Subspace Mapping of Noisy Text Documents

Axel J. Soto[1], Marc Strickert[2],
Gustavo E. Vazquez[3], and Evangelos Milios[1]

[1] Faculty of Computer Science, Dalhousie University, Canada
soto@cs.dal.ca
[2] Institute for Vision and Graphics, Siegen University, Germany
[3] Dept. Computer Science, Univ. Nacional del Sur, Argentina

Abstract. Subspace mapping methods aim at projecting high-dimensional data into a subspace where a specific objective function is optimized. Such dimension reduction allows the removal of collinear and irrelevant variables for creating informative visualizations and task-related data spaces. These specific and generally de-noised subspaces spaces enable machine learning methods to work more efficiently. We present a new and general subspace mapping method, Correlative Matrix Mapping (CMM), and evaluate its abilities for category-driven text organization by assessing neighborhood preservation, class coherence, and classification. This approach is evaluated for the challenging task of processing short and noisy documents.

Keywords: Subspace Mapping, Compressed Document Representation.

1 Introduction

Many data-oriented areas of science drive the need for faithfully representing data containing thousands of variables. Therefore, methods for considerably reducing the number of variables are desired focusing on subsets being minimally redundant and maximally task-relevant. Different approaches for subspace mapping, manifold learning, and dimensionality reduction (DR) were proposed earlier [1,2]. A current challenge in information representation is the huge amounts of text documents being produced at increasing rates. Using the well-known vector space representation, or "bag of words" model, a corpus of documents is described by the set of words that each document contains. This approach yields a document-term matrix containing thousands of unique terms, and thus is very likely to be sparse.

The text mining communities have developed methods for automatic clustering and classification of document topics using specific metrics and kernels. Yet fully developed human-in-the-loop approaches are rare to enable the user to perform visual data exploration and visual data mining. While the automatic learning of data is crucial, visualization is another key aspect for providing an intuitive interface to contained information and for interactive tuning of the

C. Butz and P. Lingras (Eds.): Canadian AI 2011, LNAI 6657, pp. 377–383, 2011.

data/text mining algorithms. This makes DR methods indispensable for interactive text corpus exploration.

In this paper, we present an application of a recent DR method, *Correlative Matrix Mapping* (CMM), which has been successfully applied in other domains [3][1] in the context of regression problems. This method is based on an adaptive matrix metric aiming at a maximum correlation of all pairwise distances in the generated subspace and the associated target distances. Preliminary work [4][1] showed some capabilities of this approach for the expert-guided visualization of labeled text corpora by integrating user feedback on the base of the interpretable low-dimensional mapped document space. Here, we provide a comprehensive comparison of CMM and other competitive DR methods for creating representative low-dimensional subspaces. Since machine learning methods rely on distance calculations, we investigate how such projections with label-driven distance metrics can improve representations of short and noisy text documents. We refer to noisy documents as the ones that are not properly written in terms of spelling and grammatical structure. Such documents are quite common in business environments such as aircraft maintenance records, online help desk or customer survey applications, and their analysis is thus highly relevant. Still, much work in the information extraction literature is focused on well-formed text documents.

2 Correlative Matrix Mapping (CMM)

Given n m-dimensional data vectors $\mathbf{x}^j \in \mathbf{X} \subset \mathbb{R}^m, 1 \leq j \leq n$, such that each $\mathbf{x}^j$ is associated to a q-dimensional vector $\boldsymbol{l}^j \in \mathbf{L} \subset \mathbb{R}^q$. For text corpora, n is the number of labeled documents in the corpus, m is the number of terms in the corpus and $\boldsymbol{l}^j$ is the vector representation of the label of the document $\mathbf{x}^j$. CMM aims at finding a subspace of $\mathbf{X}$ where the pairwise distances $\mathbf{D}_\mathbf{X}^{\boldsymbol{\lambda}}$ are in maximum correlation with those on the label space ($\mathbf{D}_\mathbf{L}$). Thus, pairwise distances in the document-term space are sought to be in maximum correlation with those distances on the label space. Here, $\mathbf{D}_\mathbf{L}$ is used as the Euclidean distance on the label space and the $\boldsymbol{\lambda}$ superscript in $\mathbf{D}_\mathbf{X}^{\boldsymbol{\lambda}}$ indicates parameters of the adaptive distance $(\mathbf{D}_\mathbf{X}^{\boldsymbol{\lambda}})_{i,j} = ((\boldsymbol{x}^i - \boldsymbol{x}^j)^\mathsf{T} \cdot \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^\mathsf{T} \cdot (\boldsymbol{x}^i - \boldsymbol{x}^j))^{1/2}$, where $\boldsymbol{\lambda}$ is an $m \times u$ matrix, and u is specified by the user. This distance matrix metric resembles a Mahalanobis distance, where $\Lambda = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^\mathsf{T}$ has a rank of u. We obtain the parameter matrix as

