Automating the Detection of Anomalies and Trends from Text

Michael W. Berry
Department of Electrical Engineering
and Computer Science
203 Claxton Complex
University of Tennessee
Knoxville, TN 37996-3450
berry@eecs.utk.edu

Abstract

Scalable and robust nonnegative matrix factorization (NMF) algorithms and software are needed for the generation of feature vectors from text corpora. By preserving nonnegativity, the NMF facilitates a sum-of-parts representation of the underlying term usage patterns in textual data. Both training and test sets of documents can be parsed and then factored by the NMF to produce a reduced-rank representation of an entire document space. The resulting feature *and* coefficient *matrix factors are then used to cluster documents. Recent studies with documents from the Aviation Safety Reporting System (ASRS) have shown that (known) anomalies of training documents can be directly mapped to NMF-generated feature vectors. Dominant features (tracking words or sentences) of test documents can then be used to generate anomaly relevance scores for those documents.*

1 Introduction

Nonnegative matrix factorization (NMF) has been widely used to approximate high dimensional data comprised of nonnegative components. Lee and Seung [14] proposed the idea of using NMF techniques to generate basis functions for image data that could facilitate the identification and classification of objects. They also demonstrated the use of NMF to extract concepts/topics from unstructured text documents. This is the context that we exploit the so-called *sum-of-parts* representation offered by the NMF for corpora such as the Aviation Safety Reporting System (ASRS) document collection [1].

Several manuscripts have cited [14], but as pointed out in [2] there are several (earlier) papers by P. Paatero [18, 19, 20] that documented the historical development of the NMF. Simply stated, the problem defining the NMF can be stated as follows:

Given a nonnegative matrix $\mathbf{A} \in \mathbf{R^{m\times n}}$ and a positive integer $k < \min\{m, n\}$, find nonnegative matrices $\mathbf{W} \in \mathbf{R^{m\times k}}$ and $\mathbf{H} \in \mathbf{R^{k\times n}}$ to minimize the functional

$$f(\mathbf{W},\mathbf{H}) = \mathbf{\frac{1}{2}}\|\mathbf{A} - \mathbf{WH}\|_\mathbf{F}^\mathbf{2}. \tag{1}$$

The product $\mathbf{WH}$ is called a nonnegative matrix factorization of $\mathbf{A}$, although $\mathbf{A}$ is not necessarily equal to the product $\mathbf{WH}$. Although the product $\mathbf{WH}$ is an approximate factorization of rank at most k, we drop the word *approximate* in our discussions below. The best choice for the rank k is certainly problem dependent, and in most cases it is usually chosen such that $k \ll \min(m, n)$. Hence, the product $\mathbf{WH}$ can be considered a *compressed* form of the data in $\mathbf{A}$.

Another key characteristic of NMF is the ability of numerical methods that minimize Equation (1) to extract underlying features as basis vectors in $\mathbf{W}$, which can then be subsequently used for identification and classification. By not allowing negative entries in $\mathbf{W}$ and $\mathbf{H}$, NMF enables a non-subtractive combination of parts to form a whole [14]. Features may be parts of faces in image data, topics or clusters in textual data, or specific absorption characteristics in hyperspectral data. The focus of this discussion is in the enhancement of NMF algorithms for the primary goal of feature extraction and identification in text and spectral data mining.

Important challenges affecting the numerical minimization of Equation (1) include the existence of local minima due to the non-convexity of $f(\mathbf{W},\mathbf{H})$ in both $\mathbf{W}$ and $\mathbf{H}$. The non-uniqueness of its solution is easily realized by noting that $\mathbf{WDD^{-1}H}$ for any nonnegative invertible matrix $\mathbf{D}$ whose inverse, $\mathbf{D^{-1}}$, is also nonnegative. Fortunately, the NMF is still quite useful for text/data mining in practice since even local minima can provide desirable data compression and feature extraction and identification of both structured and unstructured text.

Alternative formulations of the NMF problem certainly arise in the literature. As surveyed in [2], an information theoretic formulation in [15] is based on the Kullback-Leibler divergence of $\mathbf{A}$ from $\mathbf{WH}$ and the cost functions proposed in [4] are based on Csiszár's φ-divergence. A formulation in [21] enforces constraints based on the Fisher linear discriminant analysis and [9] suggest using a diagonal weight matrix $\mathbf{Q}$ in the factorization model, $\mathbf{AQ} \approx \mathbf{WHQ}$, as an attempt to compensate for feature redundancy. For other approaches using alternative cost functions see [6] and [10].

