Dimension Reduction Using Nonnegative Matrix Tri-Factorization in Multi-label Classification

Keigo Kimura[1,a)] Mineichi Kudo[1] Lu Sun[1]

Abstract: Multi-label classification problem has become more important in image processing and text analysis where an object often is associated with many labels at the same time. Recently, even in this problem setting dimension reduction aiming at avoiding *the curse of dimensionality* has gathered an attention, but it is still a challenging problem. Nonnegative Matrix Factorization (NMF) is one of promising ways for dimension reduction in unsupervised learning, and is extended from two-matrix factorization to triple-matrix factorization. In this paper, we reformulate the NMF with three factor matrices in such a way that it is solvable the problem of the combinatorial explosion of labels and incorporates the label correlation naturally in supervised learning. Experiments on web page classification datasets show the advantages of the proposed algorithm in the classification accuracy and computational time.

1. Introduction

Multi-label classification has attracted much attention in a variety of fields such as text analysis, image analysis and recommendations [10]. This is because an object often has several labels simultaneously, for example, a document may belong to *politics* and *economics*. A multi-label multi-class problem can be transformed into a set of independent single-label binary-class problems. However in that case, the relation between classes is lost.

Dimensional reduction is an essential technique in the field of machine learning and it aims to avoid *the curse of dimensionality*. The methods for dimension reduction are classified into either unsupervised or supervised method. The unsupervised methods such as Principal Component Analysis (PCA) [7] and Nonnegative Matrix Factorization (NMF) [8] reduce the dimension of the feature space ignoring the class information, while the supervised methods such as Linear Discriminant Analysis (LDA) aim to keep the class separability even in the reduced feature space. Recently, some supervised dimension reduction methods have been proposed even for multi-label classification [13, 15, 18]. The key idea in common is to keep the label dependency as possible in the reduced space.

NMF is one of the unsupervised dimension reduction methods and decomposes a given nonnegative matrix into a product of two lower-ranked nonnegative matrices [8]. It is reported that NMF outperforms PCA in the interpretability and even in the classification accuracy [3]. Its supervised version, NMF-LDA [16], is more advantageous in the classification accuracy. However, such supervised NMF algorithms are all only applicable to single-label classification and hard to be simply extended to multi-label classification for the difficulty to solve a set of binary-class problems in single dimension reduction scheme.

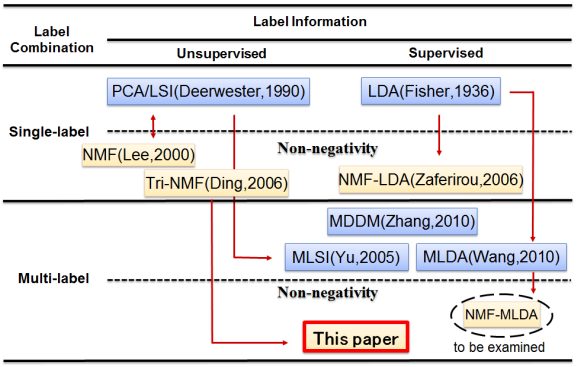


Fig. 1 The position of this study: a supervised multi-label dimension reduction method with nonnegative constraint for the elements.

In this paper, we cope with this difficulty by proposing a multi-label NMF with the idea of tri-factorization. As seen in Fig. 1. this study is the first nonnegative supervised multi-label dimension reduction method. Our goal in this study is to find an effective representation of nonnegative data matrix with corresponding label matrix, while taking into consideration the multi-label information and the label dependency at the same time. Nonnegativity is imposed from an emprical knowledge that the nonnegativity, and the sparsity induced by the nonnegative constraint, has been to the improvement of classification in the past of study of NMF [1, 6, 14]. We borrow the idea of Nonnegative Matrix Tri-Factorization (NMTF) proposed by Ding *et al.* [5], one of unsupervised algorithms, which decompose a nonnegative matrix into a product of nonnegative three matrices. In this study, we decompose a data matrix into three factor matrices that have their own roles in data approximation.

