Knowledge-Based Systems 112 (2016) 152–165

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

# Subspace learning-based graph regularized feature selection

Ronghua Shang[a,*], Wenbing Wang[a], Rustam Stolkin[b], Licheng Jiao[a]

[a] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an, Shaanxi Province 710071, China
[b] Extreme Robotics Lab, University of Birmingham UK

ARTICLE INFO

*Article history:*
Received 27 May 2016
Revised 23 August 2016
Accepted 10 September 2016
Available online 12 September 2016

*Keywords:*
Graph regularized
Subspace learning
Feature manifold
Sparse constraint
Feature selection

ABSTRACT

In recent years, a variety of feature selection algorithms based on subspace learning have been proposed. However, such methods typically do not exploit information about the underlying geometry of the data. To overcome this shortcoming, we propose a novel algorithm called subspace learning-based graph regularized feature selection (SGFS). SGFS builds on the feature selection framework of subspace learning, but extends it by incorporating the idea of graph regularization, in which a feature map is constructed on the feature space in order to preserve geometric structure information on the feature manifold. Additionally, the $L_{2,1}$-*norm* is used to constrain the feature selection matrix to ensure the sparsity of the feature array and avoid trivial solutions. The resulting method can provide more accurate discrimination information for feature selection. We evaluate SGFS by comparing it against five other state-of-the-art algorithms from the literature, on twelve publicly available benchmark data sets. Empirical results suggest that SGFS is more effective than the other five feature selection algorithms.



## 1. Introduction

The increasingly rapid growth of information technology has seen a corresponding growth in the number of dimensions of gathered data. In many high dimensional data sets, only a small subset of the available features are useful, with most features being redundant, and some features even corresponding to information-less noise [1–5]. To facilitate the subsequent processing, it is often necessary to reduce the dimension of such high dimensional data. Processing of high dimensional data has become a challenge for researchers in many different fields [6–8], including data mining, machine learning, pattern recognition and others. Dimensionality reduction methods can broadly be categorized into methods for feature selection and feature extraction [9–13]. Feature selection methods select representative features from the original set of features based on a variety of evaluation methods. In contrast, feature extraction methods map high dimensional data into a low dimensional space through a transformation matrix. Feature selection methods select a subset of the original "raw" feature data, and so retain the physical or real-world meaning of the original data. This means that the performance of the resulting classifiers can often be readily explained in terms of intuitively meaningful trends in the underlying data. In contrast, it may be difficult to explain the behaviour of feature extraction methods in terms of the relationship between the new feature and the sample class [1]. In this paper, we propose a new feature selection algorithm.

Feature selection methods can broadly be divided into: supervised [2,14], semi supervised [15], unsupervised [16–18]. In supervised feature selection problems, the data discrimination information and also the correlation between features and the class of each data sample is available during training. However, in order to obtain large amounts of such class information, need for training such methods, a large amount of human resources are typically required, e.g. for hand annotation of data [18]. Semi-supervised feature selection requires only a smaller portion of the training data to be annotated with class label information to improve the accuracy of feature selection [15]. Unsupervised feature selection, without any class label information, only relies on the inherent information of the input data to determine the importance of features [16]. In many practical applications, the true class label information is unknown, which makes unsupervised feature selection methods more widely applicable to real problems, but also engenders greater challenges for researchers. According to various possible search strategies, unsupervised feature selection can be divided into filter, wrapper and embedded [19–26] methods.

In recent years, powerful new algorithms have been proposed which exploit the advantages of matrix decomposition techniques. Well known examples of such methods include nonnegative matrix factorization (NMF) [27,28], principal component analysis (PCA) [29,30] and singular value decomposition (SVD) [30,31]. However, all of these are examples of feature extraction methods. A smaller body of literature has explored how the idea of matrix decomposition can also be applied to feature selection. Wang et al.

* Corresponding author.
*E-mail address:* rhshang@mail.xidian.edu.cn (R. Shang).

http://dx.doi.org/10.1016/j.knosys.2016.09.006
0950-7051/

[32] proposed subspace learning for unsupervised feature selection via matrix factorization (MFFS). In [33], Wang et al. proposed unsupervised feature selection via maximum projection and minimum redundancy (MPMR). These two algorithms find a suitable feature subspace through matrix decomposition, and the feature subspace is then used to represent the original feature space. By exploiting the advantages of the matrix factorization technique, MFFS and MPMR can both achieve good performance. However, MFFS and MPMR ignore the underlying geometry information of the data itself. In contrast, this paper shows how such geometry information can be used to further improve the quality of feature selection.

A variety of literature has shown that the distributions of high dimensional data are often sparse. Such data contain a lot of local information, which is important for mining the internal structure of such data and improving the performance of nonlinear learning 34,35]. Some manifold learning algorithms have been proposed to discover the underlying manifold structure of data, such as Locality Preserving Projection [36], local linear embedding (LLE) [37] and Laplacian Eigenmap [38]. By analyzing the manifold structure of the data set, we can use the underlying geometric information to improve the learning efficiency of the algorithm.

