Neural Networks 52 (2014) 1–17

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Pairwise constrained concept factorization for data representation

Yangcheng He*, Hongtao Lu, Lei Huang, Saining Xie

MOE-Microsoft Laboratory for Intelligent Computing and Intelligent Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, PR China

ARTICLE INFO

Article history:
Received 14 November 2012
Received in revised form 27 September 2013
Accepted 18 December 2013

Keywords:
Concept factorization
Clustering
Pairwise constraints
Data representation

ABSTRACT

Concept factorization (CF) is a variant of non-negative matrix factorization (NMF). In CF, each concept is represented by a linear combination of data points, and each data point is represented by a linear combination of concepts. More specifically, each concept is represented by more than one data point with different weights, and each data point carries various weights called membership to represent their degrees belonging to that concept. However, CF is actually an unsupervised method without making use of prior information of the data. In this paper, we propose a novel semi-supervised concept factorization method, called Pairwise Constrained Concept Factorization (PCCF), which incorporates pairwise constraints into the CF framework. We expect that data points which have pairwise must-link constraints should have the same class label as much as possible, while data points with pairwise cannot-link constraints will have different class labels as much as possible. Due to the incorporation of the pairwise constraints, the learning quality of the CF has been significantly enhanced. Experimental results show the effectiveness of our proposed novel method in comparison to the state-of-the-art algorithms on several real world applications.

1. Introduction

In many data analysis tasks of data mining, machine learning and pattern recognition, we often need to deal with the data in high-dimensional space. Traditional methods which perform well in low-dimensional space may break down partly in high-dimensional space (Fodor, 2002; Hua & He, 2011). Therefore, seeking a suitable low-dimensional representation for the high-dimensional data is becoming more and more important, and dimensionality reduction has been used as a principled way to do that. Many dimensionality reduction techniques can be expressed as matrix factorization problems with different objective functions. Matrix factorization aims to find two or more matrices whose product provides a good approximation to the original matrix. In matrix factorization the dimensions of the factorized matrices are generally much smaller than those of the original one. One hopes that important characteristics of the data points in the original space can be reserved in the low dimensional space. After matrix factorization we can use the low dimensional matrices to deal with classification or clustering tasks.

Among existing matrix decomposition methods, Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999) can be used to obtain new representations of the data points with non-negative constraints. That is, it requires that all elements of the decomposed matrix factors are non-negative. These non-negative constraints lead to parts-based representations of the objects because they only allow additive, not subtractive, combinations of the original data points. NMF is a helpful dimensionality reduction method for face recognition (Guillamet & Vitria, 2002), document clustering (Xu, Liu, & Gong, 2003), image processing (Kim & Park, 2008) and computer vision (Shashua & Hazan, 2005). However, one drawback of NMF is that it can only be used in the original feature space of the data, and thus cannot make use of the power of kernelization (Hua & He, 2011).

To overcome the drawback of NMF and inherit all its strengths, Xu and Gong proposed the Concept Factorization (CF) algorithm (Xu & Gong, 2004). In CF, each concept is represented by a linear combination of the data points, and each data point is represented by a linear combination of the concepts. With this model, the data clustering task is accomplished by computing the two sets of linear coefficients, and this linear coefficients computation is carried out by finding the non-negative solution that minimizes the reconstruction error of the data points. The major advantage of CF over NMF is that the powerful idea of the kernel method can be applied to CF; with the kernel function, CF can map the linearly non-separable data in the original space into linearly separable data in the transformed high-dimensional space.

However, CF is an unsupervised learning method. That is, CF does not use any prior knowledge of the data to guide the learning process; nevertheless, there is a certain amount of prior knowledge in the real world applications. Using prior knowledge of the

* Corresponding author. Tel.: +86 18817519338.
E-mail address: h331076268@126.com (Y. He).

0893-6080/$ – see front matter Crown Copyright © 2013 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.12.007

problem in hand to improve the performance of the algorithms has become one of the hot areas of machine learning. Many machine learning researchers have pointed out that when a small amount of labeled data is used in conjunction with unlabeled data, it can produce encouraging improvement in learning performance (Chapelle, Schölkopf, & Zien, 2006; Grira, Crucianu, & Boujemaa, 2005; He, Zheng, Hu, & Kong, 2011; Yang, Jin, & Sukthankar, 2008; Zhang & Yeung, 2008). Even though, it is infeasible to label all the sample points in the data set, because the cost will be high expensive, whereas obtaining a small amount of labeled data is relatively inexpensive. Under these circumstances, semi-supervised learning algorithms can play a greater performance. CF can be extended to semi-supervised manner to enhance the learning quality.

Recently, the manifold learning method (Zhang, Wang, & Zha, 2012; Zhang, Zha, & Zhang, 2008) has also been incorporated into CF. Cai, He, and Han (2011) had proposed a Locally Consistent Concept Factorization (LCCF) algorithm which encoded the geometrical information of the data space by constructing a nearest neighbor graph to model the local manifold structure. When the label information is provided, it can be directly encoded into the graph structure. For example, if two points have the same label, the algorithm assigns a large weight to the edge connecting them. Otherwise, if two points possess different labels, the corresponding weight of the edge is encoded to be 0. In doing so, LCCF turns into a semi-supervised learning algorithm. The main drawback of this algorithm is that it only focuses on the local structure in the data, which may often lead to over-fitting. Besides, it is hard to fix the nearest neighbor number of one point when constructing the nearest neighbor graph of the data and select the weights between the neighbor points.