$$\boldsymbol{\lambda} = \arg\max_{\boldsymbol{\lambda}^*} \mathsf{r}(\mathbf{D}_\mathbf{L}, \mathbf{D}_\mathbf{X}^{\boldsymbol{\lambda}}) \tag{1}$$

where r is the Pearson correlation. Locally optimal solutions for (1) can be obtained by gradient methods using its derivative with respect to $\boldsymbol{\lambda}$ [3]. It is worth noting that while the number of rows of the $\boldsymbol{\lambda}$ matrix is constrained by the number of terms in $\mathbf{X}$, i.e. the document vector dimensionality, the number of columns u, i.e. the dimensionality of the subspace, is defined by the user. Note

[1] CMM was called differently in previous works but created naming conflicts therein.

that $\boldsymbol{\lambda}^{\top} \cdot \mathbf{X}$ defines a u-dimensional subspace that is an informative representation of the input space focused on its label association. If visualization is the ultimate goal, a choice of $u \leq 3$ is recommended. New documents with unknown labels can be also projected to the new space by using the optimized $\boldsymbol{\lambda}$ matrix. An open source package with the implementation of CMM is available at [5].

3 Experiments

We selected four alternative DR methods that make use of label information and allow exact out-of-sample extension, and we used them to compare to CMM. **Linear Discriminant Analysis (LDA)** aims at finding optimal discriminant directions by maximizing the ratio of the between-class variance to the within-class variance [7]. Since its solutions require inverses of covariance matrices, it usually has ill-conditioning problems for high-dimensional data. Therefore, we also calculate a simplification of this approach based on the diagonal matrices of the covariance matrices, referred to as LDA_d. **Canonical Correlation Analysis (CCA)** is a well-known technique for finding correlations between two sets of multidimensional variables by projecting them onto two lower-dimensional spaces in which they are maximally correlated [8]. Although this method is strongly related to CMM in the sense that both look for optimal correlations, CMM does not adapt data and labels space, but adapts distances in the data space. **Neighborhood Component Analysis (NCA)** aims at learning a linear transformation of the input space such that the k-Nearest Neighbors method performs well in the transformed space [9]. The method uses a probability function to estimate the probability $p_{i,j}$ that a data point i selects a data point j as its neighbor after data mapping. The method maximizes the expected number of data points correctly classified under the current transformation matrix. **Maximally Collapsing Metric Learning (MCML)** aims at learning a linear mapping where all points in the same class are mapped into a single location, while all points in other classes are mapped to other locations, i.e. as far as possible among data points of different classes [10]. This algorithm uses a probabilistic selection rule as in NCA. However, unlike NCA the optimization problem is convex, and thus MCML transformation can be completely specified from the objective function. The Matlab Toolbox for Dimensionality Reduction [2] was used for all methods except for CCA taken from the Statistics Toolbox [6].

3.1 Data

We used the publicly available Aviation Security Reporting System (ASRS) Database Report Set [11] and extracted the narrative fields of documents belonging to 4 out of 24 topics: *Bird or animal strike records*, *Emergency medical service incidents*, *Fuel management issues* and *Inflight weather encounters*. Each topic has 50 documents, thus providing a total of 200 documents. 6048 rare terms were discarded yielding 1829 unique terms. Two major challenges are faced. First, the

average length of each document is only a few sentences, which makes it difficult to extract statistically significant terms. Second, texts are riddled with acronyms, *ad hoc* abbreviations and misspellings.

Binary representations are used for the document-term matrix, i.e. the component for to the k^{th} term of the j^{th} document $\mathbf{x}^j$ is 0 if the term is not present and 1 otherwise. This binary weighting approach is appropriate given the short length of the documents for which the frequency of a term might inflate its importance. In the case of CCA and CMM, the four label vectors (0,0,0,1), (0,0,1,0), (0,1,0,0), and (1,0,0,0) are used for class representation, thus, inducing equidistant classes. In LDA, NCA and MCML integer values are used for class assignment, because they do not quantify label dissimilarities.

For each experiment, 80% of the corpus was used for training, while the remaining documents were held-out for testing. This process was restarted 10 times, so that a new testing set was obtained in each iteration for implementing a repeated random sub-sampling validation scheme. All the applied DR algorithms showed convergence during the optimization phase, with the exception of MCML which, despite of its convex cost function, required a time-limiting stopping criterion, because of its excessive run time. Since NCA and CMM use iterative methods for optimization, different early stopping criteria were sought using a portion of the training set. Otherwise, these methods are likely to overfit training data. Since meaningful visualization is desirable for many tasks, all our experiments were deliberately constrained to 2- and 3-dimensional subspaces.

