CS6604 Project
Ensemble Classification

Project Team:
Kannan, Vijayasarathy
Soundarapandian, Manikandan
Alabdulhadi, Mohammed
Hamid, Tania

Project Client:
Yinlin Chen

VT, Blacksburg
03/06/2014
Introduction

• Project Objective:
  ▫ Developing classifiers to aid in Transfer Learning and classify educational resources for the Ensemble portal.

• Machine Learning (Text Classification)
The Big Picture

ACM METADATA -> (1) INPUT -> SOLR SERVER

RAW TRAINING DATA -> (3) ACM CCS 2012 -> PYTHON SCRIPT

RAW TRAINING DATA (FEATURES) -> (4) INPUT

ARFF FILES -> (5) FORMATTED

TRAINING DATA (FEATURES) -> (6) INPUT

WEKA

TEXT DIRECTORY LOADER

FILTERS

CLASSIFIER ALGORITHMS

TRAINED CLASSIFIER -> (7) PREPROCESS

TRAINED CLASSIFIER -> (8) INPUT

(9) BUILD MODEL

(10) TRANSFER LEARNING

TARGET DATA
## Results – All Classes

<table>
<thead>
<tr>
<th>Instance Size</th>
<th>No. of Classes</th>
<th>Filter</th>
<th>Classification Algorithm</th>
<th>% of Accuracy</th>
<th>Test Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>26695</td>
<td>54</td>
<td>String to Word Vector, SMOTE, Randomize</td>
<td>Naïve Bayes Multinomial</td>
<td>40</td>
<td>Cross-validation (3 Folds)</td>
</tr>
<tr>
<td>26695</td>
<td>54</td>
<td>String to Word Vector, SMOTE, Randomize</td>
<td>Naïve Bayes Multinomial</td>
<td>52</td>
<td>Use Training Set</td>
</tr>
<tr>
<td>26695</td>
<td>54</td>
<td>String to Word Vector, SMOTE, Randomize</td>
<td>J48</td>
<td>39</td>
<td>Cross-validation (3 Folds)</td>
</tr>
<tr>
<td>26695</td>
<td>54</td>
<td>String to Word Vector, SMOTE, Randomize</td>
<td>J48</td>
<td>67.55</td>
<td>Use Training Set</td>
</tr>
</tbody>
</table>
## Results – Reduced Classes

<table>
<thead>
<tr>
<th>Instance Size</th>
<th>No. of Classes</th>
<th>Filter</th>
<th>Classification Algorithm</th>
<th>% of Accuracy</th>
<th>Test Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>10002</td>
<td>10</td>
<td>String to Word Vector</td>
<td>Naïve Bayes Multinomial</td>
<td>75.8</td>
<td>Cross-validation (3 Folds)</td>
</tr>
<tr>
<td>12003</td>
<td>12</td>
<td>String to Word Vector</td>
<td>Naïve Bayes Multinomial</td>
<td>67.2</td>
<td>Cross-validation (3 Folds)</td>
</tr>
<tr>
<td>10002</td>
<td>10</td>
<td>String to Word Vector</td>
<td>SMO</td>
<td>76.8</td>
<td>Cross-validation (3 Folds)</td>
</tr>
<tr>
<td>12003</td>
<td>12</td>
<td>String to Word Vector</td>
<td>SMO</td>
<td>65.66</td>
<td>Cross-validation (3 Folds)</td>
</tr>
</tbody>
</table>
Future Work

• Classifier Accuracy improvement

• Adding more features
  ▫ Conference name
  ▫ Author Name
  ▫ Bibliographic references

• Include all classes of ACM CCS

• Single-Class Classifiers

• Transfer Learning to Ensemble portal
Challenges

- Size of the training data set
- Data Filtering and Preprocessing
  - Pruning the taxonomy
  - Classifier Accuracy
  - Weka Performance and Reliability
Questions ?