Dimensionality Reduction with Category Information Fusion and Non-negative Matrix Factorization for Text Categorization

Wenbin Zheng[1,2], Yuntao Qian[1,*], and Hong Tang[3,4]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2] College of Information Engineering, China Jiliang University, Hangzhou, China
[3] School of Aeronautics and Astronautics, Zhejiang University, Hangzhou, China
[4] College of Metrological Technology & Engineering, China Jiliang University, Hangzhou, China

Abstract. Dimensionality reduction can efficiently improve computing performance of classifiers in text categorization, and non-negative matrix factorization could map the high dimensional term space into a low dimensional semantic subspace easily. Meanwhile, the non-negative of the basis vectors could provide a meaningful explanation for the semantic subspace. However, it usually could not achieve a satisfied classification performance because it is sensitive to the noise, data missing and outlier as a linear reconstruction method. This paper proposes a novel approach in which the train text and its category information are fused and a transformation matrix that maps the term space into a semantic subspace is obtained by a basis orthogonality non-negative matrix factorization and truncation. Finally, the dimensionality can be reduced aggressively with these transformations. Experimental results show that the proposed approach remains a good classification performance in a very low dimensional case.

Keywords: Text Categorization, Dimensionality reduction, Non-negative Matrix Factorization, Category Fusion.

1 Introduction

Text categorization (TC) is a task of automatically assigning predefined categories to a given text document based on its content [1]. Generally, the classical text representation method based on machine learning techniques is the vector space model (VSM) [2] in which the high dimensionality of the input feature space is a major difficulty of TC [3].

The latent semantic indexing (LSI) [4] and the topic model [5] are commonly used dimensionality reduction methods which can map a term space into a latent semantic subspace; however, it is difficult to explain the physic meaning because the negative value is permitted in its basis vectors.

[*] Corresponding author.

H. Deng et al. (Eds.): AICI 2011, Part III, LNAI 7004, pp. 505–512, 2011.

Non-negative matrix factorization (NMF) is a matrix factorization method with a non-negative constrain [6], which can map the term space to the semantic subspace for TC [7]. Mathematically, a terms-by-documents matrix X can be decomposed as $X_{m\times n} \approx W_{m\times r} \times H_{r\times n}$, where m and n are the number of the terms and documents respectively, and r is a positive integer, W is called basis matrix and H is called coefficient matrix. Because each column vector of W is constituted with some non-negative values of all terms, it can be regarded as the latent semantic basis vector, and all these basis vectors span a semantic subspace with dimensionality r. When $r \ll \min\{m, n\}$, the dimensionality of the semantic subspace is far less than the dimensionality of the original term space. However, as a linear reconstructed method, it usually could not achieve a satisfied classification performance in that subspace because it is sensitive to the noise, data missing and outlier which will affect the discriminative ability of basis vectors.

This paper proposes a novel method to reduce the dimensionality aggressively by fusing category information and with basis orthogonality non-negative matrix factorization and truncation.

We give a category coding schema and fusing the weighting of training documents with their category coding into an extended matrix. Therefore the extended dimension generated by the category information can decrease the impact of the noise, data missing and outlier. After that, a NMF iteration algorithm is designed in which the basis vectors are driven to orthogonality to enhance the stability of factorization. Furthermore, a transformation matrix mapping the term space to the semantic subspace is obtained via the matrix factorization and truncation. Then a document can be represented as a point in the low dimensional semantic subspace, and TC is implemented in this subspace.

The rest of this paper is organized as follows: Section 2 reviews the related work briefly. Section 3 explains the proposed method in detail. Experimental results and analysis are shown in Section 4. Finally, we give our conclusions in Section 5.

2 Related Works

Some works have used NMF to reduce the dimensionality: Hoyer adds an explicitly constrain to control the sparseness of basis and coefficient matrices, but the sparseness does not necessarily contribute to the classification performance [8]. Guillamet et al. integrate the class information into NMF by assigning large weight values to the minority classes [9], but it is hard to suit the multi-label situation of TC (i.e. a document might belong to multiple categories in the same time). Liu, Yuan et al. utilize the NMF to reduce the dimensionality of the micro-array data but it limits the number of instances [10]. Silva and Ribeiro use NMF to extract the semantic features for TC [7], however it does not take into account the information of the category.