In order to speed up convergence of Lee and Seung's (standard) NMF iteration, various alternative minimization strategies for Equation (1) have been suggested. For example, [17] propose the use of a projected gradient bound-constrained optimization method that presumably has better convergence properties than the standard multiplicative update rule approach. However, the use of certain auxiliary constraints in Equation (1) may break down the bound-constrained optimization assumption and thereby limit the use of projected gradient methods. Accelerating the standard approach via an interior-point gradient method has been suggested in [7], and a quasi-Newton optimization approach for updating $\mathbf{W}$ and $\mathbf{H}$, where negative values are replaced with small positive ϵ parameter to enforce nonnegativity, is discussed in [23]. A complete overview of enhancements to improve the convergence of the (standard) NMF algorithm is provided in [2].

Typically, $\mathbf{W}$ and $\mathbf{H}$ are initialized with random nonnegative values to start the standard NMF algorithm. Another area of NMF-related research has focused on alternate approaches for initializing or seeding the algorithm. The goal, of course, is to speed up convergence. In [22] spherical k-means clustering is used to initialize $\mathbf{W}$ and in [3] singular vectors of $\mathbf{A}$ are used for initialization and subsequent cost function reduction. Optimal initialization, however, remains an open research problem.

1.0.1 NMF Algorithm

As surveyed in [2], there are three general classes of NMF algorithms: multiplicative update algorithms, gradient descent algorithms, and alternating least squares algorithms. Here, we describe the most basic multiplicative update method (initially described in [15]). This approach, based on a mean squared error objective function, can be illustrated using MATLAB® array operator notation:

```
MULTIPLICATIVE UPDATE ALGORITHM FOR NMF
$\mathbf{W}$ = rand(m,k);   % $\mathbf{W}$ initially random
$\mathbf{H}$ = rand(k,n);   % $\mathbf{H}$ initially random
for i = 1 : maxiter
    $\mathbf{H}$ = $\mathbf{H}$ .* ($\mathbf{W^T A}$) ./ ($\mathbf{W^T W H}$ + $\epsilon$);
    $\mathbf{W}$ = $\mathbf{W}$ .* ($\mathbf{A H^T}$) ./ ($\mathbf{W H H^T}$ + $\epsilon$);
end
```

The parameter $\epsilon = 10^{-9}$ is added to avoid division by zero. As explained in [2], if this multiplicative update NMF algorithm converges to a stationary point, there is no guarantee that the stationary point is a local minimum for the objective function. Additionally, if the limit point to which the algorithm has converged lies on the the boundary of the feasible region, we cannot conclude that it is, in fact, a stationary point. A modification of the Lee and Seung multiplicative update scheme that resolves some of the convergence issues and guarantees convergence to a stationary point in provided in [16].

Three additional parameters needed for the classification of documents (by NMF) are: α, a threshold on the relevance score or (target value) t_{ij} for document i and anomaly/label j; δ, a threshold on the column elements of $\mathbf{H}$, which will filter out the association of features with both the training ($\mathbf{R}$) and test ($\mathbf{T}$) documents; and σ, the percentage of documents used to define the training set (or number of columns of $\mathbf{R}$).

Preliminary Studies

As reported in [1], a rank-40 model (i.e., $k = 40$) has been successfully used to classify anomalies in the ASRS collection. By rank, we refer to the number of columns of the feature matrix factor $\mathbf{W}$ used to test the NMF model (with training documents only). The $\mathbf{W}$ and $\mathbf{H}$ matrix factors, in that case, were $15,722 \times 40$ and $40 \times 21,519$, respectively. The percentage of ASRS documents used for training (subset $\mathbf{R}$) was 70% (i.e., $\sigma = .70$). Hence, $15,063$ documents were used as the initial training set ($\mathbf{R}$) and $6,456$ documents were used for testing ($\mathbf{T}$) the NMF classifier. Columnwise pruning of the elements in the coefficient matrix $\mathbf{H}$ was also tested with the setting $\delta = .30$. This parameter effectively determines the number of features (among the $k = 40$ possible) that any document (training or test) can be associated with. As δ increases, so does the sparsity of $\mathbf{H}$.

