Crowdclustering with Sparse Pairwise Labels:
A Matrix Completion Approach

Jinfeng Yi, Rong Jin, Anil K. Jain, Shaili Jain
2012

Presented By: KHALID ALKOBAYER

Outline

- Introduction
  - Crowdsourcing and Crowdclustering
  - Ensemble clustering
  - Bayesian generative model
- Novel Crowdclustering approach
- Experiments
  - Data Sets
  - Baseline and evaluation metrics
- Results and Analysis
  - Full Annotations
  - Sampled annotations
- Conclusion and Comments

Introduction

- What is the Crowdsourcing?
  - It is a new business model provides an easy and inexpensive way to:
    1. Accomplish small-scale tasks (e.g., HITS).
    2. Utilize human capabilities to solve difficult problems.
  - Scenario:
    - Each human worker is asked to solve a part of a big problem.
    - Develop a computational algorithm to combine the partial solutions into an integrated one (single data partition).
  - Examples: Classification, clustering, and segmentation.

- It provides similarity measure between objects based on manual annotations → Data Clustering Problem.
Introduction

• Crowdclustering

  Scenario:
  A collection of objects need to be clustered
  → divide to subsets of objects
  → sample each subset of objects in each HIT
  → each worker annotates the subset of objects in each HIT
  → partial clustering
  → a single data partition

  Annotation types:
  grouping objects (similarity)
  Describing individual objects (keywords)

Introduction

• Ensemble clustering:
  To combine the partial clustering results, generated by individual workers, into a complete data partition (C).

  Challenges:
  1. Each worker for only a subset → partial clustering results.
  2. Different human workers → different partial clustering results.
  → Noise and inter-worker variations in the clustering results.
  → Uncertain data pairs observed.
  → Create inappropriate data partitions.

Introduction

• Bayesian Generative Model (The baseline algorithm):
  To address the large variations in the pairwise annotation labels provided by different workers.

  It explicitly models the hidden factors that are deployed by individual workers to group objects into the same cluster.

  But it requires a large number of manual annotation, or HITs, to discover the hidden factors.
  → High cost in computation and annotation.
  → limits the scalability to clustering large data set.

Novel Crowdclustering Approach

• Matrix Completion Approach (The proposed algorithm):
  o A novel approach based on the theory of matrix completion.
  o Matrix of low rank can be recovered by only a few entries.
  o To overcome the limitation of the Bayesian approach.

  Basic Idea:
  • Compute a partially observed similarity matrix based only on the reliable pairwise annotation labels. (uncertain data pairs = unobserved)
  • Use a matrix completion algorithm to complete the partially observed similarity matrix by filtering out the unobserved entries.
  • Apply a spectral clustering algorithm to the completed similarity matrix to obtain the final data partition (final clustering).
Novel Crowdclustering Approach

- Matrix Completion Approach (The proposed algorithm)

Advantages:

1. Only small number of pairwise annotations are need to construct the partially observed similarity matrix.

2. By filtering out the uncertain data pairs, this approach is less sensitive to the noisy labels → more robust clustering of data.

Novel Crowdclustering Approach

- Matrix Completion Approach (The proposed algorithm)

Similarity Matrix (SM) : from the partial clustering result of $k^{th}$ HIT

\[ W_{ij}^k = \begin{cases} 1 & \text{if objects } i \text{ and } j \text{ are assigned to the same cluster} \\ 0 & \text{if they are assigned to different clusters} \\ -1 & \text{if the pairwise label for the two objects can not be derived from the partial clustering result.} \end{cases} \]

$N$ = total number of objects. ; $i,j = 1,\ldots,N$

$A$ is a matrix of the average $W_{ij}^k$ for m HITs ; $k = 1,2,\ldots,m$

Experiments

- Image data Sets:

  1. Scenes Data Set

     1,001 images, 13 categories, 131 workers.

     HIT: To group images into multiple clusters.

     Number of clusters was determined by individual workers.

     Pairwise labels: partial clustering results generated in HITs.

     Data Source (Gomes et al. 2011).

Figure 1: Some sample images from the 13 categories in the Scenes data set.
Experiments

- Image data Sets:
  2. Tattoo Data Set
    3,000 images, 3 categories (Human, Animal, and Plant).
    HIT: To annotate tattoo images with keywords of the workers’ choice.
    On average, each image is annotated by 3 workers.
    Pairwise labels: comparing the number of matched keywords between images to a threshold (\(=1\)).

![Some sample images from the three categories in the Tattoo data set](image)

Experiments

- Baseline and evaluation metrics
  
  - Normalized mutual information metric (NMI):
    
    Ground truth partition \(C = \{C_1, C_2, ..., C_r\}\) of \(r\) clusters.
    
    Partition generated by a clustering algorithm \(C' = \{C'_1, C'_2, ..., C'_r\}\).
    The NMI of partition \(C\) and \(C'\) is:
    
    \[
    NMI(C, C') = \frac{2MI(C, C')}{H(C) + H(C')}
    \]
    
    Where,
    
    \(MI(X, Y)\): mutual information between random variables \(X\) and \(Y\).
    
    \(H(X)\): Shannon entropy of random variable \(X\).