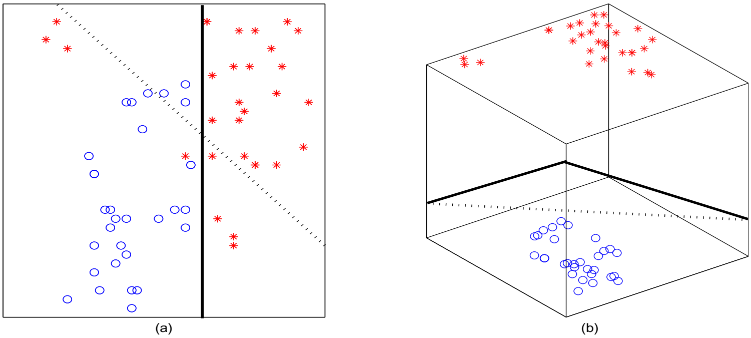

Fig. 1. Panel (a) shows there is no dimensionality extended, the solid line represents an ideally classification bound and the dash line represents the classification bound affected by three outlier points. Panel (b) shows another case of the dimensionality extended. The linear separability increases by adding some new dimension generated by category information.

3 Dimensionality Reduction Method

3.1 Category Information Fusion

Because the NMF is a linear reconstructed method, its goal is to obtain the optimal representative feature rather than the optimal classification feature. So the noise, missing data and outliers existing in the high dimensionality space also exist in the low-dimensional subspace generated by NMF. As panel (a) of Figure 1 shows, these points will affect the establishment of the correct classification boundary. Therefore, we wish to utilize category information as the extended dimensionality and fuse them into training data. So the linear separability of data increases in the new dimensionality extended space illustrated in panel (b) of Figure 1.

The category information fusion can be implemented with three stages: document weighting, category information coding, and fusion. The document weighting is presented as follows:

Given a document $d = (t_1, t_2, \cdots, t_m)$, where m is the number of dimensionality in the feature space. The $tfidf$ value [2] for each term is defined as:

$$tfidf(t_i, d) = tf(t_i, d) \times idf(t_i), \tag{1}$$

where $tf\ (t_i, d)$ denotes the number of times that t_i occurred in d, and $idf\ (t_i)$ is the inverse document frequency which is defined: $idf(t_i) = \log(n/df(t_i))$, where n is the number of documents in training set, and $df\ (t_i)$ denotes the number of documents in training set in which t_i occurs at least once. Then a document can be represented as a vector:

$$d = (w_1, w_2, \cdots, w_m)^T, \tag{2}$$

where w_i is evaluated as: $w_i = tfidf(t_i, d)/\sqrt{\sum_j^m tfidf(t_j, d)^2}$.

In order to represent the category information uniformly for multi-label or uni-label corpus, we extend the category coding scheme $1-of-K$ [11] to $k-of-K$, i.e. define the class vector corresponding to the document d as follows

$$c = (b_1, \cdots, b_i, \cdots, b_k)^T, \tag{3}$$

where k is the number of category in dataset, and b_i is equal to 1 or 0 depending on whether the related document belongs to the corresponding categories. For example, assuming there are four categories ($k = 4$), a document $d = (w_1, w_2, \cdots, w_m)^T$ belongs to the first and the forth categories, then the related class vector is $c = (1, 0, 0, 1)^T$. Then we fuse d and c into a new extended vector x:

$$x = \begin{bmatrix} d \\ \lambda \times c \end{bmatrix}, \tag{4}$$

where λ is a parameter used to control the tradeoff between train text d and its corresponding category information c. Finally, all train documents are fused into an extended matrix X, represented as

$$X = \begin{bmatrix} D \\ \lambda \times C \end{bmatrix}, \tag{5}$$

where D is the weighting matrix of all train documents, C is the class matrix related to D, and each column of X is an extended vector obtained by Eq. (4). Then X will be decomposed with the orthogonal NMF given below.

3.2 Orthogonal NMF

Given the extended matrix X, let

$$X_{m\times n} \approx W_{m\times r} \times H_{r\times n}. \tag{6}$$

Because of the uniqueness problems of scaling and permutation of NMF, we wish the basis matrix W tends to the orthogonal normalization, i.e. $W^TW - I = O$, where I is the unit matrix and O is the corresponding zero matrix. We consider it as a constrain term with parameter W, and add it into the loss function which can be constructed as follows

$$\begin{matrix} L(W, H) = ||X - WH||_F^2 + \alpha||W^TW - I||_F^2 \\ s.t. W, H \geq 0 \end{matrix}, \tag{7}$$

where α is used to balance the tradeoff between the approximation error and the orthogonal constraint. Same with [12], the multiplicative update algorithm is:

$$W_{i,j} \leftarrow W_{i,j} \frac{(WH^T + 2\alpha W)_{i,j}}{(WHH^T + 2\alpha WW^TW)_{i,j}}, \; H_{i,j} \leftarrow H_{i,j} \frac{(W^TX)_{i,j}}{(W^TWH)_{i,j}}. \tag{8}$$

