Assignment 1 Report

Tanveer Feisal Snigdho
MSc in Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
t.f.snigdho@gmail.com

I. Introduction

This report contains a brief description of assignment 1, the objectives, dataset and the procedures.

II. Objective

The objectives of the assigment are as following:

- To find hypothesis space
- To find Number of conjunctive concepts
- To find conjunctive rules for spams
- To test if those conjunctive rules can identify spams or not.

III. Dataset

The spambase dataset was used for this assignment. The source of the dataset is available in the assignment instructions manual [1]. According to this dataset, each row contains data of one email. The last column indicates if it is a spam or not. The other 57 columns represents different value of specific properties.

IV. Procedures

To Find Hypothesis Space

According to the discretization criteria, each row has only 2 possible values: 1 or 2. So the hypothesis space is: 2^57.

To Find Number of conjunctive concepts

According to the discretization criteria, each row has only 2 possible values: 1 or 2. But for cconjunctive concepts, we need to consider one more option which is data not available. So there could be 3 possible values: 1, 2 or unknown. So number od conjunctive concepts is: 3^57.

To find conjunctive rules for spams

This procedure is divided into several parts. It starts with finding the the number of spams in this dataset. Then I calculated total number of spams and the 80 percent of it. I used the first 80 percent of spams for training and the rest 20 percent for testing.

Then for each of the 57 datasets, I calculated the minimum, maximum, average and standard deviation. Here, there is a special criteria which I used for calculating average and standard deviation. It could be explained as following:

Let's consider the values: 0, 2.2, 2.7, 0, 1.6, 0, 1.3, 0, 1.9. There are 4 0s. I did not consider those 0s. After eliminating those 0s, we have 2.2, 2.7, 1.6, 1.3 and 1.9. Now, n = 5. There is a reason for eliminating those 0s. In the given dataset, 0 means absence of some words or symbols. When I tried to find and average, then the average becomes very close to 0. So I did not consider 0s while calculating average and standard deviation.

From line 18 to 66 of codes were written for finding average, minimum and maximum. From line 70 to 87 of codes were written for finding the standard deviation.

Then I added the average with 3 times of standard deviation which is: value_limit = average + (3 * standard deviation). The discretization criteria is as following:

```
if (minimum ≤ x ≤ value_limit ) :
    return 1
else:
    return 2
```

Codes from line 106 to 130 were written for algorithm 4.1 [2] and codes form line 124 to 127 were written for algorithm 4.2 [2]. For the first dataset, I saved all the discretized vaules in temp_values and set all the values_flags to 1. Then for all the next datasets (within 80 percent), I compared the discretized values and whenever I found a value is not equal, then I set values_flag to 0 which means that concept has been eliminated. After comparing all the training dataset (80 percent of spams), I checked where the values_flags are still 1. Those were the selected concepts.

Not to mention, 80 percent of spams were used for training. The rest 20 percent of spams were used for testing. This program was able to successfully detect 354 spams out of 363 and the success rate was 97.5 percent.

When I found all the concepts, then I read the spambase.names file and printed out the names of the concepts from that file. Codes from line 132 to 151 were written for printing the concepts.

Codes from line 154 to 181 were written for testing the rest 20 percent of spams. Codes 183 to 218 were written to test the whole dataset and the overall accuracy was found 63.2 percent.

V. References

[1] V. Boeva, "ML-Instructions."

[2] P. A. Flach, *Machine learning: the art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press, 2012.