Enhanced Context Recognition by Sensitivity Pruned Vocabularies

Rasmus Elsborg Madsen, Sigurdur Sigurdsson and Lars Kai Hansen
Informatics and Mathematical Modelling
Technical University of Denmark
Building 321, DK-2800 Kgs. Lyngby, Denmark
rem,siggi,lkh@imm.dtu.dk

Abstract

Document categorization tasks using the "bag-of-words" representation have been successful in instances [11]. The relatively low dimensional bag-of-words form, is well suited for machine learning methods. The pattern recognition methods suffers though, from the well-known curse of dimensionality, since the number of input dimensions (words) usually supersedes the number of examples (documents). This high dimensional representation is also containing many inconsistent words, possessing little or no generalizable discriminative power, and should therefore be regarded as noise. Using all the words in the vocabulary is therefore resulting in reduced generalization performance of classifiers. We here study the effect of sensitivity based pruning of the bag-of-words representation. We consider neural network based sensitivity maps for determination of term relevancy, when pruning the vocabularies. With reduced vocabularies documents are classified using a latent semantic indexing representation and a probabilistic neural network classifier. Pruning the vocabularies to approximately 3% of the original size, we find consistent context recognition enhancement for two mid size data-sets for a range of training set sizes. We also study the applicability of the sensitivity measure for automated keyword generation.

1 Introduction

The world wide web is an unstructured and fast growing database. Today's search tools often leave web users in frustration by the low precision and recall[1]. It is widely believed that machine learning techniques can come to play an important role in web search. Here we consider web content mining in the form of document classification - an information retrieval aspect of web-mining [8]. Our aim is to improve generalizability of supervised document classification by vocabulary pruning.

In the bag-of-words representation we summarize documents by their term histograms. The main motivation for this reduction is that it is easily automated and needs minimal user intervention beyond filtering of the term list. The term list typically contains in the range of $10^3 - 10^5$ terms, hence further reduction is necessary for most pattern recognition devices. Latent semantic indexing (LSI) [4, 3] aka principal component analysis is often used to construct low dimensional representations. LSI is furthermore believed to reduce synonymy and polysemy problems [3, 9]. Although LSI and other more elaborate vector space models have been successful in text classification in small and medium size databases, see e.g., [7, 5], we are still not at human level text classification performance. When training classifiers on relatively small databases generalizability is a key issue. How well does a model adapted on one set of data predict the labels of another test data set? Generalizability is in general a function of the number of training cases and of the effective model dimension. We are interested in investigating whether automated vocabulary reduction methods can contribute to classification accuracy by reducing the model complexity. To estimate term relevance we will use the notion of the scaled sensitivity, which is computed using a the so-called NPAIRS split-half re-sampling procedure [15]. Our hypothesis is that 'sensitivity maps' can determine which terms are consistently important. That is terms which are likely to be of general use for classification, relative to terms that are of low or highly variable sensitivity. As a side we also illustrate the feasibility of using the sensitivity to generate class specific keywords.

2 Methods

Documents are arranged in a term-document matrix $\mathbf{X}$, where X_{td} is the number of times term t occur in document d. The dimensionality of $\mathbf{X}$ is reduced by filtering and stemming. Stemming refers to a process in which words with different endings are merged, e.g., 'trained' and 'training' are merged into the common stem 'train'. This example also

indicates the main problem with stemming, namely that introducing an artificial increased polysemy. We have decided to 'live with this problem' since without stemming vocabularies would grow prohibitively large. About 500 common non-discriminative stop-words, i.e. (a, i, and, an, as, at) are removed from the term list. In addition high and low frequency words are also removed from the term list. The resulting elements in $\mathbf{X}$ are term frequency/inverse document frequency normalized,

$$X_{t,d}^{\text{tfidf}} = \frac{X_{t,d}}{\sqrt{\sum_{t'=1}^{T} X_{t',d}^2}} \log \frac{D}{F_t}. \quad (1)$$

Where $\mathbf{X}^{\text{tfidf}}$ is the normalized term document matrix, T is the number of terms, D is the number of documents and F_t is the document frequency of term t. The normalized term document matrix is reduced to a feature-document matrix using PCA, carried out by an 'economy size' singular value decomposition,

