Knowledge-Based Systems 75 (2015) 19–29

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Unsupervised feature selection via maximum projection and minimum redundancy

Shiping Wang [a,b], Witold Pedrycz [b,c], Qingxin Zhu [a], William Zhu [d,*]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, T6G2G7, Canada
[c] System Research Institute, Polish Academy of Sciences, Warsaw, Poland
[d] Lab of Granular Computing, Minnan Normal University, Zhangzhou 363000, China

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form 23 October 2014
Accepted 8 November 2014
Available online 25 November 2014

Keywords:
Machine learning
Feature selection
Unsupervised learning
Matrix factorization
Kernel method
Minimum redundancy

ABSTRACT

Dimensionality reduction is an important and challenging task in machine learning and data mining. It can facilitate data clustering, classification and information retrieval. As an efficient technique for dimensionality reduction, feature selection is about finding a small feature subset preserving the most relevant information. In this paper, we propose a new criterion, called maximum projection and minimum redundancy feature selection, to address unsupervised learning scenarios. First, the feature selection is formalized with the use of the projection matrices and then characterized equivalently as a matrix factorization problem. Second, an iterative update algorithm and a greedy algorithm are proposed to tackle this problem. Third, kernel techniques are considered and the corresponding algorithm is also put forward. Finally, the proposed algorithms are compared with four state-of-the-art feature selection methods. Experimental results reported for six publicly datasets demonstrate the superiority of the proposed algorithms.

1. Introduction

In many practical applications, such as computer vision, machine learning and image processing, and high-dimensional data are common. When dealing with learning problems, high dimensionality calls for more computational time and space requirements. Therefore, how to reduce dimensions of the data is an important and challenging issue. Feature extraction and feature selection are the two main approaches to dimensionality reduction. The former aims to map the original features into a low-dimensional space via certain transformation (often linear transformation) and then generates some new features [1], while the latter aims to find an optimal feature subset given a certain predetermined criterion [2,3]. In fact, feature selection not only reduces the number of features and makes the learning algorithm more efficient, but also improves the performance in that it may avoid the overfitting phenomenon.

Whether the class labels are available, feature selection techniques can be classified into supervised and unsupervised methods. The supervised technique is to use the label information to evaluate the significance of features and then provide rankings of these features. Some widely supervised feature selection methods include Fisher score [4,5], mutual information [6,7] and Pearson correlation coefficient [8]. The unsupervised method is to find hidden structures in unlabeled data and build a feature selector using intrinsic properties of data without error or reward signals which can be used to guide the search [9,10]. However, the label information of many samples are often difficult to obtain or the cost to label these samples are often expensive, such as medical diagnosis and object detection data. Therefore, how to make full use of the unlabeled data and reveal their inherent rules for improving learning performance is necessary and important. This is also a challenging issue in that there is no class label information to guide the dependence and relevance among features.

Feature selection is to search the most relevant feature subset with certain criterion. Many feature selection methods often select the top ranked features, which are evaluated independently by each feature. In other words, these methods consider only the relevance and dependence between individual features, not an individual feature and a feature subset. However, ranking features underlying the relation between an individual feature and a feature subset is a combinational optimization problem. In recent years, the matrix factorization method has become popular techniques to deal with this type of relation for feature selection [11–14]. Drawing support from this method, feature selection

* Corresponding author.
E-mail address: williamfengzhu@gmail.com (W. Zhu).

http://dx.doi.org/10.1016/j.knosys.2014.11.008
0950-7051/

problems are transformed into finding optimal solutions of optimization problems, and then the corresponding matrix update iterative algorithms can be developed, which implies the optimality of a feature subset as a whole, not just some individual features.

In this paper, we propose a new criterion formalized in a matrix factorization form, called maximum projection and minimum redundancy feature selection, for unsupervised learning. First, the feature selection problem is formalized by maximum projection and minimum redundancy, and then its equivalent characterizations are presented from a theoretical viewpoint. Furthermore, a concise matrix factorization for feature selection is provided. Second, an iterative algorithm is proposed for this type of matrix factorization feature selection. Inspired by the characteristics of the projection matrix, a greedy algorithm is also designed. The feature selection is also incorporated with kernel methods, and then the corresponding algorithm is also developed. Third, four state-of-the-art algorithms for unsupervised feature selection are compared with the proposed methods. Experiments reported for six publicly datasets demonstrate that the proposed algorithms outperform the four methods in most cases for all tested datasets.

The paper is arranged as follows. A brief review of recent works on feature selection is presented in Section 2. The feature selection criterion with maximum projection and minimum redundancy is provided in Section 3. We propose two iterative and greedy algorithms for this criterion, and also one iterative algorithm for its kernel methods in Section 4. Experimental results are reported and analyzed in Section 5. Finally, this paper is concluded in Section 6.

