Unsupervised Feature Selection for Multi-Cluster Data

Deng Cai et al, KDD 2010

Presenter: Yunchao Gong
Dept. Computer Science, UNC Chapel Hill
Introduction
The problem

• Feature selection for high dimensional data

• Assume high dimensional $n \times d$ data matrix $X$

• When $d$ is too high ($d > 10000$), it causes problems
  – Storing the data is expensive
  – Computational challenges
A solution to this problem

• Select the most informative features

• Assume high dimensional $n*d$ data matrix $X$
After feature selection

- More scalable and can be much more efficient
Previous approaches

• Supervised feature selection
  – Fisher score
  – Information theoretical feature selection

• Unsupervised feature selection
  – LaplacianScore
  – MaxVariance
  – Random selection
Motivation of the paper

• Improving clustering (unsupervised) using feature selection

• Automatically select the most useful feature by discovering the manifold structure of data

(a) Visualization of digit 1, 2, 3 and 4.
Motivation and Novelty

• Previous approaches
  – Greedy selection
  – Selects each feature independently

• This approach
  – Considers all features together
A toy example

Figure 1: A failed example for binary clusters/classes feature selection methods. (a)-(c) show the projections of the data on the plane of two joint features, respectively. Without the label information, both Maximum variance and LaplacianScore [17] methods rank the features as $a > b > c$. If one is asking to select two features, both Maximum variance and LaplacianScore methods will select features $a$ and $b$, which is obviously sub-optimal.
Some background—manifold learning

- Tries to discover manifold structure from data

(a) Visualization of digit 1, 2, 3 and 4.
manifolds in vision

appearance variation
Some background—**manifold learning**

- Discovers smooth manifold structure
- Map different manifold (classes) as far as possible

(b) Visualization of airplane (r), apple (g) and construction (b), sampling 5000 points.
The proposed method
Outline of the approach

1. manifold learning for cluster analysis
2. sparse coefficient regression
3. sparse feature selection
Outline of the approach

1. Generate response $Y$

2. Scatter, Correlation, and Regression

3. $MCFS(j) = \max_k |a_{k,j}|$

\[
\min_{a_k} \|y_k - X^T a_k\|^2 + \beta |a_k|
\]
1. Manifold learning for clustering

- Manifold learning to find clusters of data

(a) Visualization of digit 1, 2, 3 and 4.
1. Manifold learning for clustering

• Observations
  – Points in the same class are clustered together

(a) Visualization of digit 1, 2, 3 and 4.
1. Manifold learning for clustering

• Ideas?
  – How to discover the manifold structure?

(a) Visualization of digit 1, 2, 3 and 4.
1. Manifold learning for clustering

• Map similar points closer
• Map dissimilar points faraway

$$\min \sum_{i,j} K_{ij} (f_i - f_j)^2$$

similarity

Data points
1. Manifold learning for clustering

- Map similar points closer
- Map dissimilar points faraway

\[
\min \sum_{i,j} K_{ij} (f_i - f_j)^2
\]

\[= f^T L f, \quad L = D - K\]

\[D = \text{diag}(\text{sum}(K))\]
1. Manifold learning for clustering

• Constraining $f$ to be orthogonal ($f^T D f = I$) to eliminate free scaling

• So we have the following minimization problem

\[
(D - K) f = \lambda D f
\]
Summary of manifold discovery

• Construct graph $K$

• Choose a weighting scheme for $K$ ($e^{-\frac{||x_i-x_j||^2}{\sigma}}$)

• Perform $\left( D - K \right) f = \lambda D f$

• Use $f$ as the response vectors $Y$
Response vector $\mathbf{Y}$

• $\mathbf{Y}$ reveals the manifold structure!

(a) Visualization of digit 1, 2, 3 and 4.
2. Sparse coefficient regression

• When we have obtained $Y$, how to use $Y$ to perform feature selection?

• $Y$ reveals the cluster structure, use $Y$ as response to perform sparse regression

$$\min_{a_k} \left\| y_k - X^T a_k \right\|^2 + \beta |a_k|$$
Sparse regression

• The "Lasso"

\[
\min_{a_k} \left\| y_k - X^T a_k \right\|^2 + \beta |a_k|
\]

sparsely
Steps to perform sparse regression

- Generate \( Y \) from step 1
- Data matrix input \( X \)
- For each column of \( Y \), we denote it as \( Y_k \)
- Perform the following step to estimate \( a_k \)

\[
\min_{a_k} \left\| y_k - X^T a_k \right\|^2 + \beta |a_k|
\]
$X \quad \quad a \quad \quad \quad = \quad \quad \quad Y_k$

nonzero
3. Sparse feature selection

• But for $Y$, we will have $c$ different $Y_k$

• How to finally combine them?

• A simple heuristic approach

\[
MCFS(j) = \max_k |a_{k,j}|
\]
illustration
The final algorithm

1. manifold learning to obtain $Y$

\[(D - K)f = \lambda Df \quad , \quad Y = f\]

2. sparse regression to select features

\[
\min_{a_k} \| y_k - X^T a_k \|^2 + \beta |a_k| \]

3. final combination

\[
MCFS(j) = \max_k |a_{k,j}| \]
Step 1

Response $Y =$

(a) Visualization of digit 1, 2, 3 and 4.

$$(D - K)f = \lambda Df$$
Step 2

\[
\min_{a_k} \left\| y_k - X^T a_k \right\|^2 + \beta \left| a_k \right|
\]
Step 3

\[ MCFS(j) = \max_k |a_{k,j}| \]
Discussion
Novelty of this approach

• Considers all features

• Uses sparse regression `lasso’’ to perform feature selection

• Combines manifold learning with feature selection
Results
Face image clustering

Figure 2: C

n ORL data set.
Digit recognition

Figure 3: Clustering results for the MNIST data set.
Summary

– The method is technically new and interesting

– But not ground breaking
Summary

• Test on more challenging vision datasets
  – Caltech
  – Imagenet