Proceedings of ICML-96, 13th International Conference on Machine Learning*, pages 284–292, Bari, Italy, 1996.

[22] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[23] N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communcation*, 26:283–297, 1998.

[24] Huan Liu and Hiroshi Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, 1998.

[25] Dunja Mladenic. Feature subset selection in text-learning. In *European Conference on Machine Learning*, pages 95–100, 1998.

[26] T. Okada and S. Tomita. An optimal orthonormal system for discriminant analysis. *Pattern Recognition*, 18(2):139–144, 1985.

[27] Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.

[28] John W. Sammon Jr. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409, May 1969.

[29] George Saon and Mukund Padmanabhan. Minimum bayes error feature selection for continuous speech recognition. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 800–806. MIT Press, 2001.

[30] Hinrich Schütze, David Hull, and Jan O. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Proc. SIGIR'95*, 1995.

[31] Kari Torkkola and William Campbell. Mutual information in learning feature transformations. In *Proceedings of the 17th International Conference on Machine Learning*, pages 1015–1022, Stanford, CA, USA, June 29 - July 2 2000.

[32] H. Yang and J. Moody. Data visualization and feature selection: New algorithms for non-gaussian data. In *Proceedings NIPS'99*, Denver, CO, USA, November 29 - December 2 1999.

[33] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. 14th International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann, 1997.

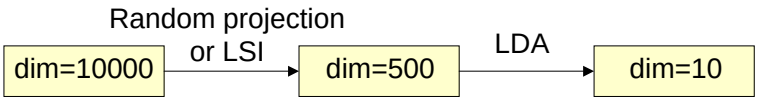

Figure 1: Dimension reduction in two steps: The first step is performed either using a random projection or by LSI, which is efficient on sparse matrices. This allows the second dimension reduction step, LDA, to operate on dense matrices of manageable size.

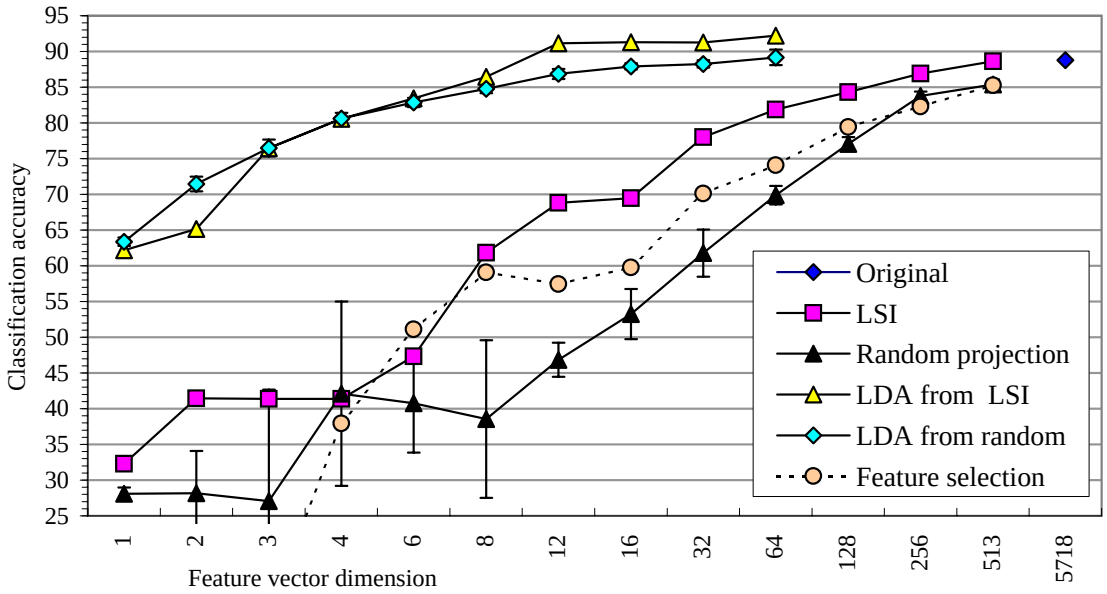

Figure 2: Classification accuracies on Reuters-21578 test data using feature transforms (LDA, LSI, random projections, and combinations of LDA with LSI and random projections), and feature selection.

