Improved Nonnegative Matrix Factorization Based Feature Selection for High Dimensional Data Analysis

Lincheng Jiang,Wentang Tan, Zhenwen Wang, Fengjing Yin,Bin Ge,Wendong Xiao
Science and Technology on Information Systems Engineering Laboratory
National University of Denfense Technology
Chang Sha, China
e-mail: linchengjiang08@163.com

***Abstract*—Feature selection has become the focus of research areas of applications with high dimensional data. Nonnegative matrix factorization (NMF) is a good method for dimensionality reduction but it can't select the optimal feature subset for it's a feature extraction method. In this paper, a two-step strategy method based on improved NMF is proposed.The first step is to get the basis of each catagory in the dataset by NMF. Added constrains can guarantee these basises are sparse and mostly distinguish from each other which can contribute to classfication. An auxiliary function is used to prove the algorithm convergent.The classic ReliefF algorithm is used to weight each feature by all the basis vectors and choose the optimal feature subset in the second step.The experimental results revealed that the proposed method can select a representive and relevant feature subset which is effective in improving the performance of the classifier.**

Keywords-feature selection; nonnegative matrix factorization; reliefF algorithm

I. INTRODUCTION

Rapid technological developments in Computer Science have resulted in increasing quantities of data, making many of the classical data analysis tools unavailable. An effective method to solve this problem is dimensionality reduction. Dimensionality reduction techniques include feature extraction and feature selection techniques. Feature extraction refers to the mapping of the original high-dimensional data onto a lower-dimensional space while feature selection is a process that chooses an optimal subset of features according to an objective function.

Existing research results show that feature extraction can be effective in improving the performance of the classifier[1]. But transform-based feature extraction makes those features which are irrelevant and redundant also play a role in the process of dimensionality reduction, thus inevitably affecting the performance of the classifier.Feature extraction loses the original meaning of the physical characteristics in the lower-dimensional feature space, so its interpretation is poor. At the same time, in some practical applications, we need to know exactly which features play a key role in the forecast.This is a typical feature selection problem. Feature selection for high-dimensional data is considered one of the current challenges in machine learning[2], especially when the dataset is huge.

NMF[3] is a good method for dimensionality reduction which has been used in many fields,such as text mining, face recognition, microarray data analysis and so on.It is always considered to be an unsupervised learning algorithms for feature extraction.This paper is to use it for feature selection not for feature extraction.

In this paper,we propose a new feature selection algorithm adopting a two-step strategy for high-dimensional data .The first step is to use improved NMF to compress the great number of data samples to several groups of vectors which can be regarded as the basis of each category in the dataset. This is the core of our algorithm. The second step is to adopt a classical feature selection tool to choose an optimal subset of feature in the basis space.In his paper we use the reliefF algorithm which is considered to be one of the most successful algorithms for assessing the quality of features.

The remainder of this paper is organized as follows. In Section 2, some related works are given. The whole algorithm of this new feature selection method is proposed in Section 3. In Section 4, the experiment and analysis of the results are given. And in the last section, we conclude our works.

II. RELATED WORKS

A. Feature Selection Method

Research on feature selection has been very attractive in the past decade. Existing algorithms are traditionally categorized as wrapper or filter methods, with respect to the criteria used to search for relevant features[4]. Wrapper methods, motivated by Kohavi and John[5] use the performance of a predetermined learning algorithm for searching an optimal feature subset,while filter methods evaluate feature subsets by their information content, typically interclass distance (e.g.Fisherscore) or statistical measures (e.g.p-value of t-test), instead of optimizing the performance of any specific learning algorithm directly.As a result, filter methods are extremely attractive for microarray analysis and text-categorization domains because of their computational efficiency and simplicity. In this paper,we uses reliefF[6] algorithm to select the optimal subset of features. The pseudo-code for reliefF algorithm is given below.

(1) Set all weights W[A] = 0.0
(2) for i = 1 to m do begin

(3) randomly select sample R_i
(4) find k nearest hit H_j
(5) for each class C $\neq$ class(R_i) do
(6) find k nearest miss M_j(C) from class C
(7) for A = 1 to all feature do

(8) $$W(A)=W(A)-\sum_{j=1}^{k} diff(A,R_i,H)/(m,k)+\sum_{C\neq class(R_i)}^{k}\left(\frac{P(C)}{1-P(class(R_i))}\sum_{j=1}^{k} diff(A,R_i,M_j(C))\right)/(m,k)$$