1.1 Notations

We use $\mathbf{X}$ for a matrix and $\boldsymbol{x}$ for a vector. In multi-label clas-

[1] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, 060-0814, Japan
[a)] kkimura@main.ist.hokudai.ac.jp

sification, a sample $\boldsymbol{x}$ is associated with a subset $\boldsymbol{y}$ of class labels. We consider N training samples and L labels. Each sample $\boldsymbol{x}_i$ belongs to an M-dimension space and the associated label subset is represented as a binary vector $\boldsymbol{y}_i \in \{0,1\}^L$. We denote $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{M\times N}$ as a data matrix and $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N] \in \{0,1\}^{L\times N}$ as a label matrix.

2. Related Work

According to Fig. 1, we review the methodology proposed so far.

Unsupervised Dimensionality Reduction Methods

Latent Semantic Indexing (LSI) is one of popular unsupervised dimension reduction methods [4]. LSI decomposes a matrix $\mathbf{X}$ into a product of two low-rank matrices $\mathbf{AB}$ by solving an eigen problem. Nonnegative Matrix Factorization (NMF) is another unsupervised method [8]. In NMF, the two factor matrices are required to be of nonnegative elements and this leads sparser matrices as a side effect. There are some reports saying that NMF outperforms PCA in the accuracy of classification by the virtue of the sparse representation [1, 3, 6, 14].

Supervised Dimensionality Reduction Methods for Single-label Classification

Linear Discriminant Analysis (LDA) is a supervised dimension reduction method for single-label multi-class classification [9]. It finds a subspace so as to maximize the ratio of between-class distance to the within-class distance. Zafeiriou *et al.* coupled NMF and LDA to produce NMF-LDA [16]. They aimed to realize both the sparsity, inheritance of NMF, and the class separability, inheritance of LDA. They constructed an objective function to achieve both in NMF-LDA.

Supervised Dimensionality Reduction Methods for Multi-label Classification

Yu *et al.* firstly conducted dimension reduction for multi-label classification problem and proposed a method called Multi-Label Informed Latent Semantic Indexing (MLSI) [15]. This method decomposes data matrix and label matrix at the same time with sharing a matrix. The shared matrix bridges the approximation information and the label information. Zhang *et al.* proposed another method called Multi-label Dimensionality reduction via Dependence Maximization (MDDM) [18]. MDDM finds a subspace so as to maximize the dependency between the features and the associated labels. Wang *et al.* proposed Multi-label LDA (MLDA) as a generalization of Linear Discriminant Analysis (LDA) [13]. They redesigned the scatter matrices so as to handle multi-label setting. Other than redesigned scatter matrices, MLDA is the same as LDA. Therefore, it is straightforward to couple NMF with MLDA, such as NMF was coupled with LDA to produce NMF-MLDA, but we leave such a trial for the future work.

3. Multi-label Informed Nonnegative Matrix Tri-Factorization

We first explain the key idea of the proposed approach. In unsupervised dimension reduction, we usually consider to approximate a data point $\boldsymbol{x} \in \mathbb{R}^M$ by a linear combination of a small number of bases $\boldsymbol{u}_j \in \mathbb{R}^M$, $j = 1, 2, \ldots, J$ ($J \ll M$), as

$$\boldsymbol{x} \cong \hat{\boldsymbol{x}} = a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2 + \cdots + a_J\boldsymbol{u}_J,$$

where $a_j \in \mathbb{R}$, $j = 1, 2, \ldots, J$ are the coefficients depending on $\boldsymbol{x}$. If samples belonging to the same class concentrate on around a representative point of the class, the number J could be identical to the number L of classes as long as single labeled samples are only considered. In multi-label problems, since such a sample $\boldsymbol{x}$ is associated with a subset of labels, the number of possible (extended) classes becomes 2^L in the same scenario. It is, therefore, infeasible to find a low-dimensional subspace, a small value of J. To cope with this problem, we take the following approach. We first assume a multi-labeled mean vector $\boldsymbol{m_y} \in \mathbb{R}^M$ is expressed by a linear combination of single-labeled mean vectors as

