Unsupervised Feature Selection

Yale Chang

1 Introduction

2 Spectral Feature Selection

Spectral feature selection [1] identifies relevant features by measuring their capability of preserving sample similarity.

3 Spectral Feature Selection with Minimum Redundancy [2]

This is an embedded model that evaluates the utility of a set of features jointly and can effectively remove redundant features. The algorithm is derived from a formulation based on multi-output regression and feature selection is achieved by enforcing sparsity through applying $L_{2,1}$-norm constraint on the solutions. The key idea is: to identify feature redundancy, features must be evaluated jointly.

MRSF: minimum redundancy spectral feature selection.

Given data matrix $X \in \mathbb{R}^{d \times n}$, similarity matrix $S \in \mathbb{R}^{n \times n}$, eigendecomposition gives low dimensional embedding matrix $Y \in \mathbb{R}^{n \times q}$, achieve feature selection by solving the following optimization problem

$$\min_{W,c} ||Y - X^T W||_F^2 + \lambda ||W||_{2,1}$$ (1)

where $W \in \mathbb{R}^{d \times q}$ is the projection matrix. [3] proposed a similar formulation

$$\min_{W} ||Y - X^T W||_{2,1} + \lambda ||W||_{2,1}$$ (2)

4 Joint Feature Selection and Subspace Learning [4]

$$\min_{W \in \mathbb{R}^{d \times q}} \text{Tr}(W^T X L X^T W) + \lambda ||W||_{2,1} \quad \text{s.t.} \quad W^T X D X^T W = I$$ (3)

5 Feature Selection via Joint Embedding Learning and Sparse Regression [5]

$$\min_{W,Y} \text{Tr}(Y^T L Y) + \beta(||X^T W - Y||_F^2 + \alpha ||W||_{2,1})$$ (4)

where $Y \in \mathbb{R}^{n \times q}$, $Y^T Y = I_{q \times q}$, $L = (I_{n \times n} - S)^T (I_{n \times n} - S)$ is the graph Laplacian of Local Linearity Embedding, $W \in \mathbb{R}^{d \times q}$. 

\[
\min_{W,Y} \text{Tr}(Y^T L Y) + \beta(||X^T W - Y||_F^2 + \alpha||W||_{2,1})
\] (5)

where \(Y \in \mathbb{R}^{n \times q} \), \(Y^T Y = I_{q \times q} \), \(Y \geq 0 \), \(L = I_{n \times n} - D^{-1/2}SD^{-1/2} \), \(W \in \mathbb{R}^{d \times q} \). When both nonnegative and orthogonal constraints are satisfied, there is only one element in each row of \(F\) is greater than zero and all the others are zeros. In that way, the learned \(F\) is more accurate, and more capable to provide discriminative information.

7 Unsupervised Feature Selection for Linked Social Media Data [7]

\[
\min_{W} \text{Tr}(W^T X L X^T W) + \beta||W||_{2,1} + \alpha\text{Tr}(W^T X(I_n - FF^T)W^T W)
\] (6)

s.t. \(W^T X X^T + \lambda I_{n \times n}W = I_{q \times q}\)

where \(W \in \mathbb{R}^{d \times q} \), \(L = D - S\) is a Laplacian matrix, \(F = H(H^T H)^{-1/2}\) is the weighted social dimension indicator matrix, \(H \in \mathbb{R}^{K \times n}\) is the social dimension indicator matrix, which can be obtained through modularity maximization.

8 Feature Selection by Joint Graph Sparse Coding [8]

\[
\min_{B,G} ||X - BG^T||_F^2 + \alpha\text{Tr}(G^T L G) + \lambda||G^T||_{2,1}
\] (7)

s.t. \(\sum_{i=1}^{d} \sum_{j=1}^{q} b_{i,j}^2 \leq 1\)

9 Robust Unsupervised Feature Selection [9]

\[
\min_{B,G,W} ||X - BG^T||_{2,1} + \nu\text{Tr}(G^T L G) + \alpha||X^T W - G||_{2,1} + \beta||W||_{2,1}
\] (8)

where \(G \in \mathbb{R}^{n \times q} \), \(B \in \mathbb{R}^{d \times q} \), \(W \in \mathbb{R}^{d \times q} \).

10 \(L_{2,1}\)-Norm Regularized Discriminative Method [10]

Supervised feature selection algorithms, e.g., Fisher score [11], robust regression [3], sparse multi-output regression [2] and trace ratio [12], usually select features according to labels of the training data. In unsupervised scenarios, label information is not available and a frequently used criterion is to select the features which best preserve the data similarity or manifold structure derived from the whole feature set [1,13,14]. Instead of evaluating the importance of each feature individually [1,13],
feature correlation should be taken into account. While [2,14] apply spectral regression and consider feature correlation in two steps, this algorithm is a one-setp approach.

$$\min_{W^TW=I} \text{Tr}(W^TMW) + \gamma \|W\|_{2,1}$$

(9)

where $M$ is constructed from local discriminative information.

11 Discriminant Analysis for Unsupervised Feature Selection [15]

12 Embedded Unsupervised Feature Selection [16]

References


