Pseudo-Label Generation For Multi-Label Text Classification

Mohammad Salim Ahmed\textsuperscript{1}  
Latifur Khan\textsuperscript{1}  
Nikunj Oza\textsuperscript{2}

\textsuperscript{1}Department of Computer Science, The University of Texas at Dallas  
\textsuperscript{2}NASA Ames Research Center

CIDU Presentation, October 2011
Outline

1 Introduction

2 Subspace Clustering
   - Objective Function
   - Components of Subspace Clustering
   - Steps

3 Experimental Setup
   - Data Sets
   - Baseline Approaches
   - Results

4 Conclusions
Outline

1. Introduction

2. Subspace Clustering
   - Objective Function
   - Components of Subspace Clustering
   - Steps

3. Experimental Setup
   - Data Sets
   - Baseline Approaches
   - Results

4. Conclusions
Outline

1. Introduction
2. Subspace Clustering
   - Objective Function
   - Components of Subspace Clustering
   - Steps
3. Experimental Setup
   - Data Sets
   - Baseline Approaches
   - Results
4. Conclusions
Outline

1. Introduction
2. Subspace Clustering
   - Objective Function
   - Components of Subspace Clustering
   - Steps
3. Experimental Setup
   - Data Sets
   - Baseline Approaches
   - Results
4. Conclusions
Text Data Classification faces two major problems:

- High and Sparse Dimensionality
- Multi-Labelity
ON ARRIVAL AT DESTINATION AIRPORT … AND FINAL WEATHER WAS RADIO FACILITY AND TURBULENCE . THE APPROACH …. MAINTAIN 2500 FEET RUNWAY HEADING . THIS WAS NOT PER THE MISSED APPROACH ALTITUDE AND FOR A MOMENT IT WAS THOUGHT TO BE A POSSIBLE ERROR . CONTROLLER THEN CALLED BACK SAYING MAINTAIN 3000 FEET RUNWAY HEADING AND HANDED US OVER TO ANOTHER CONTROLLER . DURING THIS TIME MY ALTITUDE BALLOONED UP APPROXIMATE 300-400 FEET , BUT SHORTLY WAS BROUGHT BACK TO 3000 FEET I CONTINUED FLYING VECTORS FOR A VERY HIGH FREQUENCY OMNIDIRECTIONAL RADIO RANGE RUNWAY 23 APPROACH WHICH WAS EXECUTED WITH A LANDING AND FULL STOP . WHILE WE WERE TAXIING IN THE TOWER REQUESTED CAPTAIN NAME AND AIRCRAFT REGISTRATION NUMBER . CAPTAIN CONTACTED TOWER ON THE LANDLINE AND WAS INFORMED THAT THEY HAD RECEIVED AN ALTITUDE LOW ALERT WARNING DURING APPROACH PHASE AND HAD OBSERVED ALTITUDE EXCURSION DURING THE MISSED APPROACH PROCEDURE . FOLLOWING FAC PROCEDURES … DURING MISSED APPROACH . IT IS POSSIBLE THAT THE AMOUNT OF POWER USED , LIGHT AIRCRAFT WEIGHT , GUSTY WIND CONDITIONS AND TURBULENCE , COUPLED WITH SUCH A SHORT ALTITUDE CHANGE AIRPORT ELEVATION 815 FEET , CONTRIBUTED TO MY ALTITUDE EXCursion , AS WELL AS THE REQUEST FOR A LOWER THAN PUBLISHED MISSED APPROACH ALTITUDE . MSA ALL QUADRANTS WAS 2900 FEET GREATER VIGILANCE ON ALTITUDE CONTROL BY MYSELF AND CLEARANCE TO THE PUBLISHED MISSED APPROACH ALTITUDE WOULD HAVE PREVENTED THIS UNHAPPY EVENT . NO OTHER AIRCRAFT WERE INVOLVED . WE WERE NUMBER 1 FOR THAT APPROACH . THE AIRCRAFT BEHIND US ELECTED TO TAKE APPROACH TO RUNWAY 23 . A WEATHER SHIFT OR SQUALL APPEARED TO CROSSOVER THE FIELD JUST PRIOR TO OUR APPROACH AND LANDING , FURTHER COMPLICATING BOTH OUR PROCEDURES AND EVENTS .