$$\boldsymbol{m_y} = y_1\boldsymbol{m}_1 + y_2\boldsymbol{m}_2 + \cdots + y_L\boldsymbol{m}_L,\ \boldsymbol{y} = (y_1, y_2, \ldots, y_L)^T \quad (1)$$

In addition, we consider to express the single-labeled mean vectors by J bases as

$$\boldsymbol{m}_l = s_{1l}\boldsymbol{u}_1 + s_{2l}\boldsymbol{u}_2 + \cdots + s_{Jl}\boldsymbol{u}_l, \quad l = 1, 2, \ldots, L. \quad (2)$$

From (1) and (2), we can write $\boldsymbol{m_y}$ by

$$\begin{aligned}\boldsymbol{m_y} &= (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_J)\begin{pmatrix} s_{11} & \cdots & s_{1L} \\ \vdots & \ldots & \vdots \\ s_{J1} & \cdots & s_{JL} \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{pmatrix} \\ &= \mathbf{U}\mathbf{S}\boldsymbol{y}.\end{aligned}$$

For each pair $(\boldsymbol{x}, \boldsymbol{y})$ given as a training sample, we try to approache $\hat{\boldsymbol{x}} = \boldsymbol{m_y} = \mathbf{U}\mathbf{S}\boldsymbol{y}$ to $\boldsymbol{x}$ by choosing $\mathbf{U}$ and $\mathbf{S}$ appropriately. Once bases $\mathbf{U}$ is determined, we project $\boldsymbol{x}$ on the subspace spanned by $\mathbf{U}$ for dimension reduction.

3.1 Problem Formulation

In the following objective function to minimize, we expect that all training data $\boldsymbol{x}$ associated with $\boldsymbol{y}$ are distributed near the mean vector $\boldsymbol{m_y}$. In addition, we expect the mean vectors of frequently co-occurred classes are closely located. As a result, we find $\mathbf{U}$ and $\mathbf{S}$ minimizing

$$J(\mathbf{U}, \mathbf{S}) = \|\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{Y}\|_F^2 + \lambda\mathrm{tr}(\mathbf{S}\mathbf{L}\mathbf{S}^T), \quad (3)$$

where $\|\cdot\|_F$ denotes Frobenius norm, $\mathrm{tr}(\cdot)$ denotes the trace and λ is a positive coefficient. Here, $\mathbf{L}$ is the graph Laplacian matrix defined as $\mathbf{L} = \mathbf{K} - \mathbf{D}$ where $\mathbf{K} = \mathbf{Y}^T\mathbf{Y}$, and $\mathbf{D}$ is the diagonal matrix whose lth elements is $\mathbf{D}_{ll} = \sum_{j=1}^{L}\mathbf{K}_{jl}$. This penalty term is called graph regularization which aims to preserve the locality provided by $\mathbf{K}$ on the subspace spanned by $\mathbf{U}$ [2]. That is, to minimize the second term, the mean vectors $\boldsymbol{m}_j$ and $\boldsymbol{m}_l$ should be close for frequently co-occurred jth and lth classes.