3.2 Assessing Subspace Mapping Performance

We divide different assessments applied on the methods into three types. **The first** aims at evaluating the embedding without using label information. Two performance metrics are used: the area under the extrusion/intrusion tendency curve (B) and neighborhood ranking preservation (Q) [12]. B quantifies the tendency to commit systematic neighborhood rank order errors for data pairs in the projection space (B is not bounded; the closer to zero, the better), while Q measures k-ary neighborhood preservation (Q varies between 0 and 1; the closer to one, the better). **The second** quality class considers label information namely *cohesion*, which is the ratio of the pairwise Euclidean distances of documents belonging to a same class to the pairwise distances of documents of different classes. **The third** class of assessments also uses label information. It evaluates the potential of supervised learning methods to exploit the given low-dimensional space for classification. It may be argued that the better the classification accuracy is, the better the projection is. Classifying from a low-dimensional space may produce better results due to the removal of collinear or irrelevant variables. Thereby, we used k-nearest neighbors (kNN), Decision Trees (DT), Support Vector Machines (SVM) using a Radial Basis Function kernel (rbf) and using a multi-layer perceptron kernel (mlp).

4 Results

We will focus on the results obtained on the testing set, while the training set results are still available for the reader. Table 1 shows the average of the computed metrics of the different DR methods when they are projected into a 2D space. It can be observed that most methods have an intrusive embedding, i.e. a tendency to positive rank errors in the subspace. Not surprisingly, NCA has the highest average preservation of the k-ary neighborhoods, since this is what its mapping is trying to capture. Yet the difference is not statistically significant with CMM when a Dunnett test [13] is performed with a 1% familywise probability error.

Although CMM does not have the lowest *cohesion* value, due to the variance of this metric, no significant difference can be drawn here. Looking at the performance of the classification methods, CMM significantly outperforms all the other methods with the exception of NCA with the kNN method. Nevertheless, no significant difference was found between CMM and NCA with kNN.

Results for the 3D projection show a very similar behavior as the one showed for the projections into the 2D space (Table 2). We can see that LDA_d has a good classification accuracy when DT are used, although no significant differences with CMM and NCA were found. We can also see that CMM made an improvement on most of the metrics.

In summary, LDA and CCA had poor performances in most metrics. These methods compute their optimal value in closed-form (using eigenvectors or inverse of covariance matrices), and thus the computation might get corrupted due to the large number of variables and relatively small number of documents. LDA_d has better performance than LDA. However, most of the components of the parameter matrix in LDA_d are zero. This yields a cluttered projection of the data points to a few locations, which is not convenient on most cases.

The remarkably poor performance of MCML might be due to an underfitting situation. It is worth saying that MCML is the most compute-intensive method by far and its calculations last more than 50 times the amount of time spent on any other algorithm. Moreover, delaying its stopping criterion does not seem to dramatically improve its performance. Finally, it is important to note that

Table 1. Comparison of DR methods using 2D spaces: rank-based quality measures (Q/B), cohesion, and classification accuracies of four classifiers.

	LDA		LDA_d		NCA		MCML		CCA		CMM	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Q	0.534	0.563	0.515	0.503	0.599	**0.613**	0.533	0.548	0.521	0.539	0.544	0.596
B	0.015	0.049	0.007	**0.014**	0.039	0.046	-0.058	-0.040	0.009	0.025	0.042	0.033
Cohesion	0,113	0,324	0,118	**0,134**	0,158	0,197	0,259	0,305	0,000	0,312	0,058	0,210
kNN	0.842	0.250	0.626	0.590	0.950	**0.703**	0.731	0.373	0.988	0.358	0.986	0.685
SVM_{rbf}	0.238	0.258	0.238	0.258	0.714	0.363	0.474	0.375	0.738	0.310	0.984	**0.638**
SVM_{mlp}	0.237	0.230	0.249	0.230	0.349	0.358	0.348	0.300	0.738	0.378	0.824	**0.605**
DT	0.884	0.268	0.652	0.605	0.955	0.608	0.785	0.360	0.988	0.335	0.991	**0.683**

Table 2. Comparison of DR methods using 3D spaces: rank-based quality measures (Q/B), cohesion and classification accuracies of four classifiers.