- Altitude deviation: overshoot
- Non adherence: published procedure
- Conflict: ground less severe
- Inflight encounter: weather
pseudo-LSC Overview

pseudo-LSC - pseudo-Label Based Subspace Clustering
Objective

- Classify documents belonging to multiple class labels at the same time.
- Binary relevance transformation ignores class labels relationship and differences in documents of varying class label combinations.
- Example: document $x_1$ with class labels $t_1$, $t_2$ and $t_3$ should be treated differently than another document $x_2$ with class labels $t_1$, $t_2$
Motivation

- Overlapping features among multiple classes. Hard to find class differentiating features
  - Use subspace clustering to assign weights to features based on performance
- Multi-label classification - single data point may belong to multiple classes
  - During clustering, each data point should belong to all clusters with different weights - Apply fuzzy clustering
Motivation

- Features of text data are sparse. So, clusters may form in a few dimensions during subspace clustering
  - Use Chi-Square-Statistic to allow more features to participate in clustering
- Clusters contain data points of more than one class
  - Minimize Impurity Measure within cluster
- Labels are correlated with one another
  - Consider label combinations in addition to the individual labels
### Table: Construction of Pseudo Labels In pseudo-LSC

<table>
<thead>
<tr>
<th>Data</th>
<th>Pseudo Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$p_1$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$p_2$</td>
</tr>
<tr>
<td>$x_3$</td>
<td>$p_3$</td>
</tr>
<tr>
<td>$x_4$</td>
<td>$p_1$</td>
</tr>
<tr>
<td>$x_5$</td>
<td>$p_3$</td>
</tr>
</tbody>
</table>
pseudo-LSC (pseudo-Label Subspace Clustering) is a \textit{transformation based approach} for multi-label classification:

- Uses pseudo-labels or meta-label creation to transform multi-label problem to multi-class problem
- Considers the correlation among the original classes (indirectly)
- Has all the benefits of SISC-MC (\textit{Semi-supervised Subspace Clustering - Multi-Class})(Ahmed et al., DDDM (ICDM 2009))
- Different from any multi-label to binary transformation approaches, avoiding common scenario of a data point belonging to both positive and negative classes.
Related Work

Meta-label generation approaches

- MetaLabeler (Tang et al, 2009)
  - MetaLabeler learns $|T| + 1$ classification models.
  - 1 multi-class classifier for predicting the number of class labels for a test data point
  - $|T|$ binary classifiers for the $|T|$ classes

- Ensemble of Pruned Set (Reed et al, 2008)
  - Selects frequent sets or subsets of class-labels that occur together
  - Data points may be duplicated during the label-subset generation.

- Both approaches do not use semi-supervised subspace clustering.
Outline

1. Introduction

2. Subspace Clustering
   - Objective Function
   - Components of Subspace Clustering
   - Steps

3. Experimental Setup
   - Data Sets
   - Baseline Approaches
   - Results

4. Conclusions
Objective Function

- $n$ data points, each having $m$ features are clustered into $k$ clusters.
- Following Objective Function is minimized:

\[
F(W, Z, \Lambda) = \sum_{l=1}^{k} \sum_{j=1}^{n} \sum_{i=1}^{m} w_{lj}^f \lambda_{li}^q D_{ij} \ast (1 + Imp_l) + \gamma \sum_{l=1}^{k} \sum_{i=1}^{m} \lambda_{li}^q \chi_{li}^2
\]

- $W = \text{Membership weight matrix}$
- $Z = \text{Cluster centroid matrix}$
- $\Lambda = \text{Cluster dimension weights}$
- $f = \text{fuzziness parameter (2.0 indicates fuzzy clustering)}$
- $q = \text{power over the dimension weights}$
Figure a corresponds to multi-class, whereas Figure b and c corresponds to multi-label scenario.

- Impurity calculation is harder for Figure b and c than Figure a.
- In Figure b and c, individual class label counts are same but impurities are different.
- If meta-labels or pseudo-labels are generated, then Figure c transforms to Figure a making impurity calculation is easier.

Ahmed, Khan, Oza  
pseudo-LSC
Components of Subspace Clustering

Cluster Impurity Measure ($Imp_l$):

$$Imp_l = ADC_l \times Ent_l$$

$$ADC_l = \sum_{x_i \in L_{c_l}} DC_l(x_i, y_i)$$

$$DC_l(x_i, y_i) = |L_{c_l}| - |L_{c_l}(t)|$$

$$Ent_l = \sum_{t=1}^{T} (-p^l_t \times log(p^l_t))$$

Fuzzy Formulation:

- Use the cluster membership weights instead of point counts

Ahmed, Khan, Oza  pseudo-LSC
Components of Subspace Clustering

Chi Square Statistic

- **Traditional**:
  \[
  \chi^2_{li} = \frac{m(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}
  \]
  
  - \(a\) = number of times feature \(d_i\) occurs in cluster \(c_l\)
  - \(b\) = number of times feature \(d_i\) occurs in all clusters except \(c_l\)
  - \(c\) = number of times cluster \(c_l\) occurs without feature \(d_i\)
  - \(d\) = number of times all clusters except \(c_l\) occur without feature \(d_i\)
  - \(m\) = number of dimensions