(9) end
(10)end
(11)end

B. Nonegative Matrix Factorization

NMF was first proposed by Lee and Seung[3]. NMF seeks to decompose a nonnegative n × m matrix V into a nonnegative n × r matrix W and a nonnegative r × m matrix H. where (m+n)×r << m×n. It can be described as:

$$Min\left\|V-WH\right\|^2, s.t. W>0, H>0 \tag{1}$$

W can be regarded as containing a basis that is optimized for the linear approximation of the data in V. Since relatively few basis vectors are used to represent a lot of data vectors, good approximation can only be achieved if the basis vectors find out latent structure in the data. The following iterative learning rules are used to find the linear decomposition:

$$\begin{aligned} H &\leftarrow H \bullet \frac{W^{\mathrm{T}}V}{W^{\mathrm{T}}WH} \\ W &\leftarrow W \bullet \frac{VH^{T}}{WHH^{T}} \end{aligned} \tag{2}$$

With the wide application of NMF,the researchers suggested some improved algorithms, such as local Nonnegative Matrix Factorization (LNMF), Fisher Nonnegative Matrix Factorization (FNMF), Sparse Nonnegative Matrix Factorization (SNMF), and Weighted Nonnegative Matrix Factorization (WNMF)[7].

III. IMPROVED NMF

Suppose there is a dataset which has two categories and d features in the original high-dimensional space.One category has n_1 observations represented by $X_1 \in R_+^{d,n_1}$,the other has n_2 observations represented by $X_2 \in R_+^{d,n_2}$ .The algorithm make nonnegative matrix factorization for both X_1 and X_2 .Let $X_1 = A_1S_1$ and $X_2 = A_2S_2$ .Here A_1 , S_1 , A_2 , $S_2 \in R_+$ .In the decomposition process,some other important constraints are added for a better effect.The last combined cost function which this algorithm is aimed to minimize is :

$$\begin{aligned} &Min\ (\lambda\left\|X_1-A_1S_1\right\|^2+\lambda\left\|X_2-A_2S_2\right\|^2 \\ &+\gamma\left\|A_1\right\|^2+\gamma\left\|A_2\right\|^2)\ s.t.\lambda,\gamma>0 \end{aligned} \tag{3}$$

A. Significance

This method is not simply making nonnegative matrix factorization for the matrix of each category.At first,sparse constraints are added because the data tend to be sparse in a practical problem and a sparse matrix can save storage space. Moreover,with sparse constraints, the decomposition can always result in parts-based representive basis vectors of each category.Different category has different representations,thus the proposed algorithm can make the basises of the two categories mostly distinguish from each other,which can contribute to classification. Adjusting the parameters λ,γ in (3) is useful to find the most suitable decomposition. When executing the algorithm, these parameters are input from the outside.

B. Update Rules

We have found that the following update rules are a good compromise between speed and ease of implementation for minimizing the cost function.

Theorem 1 The cost function (3) is nonincreasing under the update rules.

$$\begin{aligned} S_1 &\leftarrow S_1 \bullet \frac{A_1^{\mathrm{T}}X_1}{A_1^{\mathrm{T}}A_1S_1} \\ S_2 &\leftarrow S_2 \bullet \frac{A_2^{\mathrm{T}}X_2}{A_2^{\mathrm{T}}A_2S_2} \\ A_1 &\leftarrow A_1 \bullet \frac{\lambda X_1S_1^{T}}{\lambda A_1S_1S_1^{T}+\gamma A_1} \\ A_2 &\leftarrow A_2 \bullet \frac{\lambda X_2S_2^{T}}{\lambda A_2S_2S_2^{T}+\gamma A_2} \end{aligned} \tag{4}$$

C. Proof of Convergence

An auxiliary function method is used to prove the update rules (4) convergent.Similar method was used in [3].

Definition 1. $G(h,h^{'})$ is an auxiliary function for $F(h)$ if the conditions

$$G(h,h^{'})\geq F(h),\ \ G(h,h)=F(h) \tag{5}$$

are satisfied.

The auxiliary function is a useful concept because of the following lemma.

Lemma 1. If G is an auxiliary function, then F is nonincreasing under the update:

$$h^{t+1}=\arg\min_h G(h,h^{t}) \tag{6}$$

Proof: $F(h^{t+1})\leq G(h^{t+1},h^{t})\leq G(h^{t},h^{t})=F(h^{t})$

Note that $F(h^{t+1})=F(h^{t})$ only if h^t is a local minimum of $G(h,h^t)$ .If the derivatives of F exist and are continuous in a small neighborhood of h^t ,this also implies that the derivatives $\nabla F(h^t)=0$ .By iterating the update in (6) we

obtain a sequence of estimates that converge to a local minimum $h_{\min} = \arg\min_h F(h)$ of the objective function:

$$F(h_{\min}) \le \cdots F(h^{t+1}) \le F(h^t) \cdots \le F(h^2) \le F(h^1) \le F(h^0) \quad (7)$$

