Enhancing Semi-Supervised Clustering: A Feature Projection Perspective

Wei Tang\textsuperscript{1}, Hui Xiong\textsuperscript{2}, Shi Zhong\textsuperscript{3}, Jie Wu\textsuperscript{1}

\textsuperscript{1}Department of Computer Science & Engineering  
Florida Atlantic University

\textsuperscript{2}Management Science & Information System Department  
Rutgers University

\textsuperscript{3}Data Mining and Research Group  
Yahoo! Inc
Outline

⇒ Introduction

  • The SCREEN Algorithm

  • Experimental Results

  • Related Works

  • Conclusions
Introduction

• In many application domains:
  ◦ Large volume of unlabeled data
  ◦ Limited supervision:
    * Labeled instances
    * Pairwise instance constraints

• Semi-supervised clustering
  ◦ Combining unlabeled and labeled instances
  ◦ Improving the clustering performance through supervision
Research Motivation

- Various applications often contain high dimensional sparse data
  - text documents, market basket data

- Traditional semi-supervised clustering methods:
  - constraint-based, distance based, and hybrid methods

- Most existing methods are not designed for handling those data
  - Euclidean notion of density is not very meaningful in high-dimensional data

- There is a need to incorporate feature reduction into the process of semi-supervised clustering
Outline

• Introduction

⇒ The SCREEN Algorithm

• Experimental Results

• Related Works

• Conclusions
Problem Formulation

Given:

- A set of $d$-dimensional instances $X$
- A set of must-link constraints $C_{ML}$
- A set of cannot-link constraints $C_{CL}$
- A pre-specified reduced dimension $k \ll d$
- A desired number of clusters $K$

Find:

- $K$ clusters of instances represented in reduced $k$-dimensional vector which satisfies the given instance constraints.
The Framework of the SCREEN Algorithm

Step 1 Initialization
Step 2 Constraint-guided feature projection
Step 3 Constrained Spherical $k$-means on projected data
Initialization - An Example

- Since must-links represent an equivalence relation, it enables us to replace each transitive closure of must-links with its average.

- sets \{a_1, a_2, a_3\}, \{b_1, b_2, b_3, b_4, b_5\}, and \{c_1, c_2, c_3\} represent different transitive closures enforced by must-links.

- After the initialization:
  - The pairwise constraints $C_{ML}$ and $C_{CL}$ are reduced to $C'_{CL}$
  - The original data sets $\mathcal{X}$ are reduced to $\mathcal{X}'$ with $\mathcal{W}'$
Constraint-Guided Feature Projection - SCREEN$_{PROJ}$

• Given
  ◦ A set of cannot-link constraints $C'_{CL}$
  ◦ A set of instances $\mathcal{X}'$ with weight $\mathcal{W}'$

• Objective: find a projection matrix $F$, such that

$$f = \sum_{(x'_1, x'_2) \in C'_{CL}} \| w_1 w_2 \cdot F^T (x'_1 - x'_2) \|^2$$

is maximized subject to the constraints

$$F_i^T F_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$
Solution To the Feature Projection Problem

- The Lagrangian of the above optimization problem is

\[ L_{F_1, \ldots, F_k} = f(F_1, \ldots, F_k) - \sum_{l=1}^{k} \xi_l (F_l^T F_l - 1). \]

which can be solved as

\[ \frac{\partial L}{\partial F_l} = 2MF_l - 2\xi_l F_l = 0 \quad \forall l = 1, \ldots, k \]

\[ \Rightarrow MF_l = \xi_l F_l \quad \forall l = 1, \ldots, k. \]  

(1)

**Theorem 1** Given the desired dimensionality \( k \) \((k < d)\), the set of cannot-link constraints \( C'_{CL} \), and the covariance matrix \( M = \text{cov}(C) \), the optimal projection matrix \( F_{d \times k} \) is comprised of the first \( k \) eigenvectors of \( M \) corresponding to the \( k \) largest eigenvalues.
Constrained Spherical \( K \)-means

- Updating rule in applying pairwise constraints
  
  ◦ Given each cannot-link constraint \((x'_i, x'_j) \in C_{CL}\)
  
  ◦ Find two different cluster centroids \(\mu_{x'_i}\) and \(\mu_{x'_j}\) such that
    \[
    w_i \cdot x'_{i}^T \mu_{x'_i} + w_j \cdot x'_{j}^T \mu_{x'_j}
    \]
    is maximized.
  