Spectral graph theory [39,40] can be used to characterize the underlying manifold structure of the data. The spectral clustering method [39] exploits spectral graph theory to obtain good clustering performance. Based on nonnegative matrix factorization (NMF) [28], Cai et al. [41] proposed graph regularized nonnegative matrix factorization (GNMF), which uses the geometry information of the data itself to greatly improve performance. Compared with concept factorization (CF) [42], locally consistent concept factorization (LCCF) [43] shows better performance, because it is able to exploit the local structure of data. In recent years, new work [44–46] has shown that the manifold information of the data is not only distributed in the data space, but also in the feature space. In [44], Shang et al. proposed a graph dual regularization non-negative matrix factorization for co-clustering algorithm (DNMF). Ye et al. [46] proposed dual-graph regularized concept factorization clustering (GCF).

Some classification algorithms also use the spectral graph theory. Belkin et al. [47] proposed manifold regularization, a geometric framework for learning from labeled and unlabeled examples, which uses graph theory to exploit the manifold structure of the data. Experimental results show that this method can use unlabeled data effectively. Based on standard SVM, Chova et al. proposed semi-supervised image classification with Laplacian support vector machines (LapSVM) [48], which uses the geometry information of both labeled and unlabeled samples by using the graph Laplacian. Some experimental evidence suggests that LapSVM can outperform conventional SVM. In [49], Yang et al. proposed the Laplacian twin parametric-margin support vector machine for semi-supervised classification (LTPMSVM), which overcomes the shortcomings of conventional methods which are unable to effectively handle unlabeled data. LTPMSVM uses the geometric information of the unlabeled data to construct a better classifier, and experimental results have confirmed the strong performance of LTPMSVM.

Some feature selection algorithms which use local structure information have previously been proposed. Laplacian score (LapScor) [21], spectral feature selection (SPEC) [18], minimum redundancy spectral feature selection (MRSF) [50], unsupervised feature selection for multicluster data (MCFS) [51] are four well known algorithms. Extensions of MRSF and MCFS include clustering-guided sparse structural learning for unsupervised feature selection (CGSSL) [52] and joint embedding learning and sparse regression (JELSR) [53].

In this paper, we propose a new method called subspace learning-based graph regularized feature selection (SGFS). SGFS is based on the framework of subspace learning feature selection, which exploits the advantages of matrix factorization techniques. On this basis, we introduce the concept of graph regularization and preserve the local structure information of the feature space of the data. The local structure information of the feature space directly guides the learning of the coefficient matrix in the error reconstruction term, and indirectly guides the learning of the feature selection matrix. Additionally, we propose the use of the $\boldsymbol{L}_{2,1}$-norm to constrain the feature selection matrix, which guarantees its sparsity, so as to provide more accurate discrimination information for feature selection. We use an alternating iterative optimization mechanism to optimize the objective function and adjust the parameters to minimize the reconstruction error. Finally, we obtain the feature selection matrix. Through this matrix, we can calculate the scores of all the features, and select the most representative features.

The main contributions of this paper are as follows:

1. By using graph theory, the geometric structure information of the feature manifold is preserved. Through the guidance of geometry information, the learning of the feature selection matrix and coefficient matrix are more rapid and accurate.
2. By introducing $\boldsymbol{L}_{2,1}$-norm to constrain the feature selection term, the sparsity of the feature selection matrix is guaranteed, enabling more accurate discrimination information for feature evaluation.

The structure of this paper is organized as follows. In Section 2, we introduce the framework, the iterative update rules and convergence proof of SGFS. In Section 3, we present the experimental results of comparing the performance of SGFS against five other state-of-the-art algorithms on twelve public benchmark data sets. Section 4 provides concluding remarks.

## 2. Subspace Learning-based graph regularized feature selection

In this section, we present details of the SGFS method, which breaks down into three main parts: sparse subspace learning, local structure preserving and feature evaluation.

### 2.1. Related notations

First of all, we introduce some related notations. Denote $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \Re^{d \times n}$ as the unlabeled sample data set. Where $n$ and $d$ respectively represent the number and dimension of the samples. We use $l$ to indicate the number of selected features, $l \leq d$.

Given an arbitrary matrix $\mathbf{A} \in \Re^{e \times f}$, its $\boldsymbol{L}_{r,s}$ is defined as:

$$\|\mathbf{A}\|_{r,s} = \left( \sum_{i=1}^{e} \left( \sum_{j=1}^{f} |A_{ij}|^r \right)^{s/r} \right)^{1/s}. \quad (1)$$

According to the definition, when $r = s = 2$, it indicates Frobenius-norm or $\boldsymbol{L}_2$-*norm*. In contrast, when $r = 2$, $s = 1$, it represents sparse constraint $\boldsymbol{L}_{2,1}$-*norm*. We denote $\boldsymbol{L}_2$-*norm* and $\boldsymbol{L}_{2,1}$-*norm* respectively as $\|\cdot\|_2^2$ and $\|\cdot\|_{2,1}$ in the following.

### 2.2. Sparse subspace learning

#### 2.2.1. Distance between subspaces

According to [32], we first define the distance between subspaces. Given a vector group $\boldsymbol{X}$ in an $m$-dimensional real number space. We define $span(\mathbf{X}) = \{\mathbf{a}^T\mathbf{X}|\mathbf{a} \in \Re^{|\mathbf{X}|}\}$ as the spanning subspace of $\boldsymbol{X}$, which is the set of all combinations of elements of $\boldsymbol{X}$. Where, $|\mathbf{X}|$ is the basis of $\boldsymbol{X}$. Given two vector groups $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$