3.3 Dimensionality Reduction and TC

Assuming the matrix X (in Eq. (5)) is decomposed (with Eq. (8)) as follows

$$X = \begin{bmatrix} D_{m\times n} \\ \lambda \times C_{k\times n} \end{bmatrix} \approx W_{(m+k)\times r} \times H_{r\times n}, \tag{9}$$

let

$$W_{(m+k)\times r} = \begin{bmatrix} S_{m\times r} \\ L_{k\times r} \end{bmatrix}, \tag{10}$$

where S is obtained by truncating W. Then

$$\begin{bmatrix} D_{m\times n} \\ \lambda \times C_{k\times n} \end{bmatrix} \approx \begin{bmatrix} S_{m\times r} \times H_{r\times n} \\ L_{k\times r} \times H_{r\times n} \end{bmatrix}. \tag{11}$$

So $D_{m\times n} \approx S_{m\times r} \times H_{r\times n}$, where S can be regarded as a matrix formed with the semantic vectors. Defining

$$P = (S_{m\times r})^{\dagger}, \tag{12}$$

then

$$\hat{D}_{r\times n} \approx P \times D_{m\times n}, \tag{13}$$

where P can be regarded as the transformation matrix which maps documents from the term space into a semantic subspace spanned by S, and $\hat{D}$ can be regarded as the projected vector set in the semantic subspace.

Using Eq.(13), the train and test data all can be mapped from the term space to a semantic subspace. In the new semantic subspace, dimensionality could be reduced aggressively, and some classical classification algorithms can be applied. For example using the support vector machine (SVM) algorithm, a pseudo code for TC is given by algorithm 1.

4 Experiments

4.1 Dataset

Two popular TC benchmarks are tested in our experiments: Reuters-21578 and 20-newsgroups. The Reuters-21578 dataset[1] is a standard multi-label TC benchmark and contains 135 categories. In our experiments, we use a subset of the data collection which includes the 10 most frequent categories among the 135 topics and we call it Reuters-top10. We divide it into the train and test set with the standard 'ModApte' version. The pre-processed including: removing the stop words; switching upper case to lower case; stemming[2]); removing the low frequency words (less than three).

The 20-Newsgroups dataset[3] contains approximately 20,000 articles evenly divided among 20 usenet newsgroups. We also remove the low frequency words (less than three) in the data set.

[1] Available at http://www.daviddlewis.com/resources/testcollections/

[2] Available at http://tartarus.org/˜martin/PorterStemmer/

[3] Available at http://people.csail.mit.edu/jrennie/20Newsgroups/

Algorithm 1. TC implementation

```
Input: the training and testing set and the setting of parameters
Output: the label of testing set
Learning stage:
for each document $d$ in training set do
    evaluating the weighting of $d$ with Eq. (2)
    evaluating the extended vector $x$ with Eq. (4)
end for
evaluating the extended matrix $X$ with Eq. (5)
factoring $X$ into the form of Eq.(9) with Eq. (8)
evaluating the transformation matrix $P$ using Eq.(10) and Eq. (12)
evaluating the projection of training set in the semantic subspace with Eq.(13)
Learning SVM mode in the semantic subspace
Classification stage:
for each document $d'$ in testing set do
    evaluating the weighting of $d'$ with Eq. (2)
    evaluating $\hat{d}'$ with Eq. (13)
    classifying the document $\hat{d}'$ with SVM classifier in the semantic subspace
end for
```

4.2 Evaluation Measures

In the tests, we adapt the macroaveraged F_1 [1] as the performance measure which is defined as

$$\text{macroaveraged}\, F_1 = (\sum_{i}^{k} F_{1i})/k, \tag{14}$$

where k is the number of categories, F_{1i} denotes the F_1 value of the ith category.

4.3 Results and Analysis

To verify the performance of the proposed approach (denoting it as CONMF for convention), we compare it with the information grain (IG) (one of the most successful feature selection methods [13]), the ordinary NMF method [7] (denotes it as NMF), and LSI [14]. The classifier is implemented with SVMlight[4] and its default parameters are adapted. We set $\alpha = 0.5$ (Eq. (7)) and $\lambda = 1$ (Eq. (5)). The dimensionality reduction level is from about 1% to 0.1%. For each test dimensionality, we repeat the experimentation ten times and take their means as the result.

The results are shown in Figure 2. From this figures, we can see that the performance of IG is very awful when the dimensionality is reduced aggressively, it is because that case might induce all zero value of feature vectors using feature selection method, which makes the classification failure.