The α parameter, defined to be $.40$ in [1], is the prediction control parameter which ultimately determines whether or not document i will be given label (anomaly) j, i.e., whether $p_{ij} = +1$ or $p_{ij} = -1$ for the cost function

$$Q = \frac{1}{C}\sum_{j=1}^{C} Q_j, \qquad (2)$$

$$Q_j = (2A_j - 1) + \frac{1}{D}\sum_{i=1}^{D} q_{ij}t_{ij}p_{ij}, \qquad (3)$$

where C is the number of labels (anomalies) and D is the number of test documents. As mentioned above, $D = 6,456$ in the preliminary evaluation of the NMF classifier and $C = 22$. The cost Q given by Equation 2 in preliminary NMF testing would usually lie in the interval $[1.28, 1.30]$. To measure the quality of (anomaly) predictions across all $C = 22$ categories, a Figure of Merit (FOM) score defined by

$$FOM = \frac{1}{C}\sum_{j=1}^{C}\frac{F - F_j}{F}Q_j,\ F = \sum_{j=1}^{C} F_j, \qquad (4)$$

where F_j denotes the frequency of documents having label (anomaly) j, can be generated for each experiment. By definition, the FOM score will assign lower weights to the higher frequency labels or categories. The best FOM score for $\sigma = .70$ was 1.267 to three significant decimal digits. Keep in mind that the initial matrix factors $\mathbf{W}$ and $\mathbf{H}$ are randomly generated and will produce slightly different features (columns of $\mathbf{W}$) and coefficients (columns of $\mathbf{H}$) per NMF iteration[1].

Recent Contest Results

For the text mining contest (sponsored by NASA Ames Research Center) at the Seventh SIAM International Conference on Data Mining in Minneapolis, MN (April 26–28, 2007), all contestants were provided an additional $7,077$ ASRS unclassified documents. The top three contest entries in anomaly classification deployed Probabilistic Latent Semantic Analysis (PLSA) [8], Nonnegative Matrix Factorization (NMF) [1], and a Rich Document Representation (RDR) with the traditional Vector Space Model (VSM) [13].

For the NMF classifier described in [1], the *new* documents were considered the test subset $\mathbf{T}$ and the training subset $\mathbf{R}$ was defined by the previously available (classified) documents ($21,519$ of them). Since all of the previously classified ASRS documents were used in the term-by-document matrix $\mathbf{A}$ for the contest entry, the σ parameter was set to 1.0. The other two parameters for the NMF classifier were not changed, i.e., $\alpha = 0.40$ and $\delta = 0.30$ (see Section 1.0.1). Using 5 iterations and $k = 40$ features for the multiplicative update algorithm mentioned in Section 1.0.1, a cost of $Q = 1.27$ (see Equation 2) was reported by contest officials for the NMF classifier in mapping each of the $7,077$ test documents to any of the 22 anomaly categories was 1.27 (a second place finish). Had a tie occurred among any of the cost function values generated by contest entries, the FOM score would have been used to break it. For the NMF classifier, the average contest FOM score was 1.22 (slightly lower than what was observed in the preliminary testing phase).

ROC Curves

Figure 1 illustrates some of the Receiver Operating Characteristic (ROC) curves for the NMF classifier [1] used in the text mining competition mentioned earlier. Although not shown here, a comparison of graphs for the preliminary testing and the actual contest entry reveals similar performance for a majority of the 22 anomaly classes.

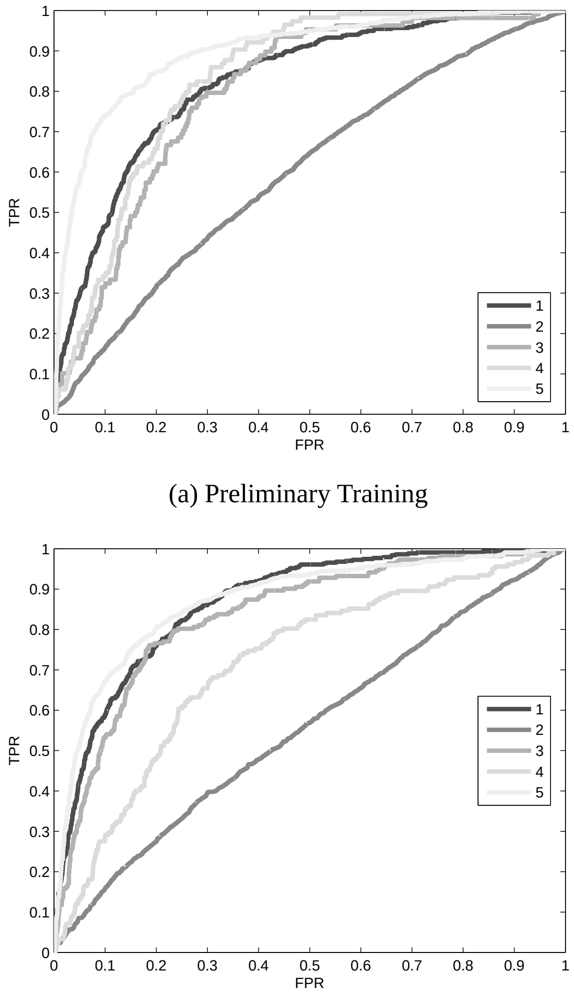
Thirteen (of the twenty-two) event types (or anomaly descriptions) listed in Table 1 were obtained from the Distributed National ASAP Archive (DNAA) maintained by the University of Texas Human Factors Research Project[2]. The generality of topics described in the ASRS reports of the *Noncompliance* (anomaly 2), *Uncommanded (loss of control)* (anomaly 10), and *Weater Issue* (anomaly 13) categories greatly contributed to the poorer performance of the NMF classifier. Additional experiments with a larger numbers of features ($k > 40$) may produce an NMF model that would better capture the diversity of contexts described by those events. More experimentation is certainly needed.