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T. \quad (2)$$

Where the orthonormal $T \times D$ matrix $\mathbf{U}$ contains the eigenvectors corresponding to the non-zero eigenvalues of the symmetric matrix $\mathbf{X}\mathbf{X}^T$. $\mathbf{\Lambda}$ is a $D \times D$ diagonal matrix of singular values ranked in decreasing order and the $D \times D$ matrix $\mathbf{V}^T$ contains eigenvectors of the symmetric matrix $\mathbf{X}^T\mathbf{X}$. The LSI representation is obtained by projecting document histograms on the basis vectors in $\mathbf{U}$,

$$\mathbf{F} = \mathbf{U}^T\mathbf{X} = \mathbf{\Lambda}\mathbf{V}^T. \quad (3)$$

Typically, the majority of the singular values are small and can regarded as noise. Consequently, only a subset of K ($K << T$) features are retained as input to the classifier algorithm. The representational potential of these LSI features is illustrated in figure 1. A wide variety of classification algorithms have been applied to the text categorization problem, see e.g., [8]. We have extensive experience with probabilistic neural network classifiers and a well tested ANN toolbox is available [12]. The toolbox adapts the network weights and tunes complexity by adaptive regularization and outlier detection using the Bayesian ML-II framework, hence, requires minimal user intervention [12, 13].

We use the definition of class specific sensitivity proposed in [16, 14] for a set of N samples,

$$\mathbf{s}_k = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{\partial P(c_k|\mathbf{f}_n)}{\partial \mathbf{x}} \right| \quad (4)$$

and where $P(c_k|\mathbf{f}_n)$ is the posterior probability of class k given the feature vector $\mathbf{f}_n$. $\mathbf{s}_k$ is the T-dimensional sensitivity vector for class k. The T-dimensional derivative is obtained using the projection (3) [14]. A split-half resampling procedure is invoked to determine the statistical significance of the sensitivity [15]. Multiple splits are generated of the original training set and classifiers trained on each of the splits. For each classifier a sensitivity map is computed. Since the two maps obtained from a given split are exchangeable the mean map is an unbiased estimate of the 'true' sensitivity map, while the squared difference is a noisy, but unbiased estimate of the variance of the sensitivity map. By repeated re-sampling and averaging the sensitivity map and its variance is estimated. We finally obtain a scaled sensitivity map by normalization trough the standard deviation.

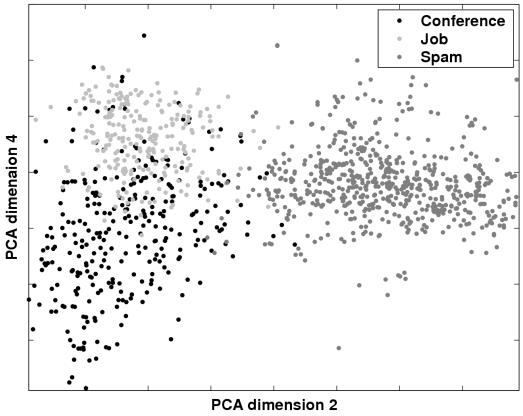

Figure 1. Illustration of the document distribution in feature space. Here we show the Email corpus projected onto the 2nd and 4th principal directions. In this projection the 'spam' class is well separated while the two other classes in the set ('conferences' and 'jobs') show some overlap.

3 Data

Three data-sets, 'Email' [10], 'Multimedia' [6] [7] and 'WebKB' [2] are used to illustrate and test the hypothesis. No less than ten split-half re-samples are used in all experiments. The Email data-set consists of texts from 1431 emails in three categories: conference (370), job (272) and spam (789). The multimedia data set consists of texts and images from 1200 web pages. Only the text part is considered here. The categories are: sports (400), aviation (400) and paintball (400). The WebKB set contains 8282 webpages from US university computer science departments. Here we have used a subset of 2240 pages from the WebKB earlier used in [5] and [9]. The WebKB categories are: project (353), faculty (483), course (553) and student (851). All html tags were removed from the data-set.