[4] Available at http://svmlight.joachims.org/

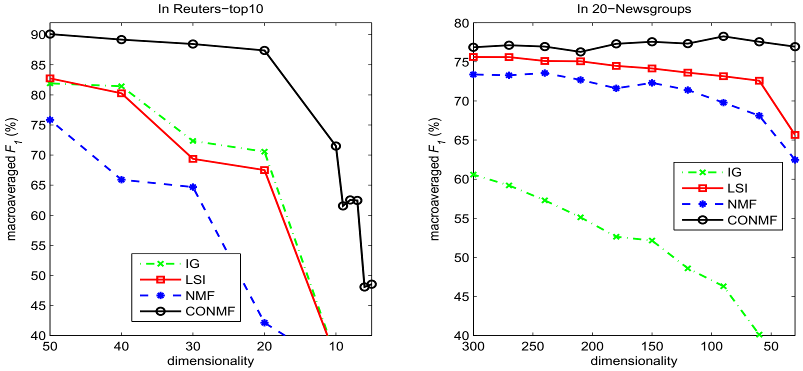

Fig. 2. Performances in Reuters-top10 and 20-newsgroups

According to the ordinary NMF method, although it has a more reasonable physical interpretation than LSI, it does not presents a good classification performance because its objective function is non-convex and it is difficult to obtain the optimal solution. Moreover, it is sensitive to the noise, data missing and outlier which will affect the establishment of the correct classification boundary.

Despite the results that the performance of IG or LSI or ordinary NMF descends drastically in the very low dimensional situation, our CONMF method obtains relatively stable performance in that case.

These figures also reveal a phenomenon that the relationship between the dimensionality of semantic subspace (or the number of semantic concepts) and the TC performance is not a linear function. In other words, the performance does not necessarily increase when the dimensionality of the semantic subspace increases and vice versa, which imply that there might exist an optimized number of semantic concepts in a specific text corpus.

5 Conclusions

This paper proposes a novel approach to reduce dimensionality aggressively. By utilizing category information as the extended dimensions, the impact of the noise, data missing and outlier could be decreased. Furthermore, with basis orthogonality non-negative matrix factorization and truncation, the data in the high dimensional term space could be mapped into a low dimensional semantic subspace. Experimental results show that the proposed approach remains a good classification performance in the very low dimensional case.

The proposed method is simply and effective, and its factorization form as well as its non-negative constrain could provide a more reasonable physical interpretation than LSI. Meanwhile, it reflects the concept about “parts form the whole” in human mind. Furthermore, the form of word-semantic-category is consistent with the cognitive process when people read articles.

For the future researches, we would like to study the dimensionality problem of the semantic subspace to enhance the stability of factorization, and try to incorporate cognitive information into the non-negative matrix factorization to improve the classification performance.

Acknowledgment. This work was supported by the Natural Science Foundation of Zhejiang Province (Grant No. Y6110147) and was supported in part by the Natural Science Foundation of Zhejiang Province (Grant No. Y1110450).

References

1. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
2. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
3. Zheng, W., Qian, Y.: Aggressive dimensionality reduction with reinforcement local feature selection for text categorization. In: Wang, F.L., Deng, H., Gao, Y., Lei, J. (eds.) AICI 2010. LNCS, vol. 6319, pp. 365–372. Springer, Heidelberg (2010)
4. Landauer, T., Foltz, P., Laham, D.: An introduction to latent semantic analysis. Discourse Processes 25(2), 259–284 (1998)
5. Zhou, S., Li, K., Liu, Y.: Text categorization based on topic model. In: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS (LNAI), vol. 5009, pp. 572–579. Springer, Heidelberg (2008)
6. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
7. Silva, C., Ribeiro, B.: Knowledge extraction with non-negative matrix factorization for text classification. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 300–308. Springer, Heidelberg (2009)
8. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research 5, 1457–1469 (2004)
9. Guillamet, D., Vitri, J., Schiele, B.: Introducing a weighted non-negative matrix factorization for image classification. Pattern Recognition Letters 24(14), 2447–2454 (2003)
10. Liu, W., Yuan, K., Ye, D.: Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. Journal of Biomedical Informatics 41(4), 602–606 (2008)
11. Bishop, C.M.: SpringerLink: Pattern recognition and machine learning, vol. 4. Springer, New York (2006)
12. Zheng, W., Zhang, H., Qian, Y.: Fast text categorization based on collaborative work in the semantic and class spaces. In: To Appear in the International Conference on Machine Learning and Cybernetics (2011)
13. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 412–420, Citeseer. Morgan Kaufmann Publishers Inc, San Francisco (1997)
14. Zhang, W., Yoshida, T., Tang, X.J.: A comparative study of tf*idf, lsi and multiwords for text classification. Expert Systems with Applications 38(3), 2758–2765 (2011)