  ◦ Assign \(x'_i\) and \(x'_j\) to these two centroids to avoid violating the constraints.
Outline

• Introduction

• The SCREEN Algorithm

⇒ Experimental Results

• Related Works

• Conclusions
Experimental Setup

- Experimental Platform
  - GNU/Linux workstation with 4 Intel Xeon 2.8 GHz CPUs and 2G main memory

- Experimental Data Sets
  - Six data sets from UCI Machine Learning Repository
  - Six data sets from TREC collection
  - Nine data sets from 20-Newsgroups corpus

- Evaluation Measure: (Normalized Mutual Information)

\[
NMI = \frac{I(\hat{Z}; Z)}{(H(\hat{Z}) + H(Z))/2}
\]

where \( I(\hat{Z}; Z) \) is the mutual information between the random variables \( \hat{Z} \) and \( Z \), \( H(Z) \) is the Shannon entropy of \( Z \).
Effectiveness of SCREEN$_{PROJ}$ (1)

- Compared with original, PCA and RCA on low dimensional data
- Measured by NMI

\[(e)\text{ Vehicle (N=846, C=4, D=18, d=5)}\]
\[(f)\text{ Wine (N=178, C=3, D=13, d=5)}\]
Effectiveness of SCREEN$_{PROJ}$ (2)

- Conclusions:
  - RCA performs the best in the low dimensional data; however is not a good choice in handling high dimensional data
  - SCREEN$_{PROJ}$ is comparable to, or better than PCA in low dimensional data; especially achieve good performance on high dimensional data
Must-links vs. Cannot-links

- Incorporate $\beta$ into the previous objective function and varies from $0.0$ to $1.0$

$$f = (1 - \beta) \cdot \sum_{(x_1, x_2) \in C_{CL}} \| F^T (x_1 - x_2) \|^2 - \beta \cdot \sum_{(x_1, x_2) \in C_{ML}} \| F^T (x_1 - x_2) \|^2$$

![Normalized Mutual Information](c) Wei Tang KDD 2007
The Choice of Dimension $K$

- The SCREEN algorithm on different values of $k$ from 10 to 100

- Clustering performance is maximized when $k$ is between 20 and 40.
Computational Performance of the SCREEN Algorithm

- SCREEN ranks third due the extra cost of feature projection.
- SCREEN is much faster than the PCSKM+M algorithm which employs metric learning in the high dimensional data.
Clustering Performance of the SCREEN Algorithm

- SCREEN is more stable compared to the other methods.
- SCREEN always outperforms the PCSKM+M via metric learning and MPCSKM via HMRF model.
Outline

• Introduction

• The SCREEN Algorithm

• Experimental Results

⇒ Related Works

• Conclusions
Related Works (1)

- From the perspective of semi-supervised clustering
  - Constraint-based methods (PCSKM)
    - guide the clustering process by supervision
  - Distance-based methods (PCSKM+M)
    - learn an adaptive distance based on constraints
  - Hybrid methods (MPCSKM)
    - combines them into an unified statistical framework
Related Works (2)

- From the perspective of feature projection

  - Principal Component Analysis (PCA)
    - without utilizing any supervision

  - Fisher’s Linear Discriminant Analysis (LDA)
    - need to get the exact class information

  - Relevant Component Analysis (RCA)
    - based only on must-link constraints

  - Many others: projected clustering, CLIQUE
Outline

- Introduction
- The SCREEN Algorithm
- Experimental Results
- Related Works

⇒ Conclusions
Conclusions

- Formulate the constraint-guided feature projection into an optimization problem and give a closed-form solution

- Propose the SCREEN algorithm which integrates feature projection into semi-supervised clustering

- Experimental comparison between the SCREEN algorithm and the other methods
Questions?

- Email: wtang@cs.utexas.edu
- URL: http://www.cs.utexas.edu/~wtang

Thank You!