  - Pairwise F-measure metric (PMF):
    
    \(A\) = a set of data pairs that share same class labels according to the ground truth.
    
    \(B\) = a set of data pairs that are assigned to the same cluster by a clustering algorithm.
    
    The PMF is:
    
    \[
    PWF = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
    \]
    
    
Experiments

- Baseline and evaluation metrics

  - Normalized mutual information NMI.
  
  - Pairwise F-measure PWF.
    
    \(NMI\) and \(PWF\) values \([0, 1]\),
    
    where \(1\) = Perfect Match, and \(0\) = Completely mismatch.

To evaluate the efficiency:

Measuring the running time.

Software: MATLAB
Results & Analysis

First experiment (with full annotations)

- Perform the algorithms on the Scenes and tattoo data sets.
- Use all the pairwise labels.
- For both data sets, $d_0 = 0$, and $d_1 = 0.9$ (Scenes), and 0.5 (Tattoo).

Two Criteria for choosing $d_1$:

- $d_1$ should be large enough to ensure that most pairwise labels are consistent with the cluster assignments.
- $d_1$ should be small enough to obtain sufficiently large number of entries with value 1 in the partially observed matrix $A$.

Results & Analysis

First experiment result:

Evaluation of the accuracy and efficiency:

**Scenes data set**: The proposed algorithm has similar, slightly lower performance as Bayesian, but significantly lower runner time.

**Tattoo data set**: The proposed algorithm outperform the Bayesian for both.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Scenes Data Set</th>
<th>Tattoo Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix Completion</td>
<td>0.758, 0.584</td>
<td>0.398, 0.595</td>
</tr>
<tr>
<td>Bayesian Method</td>
<td>0.764, 0.618</td>
<td>0.292, 0.524</td>
</tr>
</tbody>
</table>

The higher efficiency is due to that the proposal algorithm uses only a subset of reliable pairwise labels while Bayesian needs to all of them.

Results & Analysis

First experiment result:

Examine how the conditions are satisfied for both data sets:

**Condition**: A majority of the reliable pairwise labels derived from manual annotation should be consistent with the cluster assignments.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenes data set</td>
<td>95%</td>
</tr>
<tr>
<td>Tattoo data set</td>
<td>71%</td>
</tr>
</tbody>
</table>

Results & Analysis

First experiment result:

Evaluate the significance of the filtering step:

Observing: A large portion of pairwise labels derived from the manual annotation process are inconsistent with the cluster assignments. (80% for Scenes data set).

Figure 3: Sample image pairs that are grouped into the same cluster by more than 50% of the workers but are assigned to different clusters according to the ground truth.
Results & Analysis

First experiment result:

Observe how noisy labels affect the proposal algorithm:

Fix $d_0 = 0$,

Vary $d_1$ from 0.1 to 0.9 ( $d_1$ to determine the reliable pairwise labels)

From table 2 below:
The higher the percentage of consistent pairwise labels $\Rightarrow$ better performance.

Table 2: Performance of the proposed clustering algorithm as a function of different threshold values and the percentage of 1 entries in the matrix. $\hat{A}$ that are consistent with the cluster assignments for the Scenes data set.

<table>
<thead>
<tr>
<th>Threshold $d_1$</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistency percentage</td>
<td>18.02%</td>
<td>28.10%</td>
<td>35.53%</td>
<td>43.94%</td>
<td>61.79%</td>
</tr>
<tr>
<td>NMI</td>
<td>0.507</td>
<td>0.646</td>
<td>0.678</td>
<td>0.700</td>
<td>0.738</td>
</tr>
<tr>
<td>PWF</td>
<td>0.327</td>
<td>0.412</td>
<td>0.431</td>
<td>0.445</td>
<td>0.584</td>
</tr>
</tbody>
</table>

Results & Analysis

Second experiment results:

Expected result: reducing number of annotations $\Rightarrow$ lower performance

The proposed algorithm is more robust and performs better for all levels because it needs to small number of reliable pairwise labels to recover the cluster assignment matrix.

Results & Analysis

• Second experiment (with sampled annotations)

Objective: To verify how to obtain accurate clustering result even with small number of manual annotations.

Scenes data set: Using annotations that provided by 20, 10, 7, and 5 randomly sampled workers.

Tattoo data set: Randomly sample 10%, 5%, 2%, and 1% of all annotations. (3 annotators per image)

- Run both algorithms on the sampled annotations.
- Repeat the experiment 5 times and report the average performance (NMI).

Results & Analysis

Second experiment results:

The baseline algorithm needs to a large number to overcome the noisy labels and hidden factors.

As the numbers of annotations decreases $\Rightarrow$ significant reduction in the performance of the baseline algorithm.
Conclusion and Comments

- Crowdclustering uses the crowdsourcing technique to solve data clustering problems.

- The matrix completion approach improves the performance of the crowdclustering assignments.

- To derive the full similarity matrix, we need a subset of data pairs with reliable pairwise labels as the input for a matrix completion algorithm.

- A sufficient number of workers are needed to determine the reliable data pairs.

- The proposed algorithm needs just for smaller number of pairwise labels than the baseline algorithm which leads to low cost in both computation and annotation.

- To reduce the number of pairwise labels, we can reduce the number of workers or the number of HITs per worker.