Lemma 2. If $K(h^t)$ is the diagonal matrix

$$K_{ij}(h^t) = \delta_{ij}(\lambda h^t S_1 S_1^T + \gamma h^t)_i / h_i^t \quad (8)$$

where h represents row vectors in A_1 .Then

$$\begin{aligned} &G(h,h^t) = F(h^t) + (h-h^t)^T \nabla F(h^t) \\ &+\frac{1}{2}(h-h^t)^T K(h^t)(h-h^t) \end{aligned} \quad (9)$$

is an auxiliary function for

$$\begin{aligned} F(h) &= \frac{\lambda}{2}\sum_i \left((X_1)_i - \sum_k h_k (S_1)_{ki} \right)^2 \\ &+\frac{\lambda}{2}\sum_i \left((X_2)_i - \sum_k (A_2)_k (S_2)_{ki} \right)^2 \\ &+\frac{\gamma}{2}\sum_i h_i^2 + \frac{\gamma}{2}\sum_i (A_2)_i^2 \end{aligned} \quad (10)$$

Proof: Since $G(h,h) = F(h)$ is obvious, we need only show that $G(h,h') \ge F(h)$. To do this, compare

$$\begin{aligned} &F(h) = F(h^t) + (h-h^t)^T \nabla F(h^t) \\ &+\frac{1}{2}(h-h^t)^T (\lambda S_1 S_1^T + \gamma \times E)(h-h^t) \end{aligned} \quad (11)$$

with (9) to find that $G(h,h') \ge F(h)$ is equivalent to

$$0 \le (h-h^t)^T [K(h^t) - (\lambda S_1 S_1^T + \gamma \times E)](h-h^t) \quad (12)$$

To prove positive semidefiniteness, consider the matrix

$$M_{ij}(h^t) = h_i^t (K(h^t) - (\lambda S_1 S_1^T + \gamma \times E)) h_j^t \quad (13)$$

M is semipositive if and only if

$$\begin{aligned} v^T M v &= \sum_{ij} v_i M_{ij} v_j \\ &= \sum_{ij} [h_i^t (\lambda S_1 S_1^T + \gamma \times E) h_j^t v_i^2 \\ &\quad - v_i h_i^t (\lambda S_1 S_1^T + \gamma \times E) h_j^t v_j] \\ &= \sum_{ij} (\lambda S_1 S_1^T + \gamma \times E) h_i^t h_j^t [\frac{1}{2} v_i^2 + \frac{1}{2} v_j^2 - v_i v_j] \\ &= \frac{1}{2}\sum_{ij} (\lambda S_1 S_1^T + \gamma \times E) h_i^t h_j^t (v_i - v_j)^2 \ge 0 \end{aligned} \quad (14)$$

Proof of Theorem 1. Replacing $G(h,h^t)$ in (6) by (9) results in the update rule:

$$h^{t+1} = h^t - K(h^t)^{-1} \nabla F(h^t) \quad (15)$$

Since (9) is an auxiliary function, nonincreasing under this update rule according to Lemma 1. Writing the components of this equation explicitly, we obtain

$$h_a^{t+1} = h_a^t \bullet \frac{(\lambda X_1 S_1^T)_a}{(\lambda A_1 S_1 S_1^T + \gamma A_1)_a} \quad (16)$$

Other update rules can similarly be proved.

D. Feature Selection for Multiclass Problems

This method is designed to slove binary class problem,but it can be easily to extended to multiclass problem. A multiclass problem is first decomposed into several binary ones, and then feature selection is performed for each binary problem. Suppose there are k class and every class is represented by X_i(2<=i<=k).The following form is the cost function for multiclass:

$$Min\ (\lambda \sum_{i=1}^{k} \left\| X_i - A_i S_i \right\|^2 + \gamma \sum_{i=1}^{k} \left\| A_i \right\|^2) \quad (17)$$

The update rules are also similar to binary class problem,we don't list them here.

IV. EXPRIMENT

In the experiment, a two-step strategy was adopted.The first step is to use the improved NMF method to find out compressed basis of every category.Each basis is constructed by several column vectors which can be regarded as weight of each feature in the basis space. The parameters were set in this step as λ=1,γ=1.The second step is to use reliefF algorithm to weight each feature by all the basis vectors,thus we can get the feature weight vector.Sort feature weights in descending order and set a threshold.The features will be elected into the optimal feature subset if their weight is bigger than the threshold.

The experiment data are from the original TDT2 corpus(Nist Topic Detection and Tracking corpus)which has been used in CaiDeng's experiment[8]. The TDT2 corpus consists of data collected from 6 sources, including 2 newswires (APW, NYT), 2 radio programs (VOA, PRI) and 2 television programs (CNN, ABC). It consists of 11201 on-topic documents which are classified into 96 semantic categories.Here we just selected the top two categories and the data were given in matrix form.Each category consists of 1000 column vectors for training and 828 column vectors for testing.The data have 36771 features originally.Some features were removed because of a value of 0 in the observations, thus leaving 10955 features behind.