3.2 Optimization

Since NMF problems are NP-hard [12], we optimize the component matrices alternatively as EM algorithm does. We use multiplicative update rules algorithm [8]. According to [3, 8], the multiplicative update rule of $\mathbf{A}$ for minimizing $\|\mathbf{X} - \mathbf{A}\mathbf{B}\|_F^2$ are expressed in general as

Algorithm 1 Multi-label Nonnegative Matrix Tri-Factorization (MNMTF)

1: **Input:** Nonnegative matrix $\mathbf{X}$ and binary label matrix $\mathbf{Y}$; Weighting parameter for label correlation λ; The number of bases J;
2: **Output:** Nonnegative matrices $\mathbf{U}$ and $\mathbf{S}$ minimizing $\|\mathbf{X} - \mathbf{USY}\|_F^2 + \lambda \mathrm{tr}(\mathbf{SLS}^T)$;
3: Initialize $\mathbf{U}$ and $\mathbf{S}$ by random positive values;
4: **repeat**
5: $\quad \mathbf{U} = \mathbf{U} * \frac{\mathbf{XY}^T\mathbf{S}^T}{\mathbf{USYY}^T\mathbf{S}^T}$.
6: $\quad \mathbf{S} = \mathbf{S} * \frac{\mathbf{U}^T\mathbf{XY}^T + \lambda\mathbf{SK}}{\mathbf{U}^T\mathbf{USYY}^T + \lambda\mathbf{SD}}$.
7: **until** Convergence criterion is met

$$\mathbf{A} = \mathbf{A} * \frac{\nabla_A^-}{\nabla_A^+},$$

where $*$ and $/$ is the element-wise multiplication and division, respectively. Here, ∇_A^+ and ∇_A^- are the positive term and the negative term of the gradient of $\|\mathbf{X}-\mathbf{AB}\|_F^2$ in $\mathbf{A}$, respectively. The gradient of (3) in $\mathbf{U}$ and $\mathbf{S}$ are calculated as follows:

$$\nabla_U = -2\mathbf{XY}^T\mathbf{S}^T + 2\mathbf{USYY}^T\mathbf{S}^T,$$
$$\nabla_S = -2\mathbf{U}^T\mathbf{XY}^T + 2\mathbf{U}^T\mathbf{USYY}^T - 2\lambda\mathbf{SK} + 2\lambda\mathbf{SD}.$$

Hence, we update $\mathbf{U}$ and $\mathbf{S}$, respectively:

$$\mathbf{U} = \mathbf{U} * \frac{\mathbf{XY}^T\mathbf{S}^T}{\mathbf{USYY}^T\mathbf{S}^T} \quad \text{and} \quad \mathbf{S} = \mathbf{S} * \frac{\mathbf{U}^T\mathbf{XY}^T + \lambda\mathbf{SK}}{\mathbf{U}^T\mathbf{USYY}^T + \lambda\mathbf{SD}}.$$

The pseudo-code of the proposed Multi-label Nonnegative Matrix Tri-Factorization (MNMTF) algorithm is shown in Algorithm 1.

After obtaining the subspace $\mathbf{U}$, we project all training and testing samples into the subspace by solving the following minimization problem with nonnegative constraint:

$$\|\boldsymbol{x} - \mathbf{U}\boldsymbol{v}\|^2.$$

We use this $\boldsymbol{v} \in \mathbb{R}^J$ as the new representation of original sample $\boldsymbol{x} \in \mathbb{R}^M$ in both training and test phases. In the training phase, $\boldsymbol{v}$ is given by $\boldsymbol{v} = \mathbf{S}\boldsymbol{y}$ for a pair $(\boldsymbol{x}, \boldsymbol{y})$.

3.3 Computational Complexity

All traditional algorithms such as MLSI, MDDM and MDDM need $O(M^2N)$ or $O(M^3)$ to obtain the subspace $\mathbf{U}$. On the other hand, the proposed algorithm needs $O(MNL)$. In most cases, the number L of labels is smaller than both the dimension M of data and the number N of training samples L. Thus, the proposed algorithm is faster than these traditional algorithms in such cases.