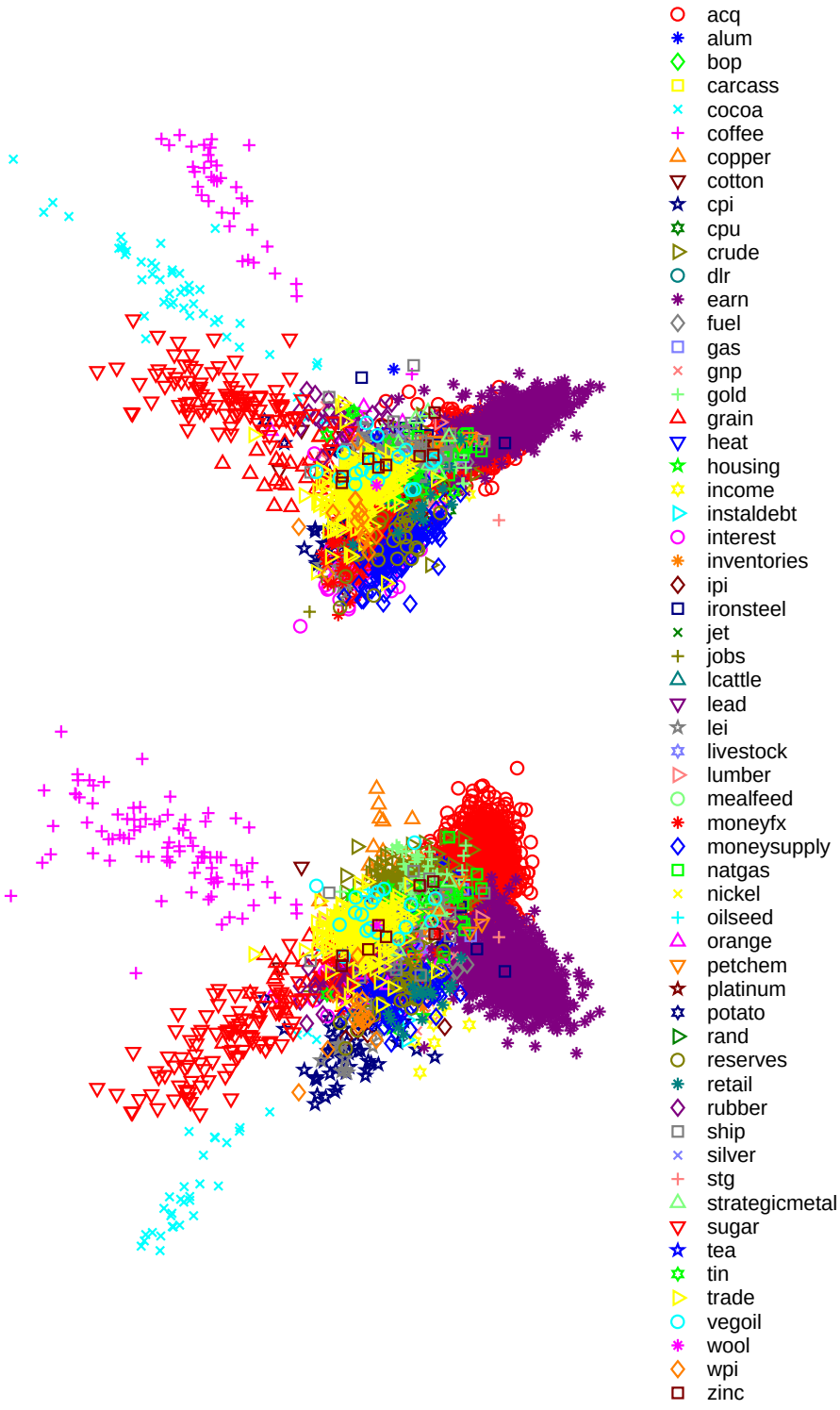

Figure 3: LDA projection of the Reuters-21578 training set onto a two-dimensional discriminative subspace. The top image depicts projection onto the two best basis vectors as determined by the LDA-criterion. The bottom part uses basis vectors ranked first and fourth by the criterion.

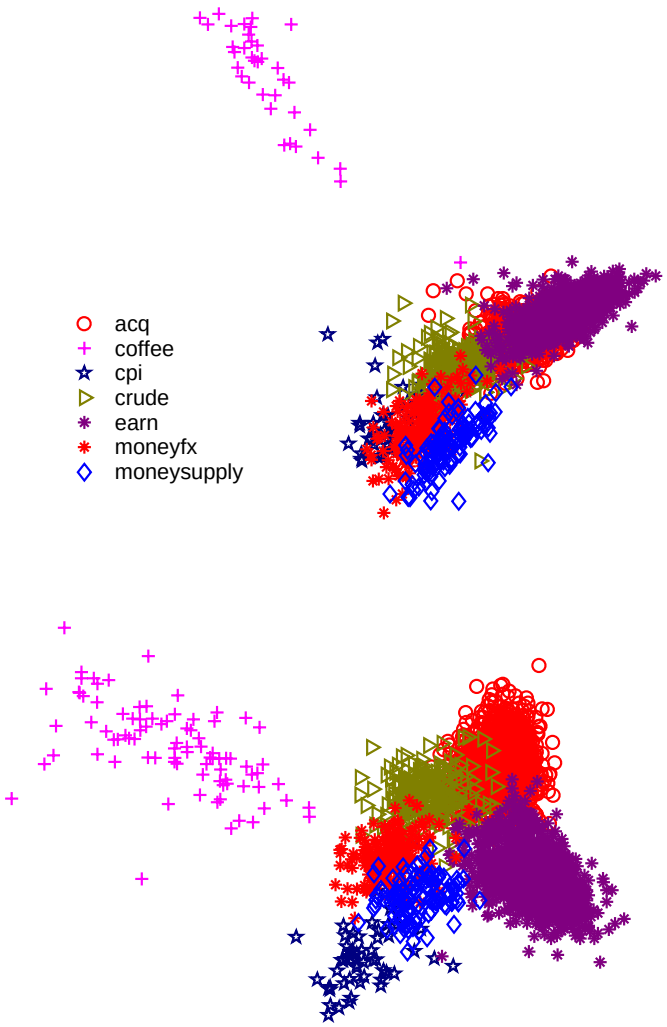

Figure 4: LDA projection of a few most numerous classes of the Reuters-21578 training set onto a two-dimensional discriminative subspace. The top image depicts projection onto the two best basis vectors as determined by the LDA-criterion. The bottom part uses basis vectors ranked first and fourth by the criterion. This demonstrates the compromise that LDA-criterion makes: LDA ranks higher basis vectors that produce more compact clusters with the expense of class separation.