After excuting the first step,the 2000 column vectors were compressed into 40 column vectors.In this text mining problem, the 40 column vectors can be regarded as 40 topics.

In order to get a better effectiveness of the algorithm, we have introduced the rate of feature reduction E and the accuracy rate F to evaluate the algorithm.E and F are defined as [9]:

$$E = (1 - \frac{N_r}{N}) \times 100\% \tag{18}$$

$$F = \frac{n_c}{n} \times 100\% \tag{19}$$

Where N is the original feature dimension, Nr is the dimension after feature selection,n_c is the number of texts which are correctly sorted, and n is the total number of texts.

After getting the feature subset,we performed a number of experiments with Support Vector Machine(SVM) as classifier.Here we used the SVM tool libsvm[10].The experiment results are shown in TABLE 1.

TABLE I. THE RESULTS OF THE EXPERIMENT

Serial No.	The Results of the Experiment			
	The number of text correct classification	The number of features after reduction	Reduction rate E	Accuracy rate F
1	1802	1000	90.87%	97.07%
2	1796	500	95.44%	96.74%
3	1743	200	98.17%	93.90%
4	1717	100	99.09%	92.52%
5	1726	50	99.54%	93.00%

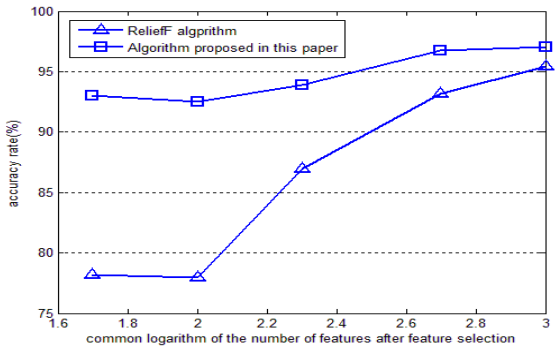

Figure 1. Results compared with ReliefF algorithm

In order to check the effectiveness of the proposed algorithm, more experiments compared with ReliefF algorithm were performed.The results were given in Fig.1. X-axis represents the common logarithm of the number of features after feature selection. Y-axis represents the accuracy rate of classfication.From the figure,we can see that the proposed algorithm works far better than ReliefF algorithm.

The experiments show that our method can significantly reduce redundant features and guarantee a high classification accuracy.It is an excellent method for feature selection.

V. CONCLUTIONS

In this paper,we have proposed a method for feature selection for high dimensional data based on improved nonnegative matrix factorization.NMF is always considered to be an unsupervised learning algorithms for feature extraction but we creatively use it as a supervised learning method for feature selection. Combining improved NMF with the classical feature selection algorithm,we get an excellent result beyond expectation .In the future, we'll discuss the influence of the parameters in our algorithm on the result.

ACKNOWLEDGMENT

This work was partially supported by National Unversity of Defence Technology.

REFERENCES

[1] Saidi Rabie,Aridhi Sabeur,Maddouri Mondher and Nguifo Engelbert Mephu, "Feature extraction in protein sequences classification: A new stability measure,"ACM Conference on Bioinformatics, Computational Biology and Biomedicine(BCB 2012), pp.683-689, 2012.

[2] J. Lafferty and L. Wasserman, "Challenges in statistical machine learning," Statistica Sinica, vol. 16, pp. 307–322, 2006.

[3] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, pp. 788–791, 1999..

[4] Sun Yijun,Todorovic Sinisa and Goodison Steve, "Local Learning Based Feature Selection for High-Dimensional Data Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI 2010), vol.32, pp. 1610-1626, 2010 .

[5] R. Kohavi and G. John, "Wrappers for Feature Subset Selection,"Artificial Intelligence, vol. 97, pp. 273-324, Dec. 1997.

[6] Igor Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF,"Peoceedings of European Conference on Machine Learning.Catania,Springer-Verlag, pp.171-182, 1994.

[7] Wu Min, Li Jia,Liao Dingan and Lin Qing, "Improved method based on NMF for face recognition," International Conference on Multimedia Technology(ICMT 2011), pp.559-562, 2011.

[8] Cai Deng,Wang Xuanhui,He Xiaofei, "Probabilistic Dyadic Data Analysis with Local and Global Consistency,"Proceedings of the 26th Annual International Conference on Machine Learning(ICML 09), vol. 382, 2009.

[9] Lei Shang, "A Feature Selection Method Based on Information Gain and Genetic Algorithm," International Conference on Computer Science and Electronics Engineering(ICCSEE 2012), vol.2,pp.355-358, 2012.

[10] C. Chang and C. Lin, LIBSVM: a library for support vector machines, 2001.Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm