Classification of Email Spam using Concept Learning(Group 53)

VAMSI SRI NAGA MANIKANTA MURUKONDA
20010620-T114
vamu21@student.bth.se

SREE LAKSHMI HIRANMAYEE KADALI
20010920-T244
srkd21@student.bth.se

I. Introduction

Based on concept learning, algorithms 4.1 and 4.3 from the main literature is implemented for this project work. We use both the algorithms to create a conjuctive least general generalization algorithm (LGG) to classify spam base dataset to distinguish between spam and non-spam mails by returning the most general version space.

II. Preprocessing of the data

The dataset given consists of 4601 instances with 57 attributes, with the last attribute being the grouping of mail as spam(1) and non-spam(0). The dataset is classified into two sets ”train_data” and ”test_data” using the sklearn and pandas library. The instances are read into one dataset ”train_data” and classification values (1-spam, 0-non spam) is read into other dataset ”test_data”. The datasets are then transformed into numpy arrays which are then transformed to train and test sets with the ratio of 70:30 where, 70% of the spam mails are used to train and 30% is used to test.Since, the data is discretized seperately for training and testing, fewer bins are used to provide greater accuracy, and n=10 bins are used. Thus KbinsDiscretizer is used. Later the datasets are mixed with non-spam emails. A hypothesis space is then created and data is calculated into the hypothesis space.

III. Algorithm selection and implementation

A combination of algorithms 4.1 and 4.3 is used to give better accuracy for the classification of data. The LGG algorithm is employed as soon as each occurence in the feature set is set up in the hypothesis space. Each one of the algorithms, LGG's internal disjunction and LGG's conjunction recieve occurences one by one and examine the hypothesis space for its particular attribute set. If the hypothesis attribute set, as of now doesn't contain the attribute, the attribute is added and the iteration is repeated until the last instance. This modified hypothesis space return the required hypothesis space.

IV. Training

For the training purpose, 70% of the dataset with spam mails is supplied to the LGG algorithm with LGG conjugate inner disjunction. The implementation builts a parameter space which contains lists of internal disjunction sets for each feature.

V. Testing

The test dataset which contains 30% of the spam messages and rest as non spam messages is provided for the algorithm for testing. Each occurence is assessed at a time to confirm it is inside the hypothesis space.The occasion set is classified according with the hypothesis space as spam and non-spam.

VI. Hypothesis Space

The calculated hypothesis space as defined in the main literature section 4.1 is with the number of possible extensions i.e set of instances 2*(10*57). The possible instances with a conjugate concept includes (10+1)*57 including the absence of feature as an addition 'value'.

The hypothesis space generated is as follows with space size ”64”: [[0.0, 1.0], [1.0, 0.0, 4.0, 3.0, 2.0], [0.0], [0.0, 1.0, 2.0, 3.0], [0.0, 2.0, 1.0], [0.0, 1.0], [0.0, 1.0], [1.0, 0.0], [0.0, 2.0, 1.0], [0.0, 1.0], [2.0, 0.0, 3.0, 1.0, 5.0, 4.0], [0.0, 1.0], [0.0], [0.0], [0.0, 1.0, 2.0], [0.0, 2.0, 1.0], [0.0, 2.0, 1.0], [6.0, 3.0, 0.0, 7.0, 2.0, 4.0, 5.0, 1.0], [0.0], [4.0, 3.0, 5.0, 0.0, 1.0, 2.0], [0.0], [0.0, 1.0], [0.0, 1.0], [0.0, 2.0, 1.0], [0.0, 1.0], [0.0], [0.0], [0.0], [0.0, 1.0], [0.0], [0.0], [0.0], [0.0], [0.0, 1.0], [0.0, 1.0], [0.0, 1.0], [0.0], [0.0], [0.0], [0.0], [0.0], [0.0], [0.0], [0.0, 1.0, 2.0], [0.0, 1.0], [0.0], [0.0], [0.0, 1.0], [1.0, 0.0, 3.0, 2.0, 4.0, 5.0], [0.0, 1.0], [0.0, 1.0, 4.0, 3.0, 2.0], [1.0, 2.0, 0.0, 3.0], [0.0, 1.0], [6.0, 7.0, 9.0, 1.0, 8.0, 4.0, 3.0, 5.0, 2.0, 0.0], [8.0, 9.0, 3.0, 4.0, 6.0, 7.0, 5.0, 2.0, 1.0, 0.0], [7.0, 6.0, 4.0, 3.0, 5.0, 8.0, 9.0, 2.0, 1.0, 0.0], [0.0]]

VII. Results

The accuracy of the model is given by (tp+tn)/(tp+tn+fn+fp) The accuracy of the model is 0.52.