Liu, Wu, Li, Cai, and Huang (2012) proposed a Constrained Non-negative Matrix Factorization (CNMF) approach which used the label information as additional hard constraints. The central idea of their algorithm is that the data points with the same class label must be strictly mapped to share the same representation in the new parts-based representations space. The method forces the new parts-based representations to have the consistent label information with the original data. Obviously, this requirement is too strict so that it will weaken the representational ability of the new parts-based representations space for other unlabeled data, because it might assign unlabeled data with totally wrong representations due to its strictly hard constraints. Kulis, Basu, Dhillon, and Mooney (2005) had proposed a Semi-supervised Kernel Kmeans (SSKK) method which constructed a similarity matrix to clustering; the penalty weights were incorporated into the similarity matrix according to pairwise constraints. A major disadvantage is that the similarity matrix adjusting to the penalty weights is defined in advance; it cannot be adjusted automatically in the clustering process to enhance or diminish the similarity among data points.

In this paper, we propose a novel semi-supervised concept factorization method, called Pairwise Constrained Concept Factorization (PCCF), which uses the pairwise must-link and cannot-link constraints as additional soft constraints which are generated among the labeled data points. Pairwise constraints have been used in semi-supervised learning (Li, Liu, & Tang, 2008); however, to our knowledge, they have not been incorporated into the CF framework. The central idea of our approach is that the data points having pairwise must-link constraints should have the same class label as much as possible. On the contrary, the data points with pairwise cannot-link constraints should have different class labels as much as possible. To achieve this, we carefully design a new concept factorization objective function incorporating the pairwise constraints information into it. We also develop an optimization scheme for the objective function to derive the iterative updating rules of the two matrices **W** and **V**, the computational complexity of our algorithm is qualitatively analyzed, and the convergence proof of our algorithm is provided. Our experimental evaluations show that the proposed approach achieves good performance and outperforms other state-of-the-art methods.

2. A brief review of NMF and CF

Given a set of sample points, we form a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$; $\mathbf{x}_j$ $(j = 1, \ldots, n)$ is an m-dimensional non-negative vector, denoting the jth sample point. NMF aims to factorize X into the product of two non-negative matrices **U** and **V**, such that the product of **U** and **V** is a good approximation to the original matrix.

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^T. \tag{1}$$

In order to obtain the two non-negative matrices **U** and **V**, we can quantify the quality of the approximation by using a cost function with some distance metric. For example, if the Euclidean distance between two matrices is used, the problem turns to minimize the following objective function.

$$J = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \sum_{c=1}^{k} u_{ic}v_{jc}\right)^2 \tag{2}$$

where $\|\cdot\|$ is the matrix Frobenius norm denoting the square root of the squared sum of all the entries in the matrix. The dimensions of the factorized matrices **U** and **V** are $m \times k$ and $n \times k$, respectively. Usually, k is chosen such that $k \ll \min\{m, n\}$. Each column vector $\mathbf{u}_c$ of matrix **U** can be regarded as a basis of the new representations space (Das Gupta & Xiao, 2011; Xu & Gong, 2004), while each row vector of matrix **V** contains the coefficients of a linear combination of the column vectors of **U**; the linear combination of the columns of the matrix **U** with the jth row vector of the matrix **V** is used to approximate the jth column vector $\mathbf{x}_j$ of the matrix **X**. In fact, the jth row vector of the matrix **V** is the low-dimensional representation of the original high-dimensional data $\mathbf{x}_j$. The new representations space only contains k bases which is much less than the dimension of the original space. As $k \ll \min\{m, n\}$, the high-dimensional vector is represented by a low-dimensional vector in the low-dimensional coordinate space. One expect that through this process, a lot of redundant information can be removed from the original data, and the underlying structure in the original data can be captured. In contrast to other dimension reduction methods such as PCA and LDA, the factorized matrices are not allowed to contain negative entries and only permitted the non-negative combination of the basis vectors in the new representations space, this is why NMF can be treated as parts-based representations method.

NMF can only be used in the original feature space of the data; when the data are highly non-linear distributed, it cannot make use of the power of kernelization. To overcome this drawback of NMF, Xu and Gong (2004) proposed Concept Factorization (CF) which is an extension of NMF. In CF, each basis $\mathbf{u}_c$ which is the center of concept c is modeled as a linear combination of the data point vectors $\mathbf{x}_j$, and each data point is modeled as a linear combination of the basis vectors, that is

$$\mathbf{u}_c = \sum_{j=1}^{n} w_{jc}\mathbf{x}_j \tag{3}$$

$$\mathbf{x}_j = \sum_{c=1}^{k} v_{jc}\mathbf{u}_c \tag{4}$$

where $j = 1, 2, \ldots, n$, $\mathbf{u}_c$ is the basis vector, also called the center of concept c, $c = 1, 2, \ldots, k$. w_{jc} is a non-negative weight to