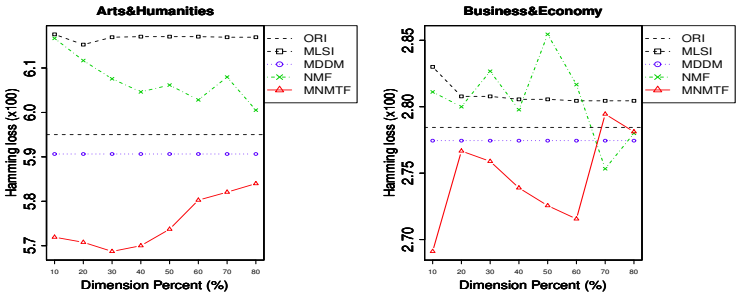
In the projection step, all traditional algorithms need $O(M(N+K)J)$ to project N training samples and K test samples. The proposed algorithm also needs $O(M(N+K)J)$, however, in practice it needs several repeatations of matrix multiplications due to the nonnegativity constraint and non-orthogonality of $\mathbf{U}$. Thus, the actual computation time is a little more than those of the traditional algorithms.

4. Experiments

We evaluated the performance of the proposed algorithm through experiments on web-page classification.

Table 1 A Summary of Dataset

Top-category	#Labels (L)	#Words (M)
Arts&Humanities	26	2315
Business&Economy	30	2192
Computers&Internet	33	3410



(a) *Arts&Humanities* (b) *Business&Economy*

Fig. 2 Classification performance in dimension reduction (Hamming loss comparison). We omitted the result of MLDA due to the worse performance.

4.1 Dataset

We used a yahoo web page classification dataset [11]. The summary of dataset are shown in Table 1. We followed the setting used in [17]. For each dataset, we randomly picked up 2,000 samples for training and 3,000 for testing. Top 10% most frequently occurred words were chosen as features. For more detail, see [17].

4.2 Results

We compared the proposed algorithm (MNMTF) with the other three multi-label dimension reduction methods, MLSI [15], MDDM [18], and MLDM [13]. In addition, we also compared with the original feature space (ORI) without dimension reduction and an unsupervised standard NMF without multi-label information (NMF) [8]. The weighting parameter β of MLSI β was set to the recommended value $\beta = 0.5$ [15]. In the proposed algorithm (MNMTF), we set $\lambda = 0.1$. We used a Multi-label k-Nearest Neighbor (ML-kNN) for the classifier after dimension reduction [17] with default settings.

In multi-label classification, several measures of performance are used at the same time instead of a single measure,*e.g.* the error rate used in single-label classification. We used four popular criteria of *hamming loss, one-error, coverage* and *average precision* [17]. We averaged the results of five-subsets in each dataset.

We varied the number of dimensionality $J = 0.1M, 0.2M, \ldots, 0.8M$. The results on "Arts&Humanities" dataset and "Business&Economy" are shown in Fig. 2. We note that NMF, MLSI and MLDA failed to improve the classification performance by reducing the dimensionality. The proposed MNMTF succeeded in total, although it is a little sensitive the value of J. We show the result with 30% dimension on Table 2. This was the best setting on "Arts&Humanities" dataset. We can see that the proposed algorithm MNMTF is almost the best in all the criteria, although the degree of improvement is only slightly more.

Table 2 Results on yahoo web page classification datasets. the bold figures show the best score. The symbol +/− larger/smaller is better in the criterion.

Dataset	Evaluation criterion	Compared methods					
		ORI	MLDA [13]	MLSI [15]	MDDM [18]	NMF [8]	MNMTF(Proposal)
Arts&Humanities	Hamming loss (−)	0.060	0.111	0.061	0.059	0.061	**0.058**
	One-error (−)	0.608	0.847	0.618	**0.565**	0.609	0.608
	Coverage (−)	6.269	16.815	6.468	6.395	6.364	**5.978**
	Average precision (+)	0.360	0.154	0.353	0.369	0.358	**0.383**
Business&Economy	Hamming loss	0.028	0.065	0.028	0.028	0.028	**0.027**
	One-error	0.119	0.596	0.123	0.126	0.122	**0.116**
	Coverage	4.023	18.598	4.053	4.125	4.030	**3.930**
	Average precision	0.394	0.193	0.393	0.391	0.394	**0.396**
Computers&Internet	Hamming loss	0.038	0.070	0.042	0.041	0.040	**0.036**
	One-error	0.421	0.702	0.429	0.404	0.413	**0.402**
	Coverage	5.343	18.193	5.606	5.607	5.491	**5.285**
	Average precision	0.389	0.193	0.382	0.388	0.388	**0.396**

