Multi-label Classification for Web Services
Mandar Badave (mrb5960@g.rit.edu)
Advisor: Dr. Xumin Liu

INTRODUCTION

- **Motivation:** Web services can belong to more than one category and it is necessary to categorize them according to their functionalities. This will help in making web service discovery and replacement more efficient.
- **Background:** Multi-label classification is still being explored in the domain of web services and there are several interesting strategies that could be advantageous.
- **Goal:** To perform multi-label classification on the web services dataset using different combinations of algorithms and feature sets and compare their performances.

MULTI-LABEL CLASSIFICATION TYPES

- **Problem Transformation (PT):** Transforms a multi-label classification problem into several single label classification problems which use a base classifier. We have used Decision Trees (DT), Naïve Bayes (NB), Logistic Regression (LR) and Support Vector Machine (SVM).
- **Binary Relevance (BR):** A classifier is trained for every label which is used for predicting that label for every test instance.
- **Classifier Chain (CC):** Similar to BR but once a label is predicted, it is used as a feature to train the next classifier.
- **Label Powerset (LP):** Transforms the multi-label classification problem to multi-class classification problem.
- **Algorithm Adaptation (AA):** Classification algorithms are modified to handle multi-label data.

DATA FORMATTING

<table>
<thead>
<tr>
<th>Name</th>
<th>Desc</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>desc 1</td>
<td>X, Y</td>
</tr>
<tr>
<td>B</td>
<td>desc 2</td>
<td>Y, Z</td>
</tr>
<tr>
<td>C</td>
<td>desc 3</td>
<td>X, Y, Z</td>
</tr>
</tbody>
</table>

IMPLEMENTATION

- **Data Collection:** Scraped [www.programmableweb.com](http://www.programmableweb.com). The dataset has 16441 web services and 477 unique labels.
- **Data Preprocessing:** Cleaning the data, converting the data to proper format, selecting web services having top 20 most frequent labels.
- **Feature Extraction:** Generating the TF-IDF matrix, performing POS tagging, training Word2Vec model or using a pre-trained Word2Vec model to generate the document vectors.
- **Classification:** Multi-label classification models are trained using all the feature sets separately and tested using 10-fold cross validation.
- **Comparison:** F1 score is used as the evaluation metric and the performance of the models are compared using the F1 scores.

RESULTS

- Google D2V performs better than D2V.
- Weighted D2V may perform better if descriptions are larger.
- LP is the most consistent multi-label classification algorithm.
- SVM performs better than other base classifiers irrespective of the multi-label classifier used in PT.

CONCLUSIONS

- The performance of the multi-label classification algorithms is better when TF-IDF is used as the feature set.
- The performance of Problem Transformation approaches largely depends on the type of base algorithm used.
- Multi-label classification not only help in assigning labels to web services but also in identifying incorrectly assigned labels.