4 Results

Preliminary experiments indicated that a reduced feature space of $K = 48$ projections and a neural network classifier with five hidden units were sufficient for the task (data not shown). All results have been validated using 10 fold split half re-sampling cross validation. The neural network based term sensitivity is a function of the given training set. Terms for which the sensitivity is high but also highly variable are less likely to support generalizability compared to terms that have a consistent high or medium sensitivity. The empirical distribution of mean and standard deviations the terms in the Email set are shown in figure 2. The empirical

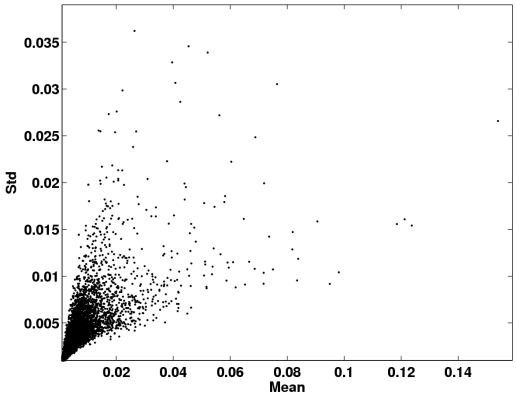

Figure 2. Mean and standard deviation of the term sensitivity. The most relevant terms have consistent high sensitivity in each re-sampling split, i.e., a high mean and relatively low standard deviation. These terms occupy the lower right part of the plot.

scaled sensitivities $Z_t = \mu_t/\sigma_t$ of the terms in the Email data, are shown in figure 3. The scaled sensitivities can be used also to select relevant keywords for the text categories. For the Email data the five highest scores for the Conference category are (*Paper, Conference, Deadline, Neural, Topic*) and for the Job category (*Research, Position, Candidate, University, Edt*) and for the Spam category (*Money, Remove, Free, Thousand, Simply*). We then remove increasing fractions of the terms using the scaled sensitivity ranking. Using all the terms, the generalization classification error rate is 16.9% in the WebKB and 2.7% in Email data. Removing 97% of the terms with the lowest scaled sensitivity, the generalization error for the WebKB is reduced to 14.3% and to 2.3% for the Email data. This is a reduction of about 15% for both data-sets. For the Multimedia data the pruning does not lower the generalization error, but about 80% of the multimedia vocabulary can be removed without loss of generalizability. The generalizability as function of pruning fraction is presented in figure 4. The multimedia

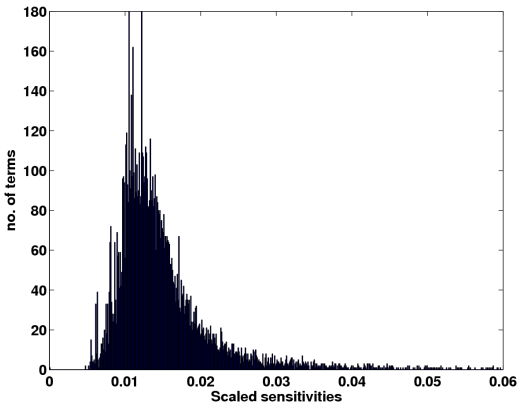

Figure 3. Scaled sensitivities Email data set terms. Many of the terms have a standard deviation on the order of the mean value indicating that they are not consistently sensitive. 3% of the terms has a scaled sensitivity higher than 3.75.

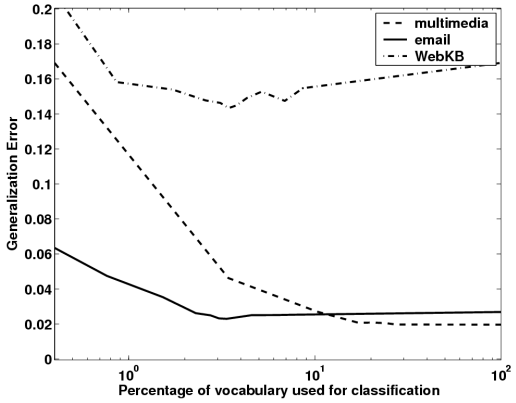

Figure 4. Generalization error using subsets of the vocabulary for the WebKB, Email and Multimedia data. The terms retained are those with the highest scaled sensitivity. For the WebKB and Email data-sets pruning to 3% of the vocabulary gives the lowest generalization error with error rates reduced about 15%. The relatively limited vocabulary of the Multimedia data set can be reduced to about 50% without decreasing performance but does not improve performance.

data has the sparsest vocabulary with only 3500 terms after filtering while the Email data set has 9500 terms, and the WebKB data has 12950 terms after filtering. In figure 5 we show that learning curves are consistently improved for a range of training sets for the WebKB and the Email data

based on a fixed reduction to 3% of their original vocabulary. Pruning the vocabulary to 3% of the original size, results in better generalization in the whole range of training-set sizes and has the largest effect for small training sets.