2. Related works

By whether the feature selection is independent to the learning algorithm, feature selection techniques can be categorized into wrapper methods and filter methods. Wrapper methods evaluate the importance of features using the learning algorithms [15]. In other words, they "wrap" the feature selection process around the mining algorithms, such as the Bayes classifier [16], support vector machine [17] and clustering [18]. However, wrapper methods are usually computationally expensive [19] in that they need to repeatedly train learning algorithms and then may not be effective for high-dimensional data mining. Filter methods analyze the intrinsic characteristics of the data and select the top ranked features underlying certain criterion before doing learning task [20]. They are viewed as a pure preprocessing tool and fully irrelevant to the learning algorithms. In this paper, we are much interested in filter methods in that they are more efficient than wrapper methods.

Wrapper methods and filter methods can be supervised or unsupervised (even semi-supervised). We focus particularly on unsupervised filter methods since there is no label information to guide the search for features. Many feature selection methods score features using the dependence between two features, not between an individual feature and a feature subset. Therefore, matrix factorization techniques have been introduced for feature selection, which provides a batch mode for searching features. For example, Nie et al. proposed a robust feature selection criterion using joint $\ell_{2,1}$-norm minimization on both function loss and regularization [12]. Farahat et al. presented a matrix factorization criterion for feature selection and also provided an efficient greedy algorithm [21]. Yang et al. proposed an efficient online learning algorithm for multitask feature selection, which showed its big advantages in saving time complexity and memory cost [22]. Song et al. constructed a framework for feature selection using the Hilbert–Schmidt independence criterion, which was based on the assumption that good features should be highly relevant to the labels [23]. Liu et al. proposed a large margin subspace learning algorithm for feature selection, which aimed to maximize the margin of the given samples [24]. Zhao et al. introduced a similarity preserving criterion for feature selection, which encompassed many widely used criteria [25]. Li et al. proposed a nonnegative spectral analysis method to select the most discriminative features, which took into account both the discriminative information and feature correlation simultaneously [11].

Before proceeding with a detailed discussion, we summarize the notation used in this paper as Table 1.

3. Unsupervised feature selection criterion with maximum projection and minimum redundancy

A projection is a linear transformation from a vector space to itself with its image preserved. Feature selection can be viewed to select a feature subset such that all features are projected into this feature subspace with minimum reconstruction error. Specifically, the solution to the feature selection problems can be represented in the following form:

$$\arg\min_{I}\|X - P^{I}X\|_F^2 \tag{1}$$

where $P^I \in \mathbb{R}^{n\times n}$ is a projection matrix induced by the feature subset whose column index set is I. For example, we can specify projections as the orthogonal or the oblique projections. In fact, a matrix P can be viewed as a projection if and only if $P^2 = P$. The orthogonal projection of a feature subset is given as

$$P^I = X_I(X_I^TX_I)^{-1}X_I^T, \tag{2}$$

Here, if $X_I \in \mathbb{R}^{n\times|I|}$ is orthogonal by columns, then $X_IX_I^T$ is an isometry embedding the whole feature space into the selected feature subspace. However, when X_I is just linearly independent (not necessarily orthogonal), it still embeds the whole feature space into the underlying subspace but it is no longer an isometry in general. At that time, the matrix $(X_I^TX_I)^{-1}$ serves as a normalizing factor that recovers the isometry. It is noted that the independence condition of X_I can also be dropped, and we only need to replace $(X_I^TX_I)^{-1}$ with the Moore–Penrose pseudoinverse $(X_I^TX_I)^{+}$. The oblique projection of a feature subset is

$$P^I = X_I(B^TX_I)^{-1}B^T \tag{3}$$

where B is composed by a basis of the orthogonal complement space of the null space of X_I. It is evident that the oblique projection is an extension of the orthogonal projection. It is noted that there are many ways to construct a projection matrix and the above feature selection criterion is just a framework constructed by projections.

Table 1
Notation used in the study.

Notation	Representation
n	The number of instances
d	The number of features
$[k]$	$\{1,\cdots,k\}$
I	An index set contained in $[d]$
I^c	The complement of set I
$\|A\|$	The cardinality of set A
E_k	k-by-k unit matrix
M^T	The transpose of matrix M
M_I	The sub-matrix of M consisting of the set of I columns
$Tr(M)$	The trace of square matrix M, i.e., $Tr(M)=\sum_{i=1}^{n}M_{ii}$
$\|M\|_F$	Frobenius norm of M, i.e., $\|M\|_F=\left(\sum_{j=1}^{m}\sum_{i=1}^{n}M_{ij}^2\right)^{1/2}$