Table 1. ROC Areas Versus DNAA Event Types for Selected Anomalies [1]

		ROC Area	
Anom.	**DNAA Event Type**	**Training**	**Contest**
22	Security Concern/Threat	.904	.892
5	Incursion (collision hazard)	.897	.871
4	Excursion (loss of control)	.829	.715
21	Illness/Injury Event	.820	.817
12	Traffic Proximity Event	.795	.775
7	Altitude Deviation	.793	.808
15	Approach/Arrival Problems	.751	.672
18	Aircraft Damage/Encounter	.725	.726
11	Terrain Proximity Event	.723	.757
9	Speed Deviation	.706	.689
10	Uncommanded (no control)	.678	.650
13	Weather Issue	.628	.601
2	Noncompliance (policy)	.600	.555

[1] Only 5 iterations were used in the preliminary study documented in [1].

[2] See http://homepage.psy.utexas.edu/HomePage/Group/HelmreichLAB.



(a) Preliminary Training

(b) Contest Performance

Figure 1. ROC curves for the NSF classifier applied to anomalies (labels) 1 through 5.

1.0.2 Control Vocabulary Development

Nonnegative matrix factorization (NMF) is a viable alternative for automated document classification problems. As the volume and heterogeneity of documentation continues to grow, the ability to discern common themes and contexts can be problematic. Current research has demonstrated that NMF can be used to both learn and assign (anomaly) labels for collections such as the Aviation Safety Reporting System (ASRS). However, there is room for improvement in both the performance and interpretability of the NMF. In particular, the the summarization of anomalies (document classes) using k NMF features needs further work. Alternatives to the filtering of elements of the coefficient matrix $\mathbf{H}$ (based on the parameter δ) could be the use of sparsity or smoothing constraints (see [2]) on either (or both) factors $\mathbf{W}$ and $\mathbf{H}$. Of particular interest are the effects that sparsity and/or smoothing constraints may have on the conservation of terms/tokens of *high* information content (e.g., larger entropy global weight).

Penalty terms can be used to enforce smoothing constraints via the modified objective function

$$f(\mathbf{W},\mathbf{H}) = \|\mathbf{A}-\mathbf{WH}\|_{\mathbf{F}}^{\mathbf{2}} + \beta\mathbf{J_1}(\mathbf{W}) + \gamma\mathbf{J_2}(\mathbf{H}), \quad (5)$$

where $J_1(\mathbf{W})$ and $J_2(\mathbf{H})$ are the penalty terms introduced to enforce certain application-dependent constraints, and β and γ are small regularization parameters that balance the trade-off between the approximation error and the constraints. Possible measures for sparsity include, for example, the ℓ^p norms for $0 < p \leq 1$ [12] and Hoyer's measure [11],

$$\text{sparseness}(\mathbf{x}) = \frac{\sqrt{\mathbf{n}} - \|\mathbf{x}\|_{\mathbf{1}}/\|\mathbf{x}\|_{\mathbf{2}}}{\sqrt{\mathbf{n}}-\mathbf{1}}.$$

The latter can be imposed as a penalty term of the form

$$J_2(\mathbf{H}) = (\omega\|\text{vec}(\mathbf{H})\|_{\mathbf{2}} - \|\text{vec}(\mathbf{H})\|_{\mathbf{1}})^{\mathbf{2}}, \quad (6)$$

where $\omega = \sqrt{kn} - (\sqrt{kn}-1)\gamma$ and $\text{vec}(\cdot)$ is the vec operator that transforms a matrix into a vector by stacking its columns. The desired sparseness in $\mathbf{H}$ is specified by setting γ to a value between 0 and 1. Initial testing with both smoothing and sparsity constraints on NMF models of the Enron email collection [5] suggests that small control vocabularies of conserved high-entropy (weighted) terms can discern important topics in time-sensitive documents (such as email and blog entries).