	LDA		LDA_d		NCA		MCML		CCA		CMM	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Q	0.537	0.573	0.536	0.527	0.608	**0.618**	0.540	0.559	0.517	0.548	0.560	0.617
B	0.017	0.046	0.014	**0.019**	0.049	0.043	-0.071	-0.043	-0.004	0.026	0.079	0.040
Cohesion	0,085	0,324	0,149	**0,173**	0,157	0,202	0,253	0,298	0,006	0,314	0,081	0,213
kNN	0.872	0.268	0.716	0.673	0.976	**0.750**	0.773	0.450	0.991	0.415	0.990	0.743
SVM_{rbf}	0.238	0.258	0.233	0.263	0.622	0.310	0.755	0.405	0.811	0.265	0.974	**0.650**
SVM_{mlp}	0.237	0.230	0.246	0.248	0.363	0.328	0.411	0.305	0.833	0.325	0.971	**0.658**
DT	0.928	0.263	0.761	**0.723**	0.972	0.653	0.828	0.415	0.991	0.343	0.990	0.695

CMM either on the 2D or 3D projection is the most stable method, since it gets the first or second best values for all the metrics. More specifically, CMM is the only method that has a consistent classification accuracy when SVM is used. Additional results that were not included here for a matter of space can be looked up in [14].

5 Conclusions

Subspace mapping allows visualization of high-dimensional spaces on an informative plotting space, suitable for visual data mining methods. Additionally, projections into low-dimensional spaces allow a reduction of the storage of data points and lead to improved prediction capacity of a subsequently applied supervised method. We emphasize the advantages of applying linear subspace transformations, since they provide a simple interpretation of the new space. Moreover, they guarantee exact out-of-sample extensions. Methods that make use of calculation of eigenvectors may not be the best option when the input data dimensionality is considerably high.

This paper described the applicability of different DR methods for short and noisy text documents. This is the first work where CMM is compared against other well-established DR methods. From the results showed in Section 4 we can state that our proposed method CMM represents a competitive subspace mapping method, with the advantage of more stable behavior than the other methods tested in this work. NCA was its closest competitor, especially for Q and k-NN.

As future work, we plan to extend this development by considering a semi-supervised scenario. In this case the system can automatically classify documents and, at the same time, the user can provide its feedback about reclassifying a document or indicating the irrelevance of a term. Moreover, the system should adapt its behavior from the user feedback and correct future actions.

We thank NSERC, PGI-UNS (24/ZN16), the DFG Graduate School 1564, and MINCyT-BMBF (AL0811 - ARG 08/016) for their financial support.

References

1. Zhang, J., Huang, H., Wang, J.: Manifold Learning for Visualizing and Analyzing High-Dimensional Data. IEEE Intel. Syst. 25, 54–61 (2010)
2. van der Maaten, L., Postma, E., van den Herik, J.: Dimensionality Reduction: A Comparative Review. Tilburg University, TiCC TR 2009–005 (2009)
3. Strickert, M., Soto, A.J., Vazquez, G.E.: Adaptive Matrix Distances Aiming at Optimum Regression Subspaces. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2010, pp. 93–98 (2010)
4. Soto, A.J., Strickert, M., Vazquez, G.E., Milios, E.: Adaptive Visualization of Text Documents Incorporating Domain Knowledge. In: Challenges of Data Visualization, NIPS 2010 Workshop (2010)
5. Machine Learning Open Source Software, `http://mloss.org`
6. Matlab Statistics Toolbox, `http://www.mathworks.com/products/statistics/`
7. McLachlan, G.: Discriminant Analysis and Statistical Pattern Recognition. Wiley-Interscience, Hoboken (2004)
8. Hardoon, D.R., Szedmak, S.R., Shawe-Taylor, J.R.: Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Comput. 16, 2639–2664 (2004)
9. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighborhood Components Analysis. Adv. Neural Inf. Process. Syst. 17, 513–520 (2005)
10. Globerson, A., Roweis, S.: Metric Learning by Collapsing Classes. Adv. Neural Inf. Process. Syst. 18, 451–458 (2006)
11. Aviation Safety Reporting System, `http://asrs.arc.nasa.gov/`
12. Lee, J.A., Verleysen, M.: Quality Assessment of Dimensionality Reduction: Rank-Based Criteria. Neurocomputing 72, 1431–1443 (2009)
13. Dunnet, C.W.: A Multiple Comparisons Procedure for Comparing Several Treatments with a Control. J. Am. Stat. Assoc. 50, 1096–1121 (1955)
14. Soto, A.J., Strickert, M., Vazquez, G.E., Milios, E.: Technical Report, Dalhousie University (in preparation), `http://www.cs.dal.ca/research/techreports`