- **Fuzzy Formulation**:
  
  \[
  a = \sum_{j=1}^{n} \sum_{d_j \in x_j} w_{lj}, \quad b = 1 - \sum_{j=1}^{n} \sum_{d_j \in x_j} w_{lj}, \quad c = \sum_{j=1}^{n} \sum_{d_j \notin x_j} w_{lj}, \quad d = 1 - \sum_{j=1}^{n} \sum_{d_j \notin x_j} w_{lj}
  \]

  \(m = \text{total number of labeled points}\)
Subspace Clustering

Clustering Steps

- **E-Step:**
  - Dimension weight and cluster memberships are updated.
  - For each cluster the sum of all dimension/feature weights equals 1.
  - Each data point is a member of every cluster, but with different weights. For a single training data point, the sum of membership weights across all clusters is 1.
Clustering Steps

- **M-Step:**
  - Cluster centroids and summary statistics are updated.
  - Contribution of a label is found by summing up the cluster membership weights of data points belonging to that label.
  - Label contribution is then normalized using the all the label contributions in the cluster.
  - Normalized label contribution is saved as summary statistics.
Classification

Classification Step

- **K-NN Formulation:**
  - Test point is assigned labels of the k-nearest clusters.
  - The distance is measured using the dimension weights of each of the clusters.
  - From the summary statistics, the probability of each label in each of the k-nearest clusters is determined.
  - The probabilities are weighted-averaged across the k-clusters.
Outline

1. Introduction

2. Subspace Clustering
   - Objective Function
   - Components of Subspace Clustering
   - Steps

3. Experimental Setup
   - Data Sets
   - Baseline Approaches
   - Results

4. Conclusions
Experimental Setup

Evaluation
- On *three* real world datasets:
  - NASA ASRS
  - Reuters
  - 20 Newsgroups

Performance Metric
- *Area under the ROC curve (AUC)*
- *(AUC calculated for each class separately and then macro-average is taken.)*
Outline

1. Introduction

2. Subspace Clustering
   - Objective Function
   - Components of Subspace Clustering
   - Steps

3. Experimental Setup
   - Data Sets
   - Baseline Approaches
   - Results

4. Conclusions
Baseline Approaches

5 Approaches

- $\kappa$ Nearest Neighbor ($\kappa$-NN)

- SCAD2

- Entropy based K-Means

- MetaLabeler

- Pruned Set
Outline

1. Introduction
2. Subspace Clustering
   - Objective Function
   - Components of Subspace Clustering
   - Steps
3. Experimental Setup
   - Data Sets
   - Baseline Approaches
   - Results
4. Conclusions
## Results

### Area Under The ROC Curve Comparison Chart For Multi-Label Classification

<table>
<thead>
<tr>
<th>Methods</th>
<th>ASRS</th>
<th>Reuters</th>
<th>20 Newsgroups</th>
</tr>
</thead>
<tbody>
<tr>
<td>pseudo-LSC</td>
<td>0.637</td>
<td>0.821</td>
<td>0.874</td>
</tr>
<tr>
<td>$\kappa$-NN</td>
<td>0.552</td>
<td>0.585</td>
<td>0.698</td>
</tr>
<tr>
<td>SCAD2</td>
<td>0.482</td>
<td>0.533</td>
<td>0.643</td>
</tr>
<tr>
<td>K-Means Entropy</td>
<td>0.47</td>
<td>0.538</td>
<td>0.657</td>
</tr>
<tr>
<td>MetaLabeler</td>
<td>0.58</td>
<td>0.762</td>
<td>0.766</td>
</tr>
<tr>
<td>Pruned Set</td>
<td>0.469</td>
<td>0.56</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Ahmed, Khan, Oza

pseudo-LSC
Results

Reuters

20 Newsgroups
Impact of Chi-Square Statistic on pseudo-LSC

<table>
<thead>
<tr>
<th>Methods</th>
<th>ASRS</th>
<th>Reuters</th>
<th>20 Newsgroups</th>
</tr>
</thead>
<tbody>
<tr>
<td>pseudo-LSC</td>
<td>0.637</td>
<td>0.821</td>
<td>0.874</td>
</tr>
<tr>
<td>pseudo-LSC Without Chi Square</td>
<td>0.455</td>
<td>0.532</td>
<td>0.582</td>
</tr>
</tbody>
</table>
Conclusions

Directions

- pseudo-LSC classifies multi-label data using *semi-supervised subspace clustering*
  - Transforms the data to multi-class using pseudo-labels
  - Utilizes both labeled and unlabeled data

- Future Goal: predict correct number of labels for multi-label test data point instead of merging the predictions.