Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 1440–1451
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

# Constraint Score: A new filter method for feature selection with pairwise constraints

Daoqiang Zhang[a,*], Songcan Chen[a], Zhi-Hua Zhou[b]

[a]*Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China*
[b]*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China*

Received 17 April 2007; received in revised form 9 October 2007; accepted 12 October 2007

**Abstract**

Feature selection is an important preprocessing step in mining high-dimensional data. Generally, supervised feature selection methods with supervision information are superior to unsupervised ones without supervision information. In the literature, nearly all existing supervised feature selection methods use class labels as supervision information. In this paper, we propose to use another form of supervision information for feature selection, i.e. pairwise constraints, which specifies whether a pair of data samples belong to the same class (*must-link* constraints) or different classes (*cannot-link* constraints). Pairwise constraints arise naturally in many tasks and are more practical and inexpensive than class labels. This topic has not yet been addressed in feature selection research. We call our pairwise constraints guided feature selection algorithm as Constraint Score and compare it with the well-known Fisher Score and Laplacian Score algorithms. Experiments are carried out on several high-dimensional UCI and face data sets. Experimental results show that, with very few pairwise constraints, Constraint Score achieves similar or even higher performance than Fisher Score with full class labels on the whole training data, and significantly outperforms Laplacian Score.


*Keywords:* Feature selection; Pairwise constraints; Filter method; Constraint Score; Fisher Score; Laplacian Score

## 1. Introduction

With the rapid accumulation of high-dimensional data such as digital images, financial time series and gene expression microarrays, feature selection has been an important preprocessing step to machine learning and data mining. In many real-world applications, feature selection has shown very effective in reducing dimensionality, removing irrelevant and redundant features, increasing learning accuracy, and enhancing learning comprehensibility [1–3]. Typically, feature selection methods can be categorized into two groups, i.e., (1) filter methods [3] and (2) wrapper methods [4]. The filter methods evaluate the goodness of features by using the intrinsic characteristics of the training data and are independent on any learning algorithm. On the contrary, the wrapper methods directly use predetermined learning algorithms to evaluate the features. Generally, the wrapper methods outperform the filter methods in terms of accuracy, but the former are computationally more expensive than the latter. When dealing with data with huge number of features, the filter methods are usually adopted due to their computational efficiency. In this paper, we are particularly interested in the filter methods.

According to whether the class labels are used, feature selection methods can be divided into supervised feature selection [1] and unsupervised feature selection [5,6]. The former evaluates feature relevance by the correlation between features and class labels, while the latter evaluates feature relevance by the capability of keeping certain properties of the data, e.g., the variance or the locality preserving ability [7,8]. When labeled data are sufficient, supervised feature selection methods usually outperform unsupervised feature selection methods [9]. However, in many cases obtaining class labels is expensive and the amount of labeled training data is often very limited. Most traditional supervised feature selection methods may fail on such 'small labeled-sample problem' [10]. A recent important advance on this direction is to use both labeled and unlabeled data for feature selection, i.e. semi-supervised feature selection

* Corresponding author.
*E-mail address:* dqzhang@nuaa.edu.cn (D. Zhang).

0031-3203/$30.00 
doi:10.1016/j.patcog.2007.10.009

[11], which introduces the popular semi-supervised learning technique [12] into feature selection research. However, like in supervised feature selection, the supervision information used in semi-supervised feature selection is still class labels.

In fact, besides class labels, there exist other forms of supervision information, e.g. the *pairwise constraints*, which specifies whether a pair of data samples belongs to the same class (*must-link* constraints) or different classes (*cannot-link* constraints) [13–15]. Pairwise constraints arise naturally in many real-world tasks, e.g. image retrieval [13]. In those applications, considering the pairwise constraints is more practical than trying to obtain class labels, because the true labels may be unknown *a priori*, while it can be easier for a user to specify whether some pairs of examples belong to the same class or not, i.e. similar or dissimilar. Besides, the pairwise constraints can be derived from labeled data but not vice versa. Finally, unlike class labels, the pairwise constraints can sometimes be automatically obtained without human intervention. For those reasons, pairwise constraints have been widely used in distance metric learning [16] and semi-supervised clustering [12–14]. In one of our recent work, we have proposed to use pairwise constraints for dimension reduction [15].