References

[1] E. Allan, M. Horvath, C. Kopek, B. Lamb, T. Whaples, and M. Berry. Anomaly Detection Using Non-negative Matrix Factorization. In M. Berry and M. Castellanos, editors, *Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition*. Springer, New York, 2007. to appear.

[2] M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons. Algorithms and Applications for Approximate Nonnegative Matrix Factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.

[3] C. Boutsidis and E. Gallopoulos. On SVD-based initialization for nonnegative matrix factorization. Technical Report HPCLAB-SCG-6/08-05, University of Patras, Patras, Greece, 2005.

[4] A. Cichocki, R. Zdunek, and S. Amari. Csiszar's Divergences for Non-Negative Matrix Factorization: Family of New Algorithms. In *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation*, Charleston, SC, March 5-8 2006.

[5] W. W. Cohen. Enron email dataset. Webpage. `http://www.cs.cmu.edu/~enron/`.

[6] I. Dhillon and S. Sra. Generalized Nonnegative Matrix Approximations with Bregman Divergences. In *Proceeding of the Neural Information Processing Systems (NIPS) Conference*, Vancouver, B.C., 2005.

[7] E. Gonzalez and Y. Zhang. Accelerating the Lee-Seung Algorithm for Nonnegative Matrix Factorization. Technical Report TR-05-02, Rice University, March 2005.

[8] C. Goutte. A Probabilistic Model for Fast and Confident Categorisation of Textual Documents. In M. Berry and M. Castellanos, editors, *Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition*. Springer, New York, 2007. to appear.

[9] D. Guillamet, M. Bressan, and J. Vitria. A Weighted Nonnegative Matrix Factorization for Local Representations. In *Proc. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 942–947, Kavai, HI, 2001.

[10] A. Hamza and D. Brady. Reconstruction of Reflectance Spectra Using Robust Non-Negative Matrix Factorization. *IEEE Transactions on Signal Processing*, 54(9):3637–3642, 2006.

[11] P. Hoyer. Non-negative Matrix Factorization with Sparseness Constraints. *J. of Machine Learning Research*, 5:1457–1469, 2004.

[12] J. Karvanen and A. Cichocki. Measuring Sparseness of Noisy Signals. In *Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, 2003.

[13] M. Keikha, N. Razavian, F.Oroumchian, and H. Razi. Document Representation and Quality of Text: An Analysis. In M. Berry and M. Castellanos, editors, *Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition*. Springer, New York, 2007. to appear.

[14] D. Lee and H. Seung. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 401:788–791, 1999.

[15] D. Lee and H. Seung. Algorithms for Non-Negative Matrix Factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.

[16] C.-J. Lin. On the Convergence of Multiplicative Update Algorithms for Non-negative Matrix Factorization. Technical Report Information and Support Services Techincal Report, Department of Computer Science, National Taiwan University, 2005.

[17] C.-J. Lin. Projected gradient methods for non-negative matrix factorization. Technical Report Information and Support Services Technical Report ISSTECH-95-013, Department of Computer Science, National Taiwan University, 2005.

[18] P. Paatero. Least Squares Formulation of Robust Non-negative Factor Analysis. *Chemometrics and Intelligent Laboratory Systems*, 37:23–35, 1997.

[19] P. Paatero. The Multilinear Engine — A Table-Driven Least Squares Program for Solving Multilinear Problems, including the n-Way Parallel Factor Analysis Model. *Journal of Computational and Graphical Statistics*, 8(4):1–35, 1999.

[20] P. Paatero and U. Tapper. Positive Matrix Factorization: A Non-negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics*, 5:111–126, 1994.

[21] Y. Wang, Y. Jiar, C. Hu, and M. Turk. Fisher non-negative matrix factorization for learning local features. In *Asian Conference on Computer Vision*, Korea, January 27-30 2004.

[22] S. Wild, J. Curry, and A. Dougherty. Motivating Non-Negative Matrix Factorizations. In *Proceedings of the Eighth SIAM Conference on Applied Linear Algebra, July 15-19*, Williamsburg, VA, 2003. SIAM. http://www.siam.org/meetings/la03/proceedings.

[23] R. Zdunek and A. Cichocki. Non-Negative Matrix Factorization with Quasi-Newton Optimization. In *Proc. Eighth International Conference on Artificial Intelligence and Soft Computing, ICAISC*, Zakopane, Poland, June 25-29 2006.