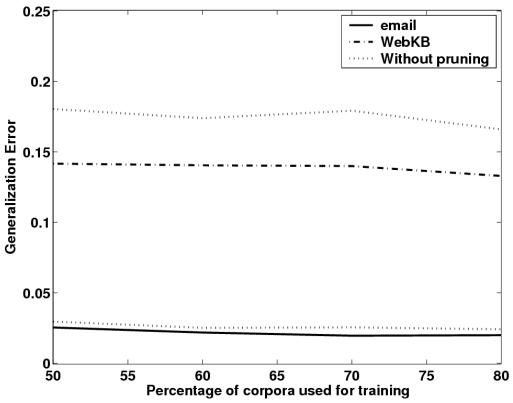

Figure 5. Learning curves for full and reduced the vocabularies for the WebKB and Email data sets. Both vocabularies are reduced to 3% of their original sizes. While the effect of pruning is consistently positive for all training set sizes, the effect is most pronounced for small samples.

5 Conclusion

Neural network sensitivity maps has been applied to a latent semantic indexing context recognition framework. The use of scaled sensitivities for reducing vocabularies, has resulted in use of only 3% of the original vocabulary. The vocabulary reduction has lead to a simpler modeling while also significantly improving classification performance for two large vocabulary data sets. This indicates that these context classification problems are generalization limited by model complexity for the sample sizes studied. Classification of a third data-set with a more limited vocabulary was not enhanced by vocabulary pruning. However, the vocabulary could be pruned to 80% of the initial size without loss of performance. We thus recommend to monitor the test set classification performance as function of the pruning fraction. The scaled sensitivity has also been useful for identifying class specific keywords.

References

[1] S. Chakrabarti. Data mining for hypertext: a tutorial survey. *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM*, 1:1–11, 2000.

[2] CMU-WebKB. The 4 universities data set. http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/, 1997.

[3] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41:391–407, 1990.

[4] G. Furnas, S. Deerwester, S. Dumais, T. Landauer, R. Harshman, L. Streeter, and K. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *The 11th International Conference on Research and Development in Information Retrieval*, pages 465–480, Grenoble, France, 1988. ACM Press.

[5] L. Hansen, S. Sigurdsson, T. Kolenda, F. Nielsen, U. Kjems, and J. Larsen. Modeling text with generalizable gaussian mixtures. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3494–3497. IEEE, 2000.

[6] T. Kolenda. Multimedia dataset. http://mole.imm.dtu.dk/faq/MMdata/, 2002.

[7] T. Kolenda, L. Hansen, J. Larsen, and O. Winther. Independent component analysis for understanding multimedia content. In S. B. J. L. H. Bourlard, T. Adali and S. Douglas, editors, *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*, pages 757–766, Piscataway, New Jersey, 2002. IEEE Press.

[8] R. Kosala and H. Blockeel. Web mining research: A survey. In *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM*, pages 1–15. ACM Press, 2000.

[9] J. Larsen, L. Hansen, A. Have, T. Christiansen, and T. Kolenda. Webmining: learning from the world wide web. *Computational Statistics and Data Analysis*, 38:517–532, 2002.

[10] F. Nielsen. Email data-set. http://www.imm.dtu.dk/∼rem/, 2001.

[11] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.

[12] S. Sigurdsson. The dtu: Artificial neural network toolbox. http://mole.imm.dtu.dk/toolbox/ann/, 2002.

[13] S. Sigurdsson, J. Larsen, L. Hansen, P. A. Philipsen, and H. C. Wulf. Outlier estimation and detection: Application to skin lesion classification. In *International conference on acoustics, speech and signal processing*, pages 1049–1052, 2002.

[14] S. Sigurdsson, P. Philipsen, L. Hansen, J. Larsen, M. Gniadecka, and H. Wulf. Detection of skin cancer by classification of Raman spectra. *Accepted for IEEE Transactions on Biomedical Engineering*, 2004.

[15] S. Strother, J. Anderson, L. Hansen, U. Kjems, R. Kustra, J. Siditis, S. Frutiger, S. Muley, S. LaConte, and D. Rottenberg. The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *Neuroimage*, 15:747–771, 2002.

[16] J. Zurada, A. Malinowski, and C. I. Sensitivity analysis for minimization of input data dimension for feedforward neural network. In *Proceedings of the IEEE Symposium on Circuits and Systems*, pages 447–450, 1994.