DV2578-Assignment-1

Patamsetti Naga Vishnu
Chakravarthi

*Department of Computer
Science Blekinge Institute of
Technology* Karlskrona, Sweden.
napa18@student.bth.se

I. Introduction

To implement the concept learning, detection of spam emails from email collection is the main of the report. For that concept learning algorithms have been used to create hypothesis space. This contains all the positive instance features from training data. This is used for testing. And later, data information and implementation results will be presented.

II. Data

Here, the data is collected from the UCI Machine learning repository. This consists of instances in the form of rows and attributes in columns. The values in it are continuous and assigned to each row. It is in the form of 1's and 0's of the 58th column.

III. Methodology

Here by using the cut function in pandas, continuous data is converted to data description.

The continuous data is converted to discrete data by using the cut function in pandas. So here 6 bins are considered for data discretization. After the completion of data discretization, segregated into training and testing purposes. Here 70 percent of data is considered for training. The remaining 30 percent of data is considered for testing.

A. Algorithms

Algorithms 4.1 and 4.3 considered from the concept learning. From the data each of the instance is considered in finding out the concepts of generalization and conjuctions from the data that are added to hypothesis. By using of algorithm 4.3 conjuctions is gathered to add all the unique features by using of internal disjunction. This is done in the hypothesis space considering the reference [1]. More accuracy can be tested in hypothesis space by addition of all features. The data is trained in such a way that only the values from features which are not added in the hypothesis before, are added and is followed similarly to all the other features.

IV. Results

With training data of rows that consists of positive instances with the length of 57 features the hypothesis is formed. After the completion of the implementation among algorithms 4.1 and 4.3. As 6 bins are considered for data discretization the total number of all possible connective concepts is 7^{57} and the size of hypothesis space is $2^{6^{57}}$ Here accuracy is calculated by using of formula (tp+tn)/(tp+tn+fp+fn).

The Hypothesis is in the form of Ho, H1, H2,H3. H57,
Ho= (1 or 2 or 3 or 5 or 8) similarly to all features up to H56.

(1 ∨ 3 ∨ 2 ∨ 6 ∨ 4) ^ (1 ∨ 2) ^ (1 ∨ 2 ∨ 3 ∨ 4 ∨ 5) ^ (1 ∨ 2 ∨ 6 ∨ 3 ∨ 5) ^ (1 ∨ 2 ∨ 4 ∨ 3) ^ (1 ∨ 2 ∨ 3) ^ (1 ∨ 2 ∨ 3 ∨ 5 ∨ 6 ∨ 4) ^ (1 ∨ 2 ∨ 3 ∨ 6) ^ (1 ∨ 2 ∨ 3) ^ (1 ∨ 2 ∨ 3) ^ (1 ∨ 2 ∨ 4 ∨ 3) ^ (2 ∨ 1 ∨ 3 ∨ 4) ^ (1 ∨ 2 ∨ 6 ∨ 3) ^ (1 ∨ 2 ∨ 3) ^ (1 ∨ 4 ∨ 2 ∨ 3 ∨ 6) ^ (1 ∨ 2 ∨ 3 ∨ 4) ^ (1 ∨ 2 ∨ 3 ∨ 5 ∨ 4 ∨ 6) ^ (1 ∨ 4 ∨ 2 ∨ 5 ∨ 3 ∨ 6) ^ (2 ∨ 1 ∨ 3 ∨ 4) ^ (1 ∨ 2 ∨ 6 ∨ 3) ^ (1 ∨ 2 ∨ 3 ∨ 4 ∨ 6) ^ (1 ∨ 4 ∨ 3 ∨ 2 ∨ 6 ∨ 5) ^ (1 ∨ 2 ∨ 3 ∨ 6 ∨ 5 ∨ 4) ^ (1 ∨ 3 ∨ 5 ∨ 2 ∨ 4 ∨ 6) ^ (1 ∨ 2) ^ (1) ^ (1) ^ (1 ∨ 2 ∨ 6) ^ (1) ^ (1 ∨ 3) ^ (1) ^ (1) ^ (1) ^ (1) ^ (1) ^ (1 ∨ 2) ^ (1 ∨ 2 ∨ 3 ∨ 5) ^ (1) ^ (1 ∨ 2) ^ (1 ∨ 2 ∨ 3) ^ (1) ^ (1) ^ (1 ∨ 2) ^ (1) ^ (1 ∨ 2) ^ (1) ^ (1 ∨ 2) ^ (1) ^ (1 ∨ 2) ^ (1 ∨ 2 ∨ 3) ^ (1 ∨ 2) ^ (1 ∨ 2) ^ (1 ∨ 2 ∨ 6 ∨ 4 ∨ 5) ^ (1 ∨ 6 ∨ 2) ^ (1 ∨ 6 ∨ 4 ∨ 2 ∨ 3) ^ (1 ∨ 2) ^ (1 ∨ 2 ∨ 6 ∨ 4)

V. Conclusion:

From the Hypothesis, the generated conjunctive rule LGG-Conj. If all the conditions are satisfied when tested it with hypothesis, it is detected as spam, or else not as spam. The accuracy obtained by testing the hypothesis on the testing data is 57. 3099.

VI. REFERENCES

[1] P. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012.