5. Conclusion

In this paper, we have proposed a supervised Nonnegative Matrix Factorization algorithm for multi-label classification problems. The key idea is to formulate a supervised multi-label problem as a factorization problem of a given data matrix into three nonnegative factor matrices one of which is a give label matrix. The results on text classification showed the advantages in classification accuracy and computational time compared with the state-of-the-art methods.

References

[1] Cai, D., He, X. and Han, J.: Document clustering using locality preserving indexing, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 12, pp. 1624–1637 (2005).

[2] Cai, D., He, X., Han, J. and Huang, T. S.: Graph regularized nonnegative matrix factorization for data representation, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 33, No. 8, pp. 1548–1560 (2011).

[3] Cichocki, A., Zdunek, R., Phan, A. H. and Amari, S.-i.: *Nonnegative matrix and tensor factorizations: applications to exploratory multiway data analysis and blind source separation*, Wiley. com (2009).

[4] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A.: Indexing by latent semantic analysis, *JAsIs*, Vol. 41, No. 6, pp. 391–407 (1990).

[5] Ding, C., Li, T., Peng, W. and Park, H.: Orthogonal nonnegative matrix t-factorizations for clustering, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 126–135 (2006).

[6] Guillamet, D., Schiele, B. and Vitria, J.: Analyzing non-negative matrix factorization for image classification, *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, Vol. 2, IEEE, pp. 116–119 (2002).

[7] Jolliffe, I.: *Principal component analysis*, Wiley Online Library (2002).

[8] Lee, D. D. and Seung, H. S.: Learning the parts of objects by nonnegative matrix factorization, *Nature*, Vol. 401, No. 6755, pp. 788–791 (1999).

[9] Scholkopft, B. and Mullert, K.-R.: Fisher discriminant analysis with kernels, *Proceedings of the 1999 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing IX, Madison, WI, USA*, pp. 23–25 (1999).

[10] Tsoumakas, G. and Katakis, I.: Multi-label classification: An overview, *Dept. of Informatics, Aristotle University of Thessaloniki, Greece* (2006).

[11] Ueda, N. and Saito, K.: Parametric mixture models for multi-labeled text, *Advances in neural information processing systems*, pp. 721–728 (2002).

[12] Vavasis, S. A.: On the complexity of nonnegative matrix factorization, *SIAM Journal on Optimization*, Vol. 20, No. 3, pp. 1364–1377 (2009).

[13] Wang, H., Ding, C. and Huang, H.: Multi-label linear discriminant analysis, *Computer Vision–ECCV 2010*, Springer, pp. 126–139 (2010).

[14] Xu, W., Liu, X. and Gong, Y.: Document clustering based on nonnegative matrix factorization, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, ACM, pp. 267–273 (2003).

[15] Yu, K., Yu, S. and Tresp, V.: Multi-label informed latent semantic indexing, *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 258–265 (2005).

[16] Zafeiriou, S., Tefas, A., Buciu, I. and Pitas, I.: Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification, *IEEE Transactions on Neural Networks*, Vol. 17, No. 3, pp. 683–695 (2006).

[17] Zhang, M.-L. and Zhou, Z.-H.: ML-KNN: A lazy learning approach to multi-label learning, *Pattern recognition*, Vol. 40, No. 7, pp. 2038–2048 (2007).

[18] Zhang, Y. and Zhou, Z.-H.: Multilabel dimensionality reduction via dependence maximization, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 4, No. 3, p. 14 (2010).