It's worthy to note that one should neither confuse the pairwise constraints mentioned in this paper with the pairwise similarity or dissimilarity value used in spectral graph based algorithms [17–20], nor with some class pairwise methods [21]. In spectral graph based algorithms, one first computes the pairwise similarity or dissimilarity between samples to form the similarity or dissimilarity matrix, and then perform subsequent operations on it. On the other hand, in class pairwise methods, e.g. class pairwise feature selection [21], one takes the subsets of features which are the most effective in discriminating between all possible pairs of classes. Apparently, both are very different from the pairwise constraints mentioned in this paper.

In this paper, we propose to use pairwise constraints for feature selection. To the best of our knowledge, we have not noticed any similar work on this topic before. We devise two novel score functions based on pairwise constraints to evaluate the feature goodness and name the corresponding algorithms as Constraint Score. Experiments are carried out on several high-dimensional UCI and face data sets to compare the proposed algorithm with established feature selection methods such as Fisher Score [9] and Laplacian Score [7], etc. Experimental results show that, with a few pairwise constraints, Constraint Score achieves similar or even higher performance than Fisher Score with full class labels on the whole training data, and significantly outperforms Laplacian Score.

The rest of this paper is organized as follows. Section 2 first introduces the background of this paper and briefly shows several existing score functions used in supervised and unsupervised feature selection. Then we present the Constraint Score algorithm in Section 3. Section 4 reports on the experimental results. Finally, Section 5 concludes this paper with some future work.

## 2. Background

In this section, we briefly introduce several score functions popularly used in feature selection methods, including Variance [9], Laplacian Score [7] and Fisher Score [9]. Among them, Variance and Laplacian Score are unsupervised, while Fisher Score is supervised.

Variance might be the simplest unsupervised evaluation of the features. It uses the variance along a dimension to reflect its representative power and those features with the maximum variance are selected. Let $f_{ri}$ denote the $r$th feature of the $i$th sample $\mathbf{x}_i$, $i = 1, \ldots, m$; $r = 1, \ldots, n$. Define $\mu_r = \frac{1}{m}\sum_i f_{ri}$. Then, the Variance score of the $r$th feature $V_r$, which should be maximized, is computed as follows [9]:

$$V_r = \frac{1}{m}\sum_{i=1}^{m}(f_{ri} - \mu_r)^2. \tag{1}$$

Another unsupervised feature selection method, i.e. Laplacian Score, makes a further step on Variance. It not only prefers to those features with larger variances which have more representative power, but also prefers to selecting features with stronger locality preserving ability. A key assumption in Laplacian Score is that data from the same class are close to each other. The Laplacian score of the $r$th feature $L_r$, which should be minimized, is computed as follows [7]:

$$L_r = \frac{\sum_{i,j}(f_{ri} - f_{rj})^2 S_{ij}}{\sum_i (f_{ri} - \mu_r)^2 D_{ii}}, \tag{2}$$

where $D$ is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$, and $S_{ij}$ is defined by the neighborhood relationship between samples $\mathbf{x}_i$ $(i = 1, \ldots, m)$ as follows:

$$S_{ij} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t}} & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are neighbors,} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $t$ is a constant to be set, and '$\mathbf{x}_i$ and $\mathbf{x}_j$ are neighbors' means that either $\mathbf{x}_i$ is among $k$ nearest neighbors of $\mathbf{x}_j$, or $\mathbf{x}_j$ is among $k$ nearest neighbors of $\mathbf{x}_i$.

In contrast to Variance and Laplacian Score, Fisher Score is supervised with class labels and it seeks features with best discriminant ability. Let $n_i$ denote the number of samples in class $i$. Let $\mu_r^i$ and $(\sigma_r^i)^2$ be the mean and variance of class $i$, $i = 1, \ldots, c$, corresponding to the $r$th feature. The Fisher Score of the $r$th feature $F_r$, which should be maximized, is computed as follows [9]:

$$F_r = \frac{\sum_{i=1}^{c} n_i(\mu_r^i - \mu_r)^2}{\sum_{i=1}^{c} n_i(\sigma_r^i)^2}. \tag{4}$$

## 3. Constraint Score

In this paper, we formulate the pairwise constraints guided feature selection as follows: Given a set of data samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]$, and some supervision information in the form of pairwise must-link constraints $M = \{(\mathbf{x}_i, \mathbf{x}_j) | \mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class$\}$ and pairwise