Data Mining Assignment 4

Ankita Kulkarni

May 8, 2016

I declare that all material in this assessment task is my own work except where there is clear acknowledgement or reference to the work of others and I have complied and agreed to the UIS Academic Integrity Policy at the University website URL: http://www.uis.edu/academicintegrity

Student Name: Ankita Kulkarni UID: 672369586 Date: May 8, 2016

Introduction

In the PRISM algorithm a rule is first identified. After the identification of the rule the instances that lie in that rule are separated from the rest of the rules. The instances that are left out are then worked upon. Once an instance has been brought under a particular rule then there is no further work required on it. Support based pruning is not provided in prism. While classifying data with PRISM classifier there might be a number of instances which are left unclassified. These instances are generally the outliers.

Classification of weather dataset using PRISM

- I experimented the PRISM algorithm on the weather dataset. In order to perform the experiment, I implemented the PRISM classifier on the weather dataset for the following three test cases: Use training set, Cross-Validation and Percentage Split. When the test case 'Use training set' was used, an accuracy of 100

S. No	No. of Folds	Correctly classified instances (%)	Unclassified instances out of total 14 instances
1	5	78.5714	1
2	10	64.2857	2
3	13	57.1429	3

As mentioned earlier the PRISM classifier leaves some instances as unclassified, the instances being outliers. I could observe that the performance of the PRISM classifier increased with the decrease in the number of folds. There was also a decrease in the number of unclassified instances. For the test case Percentage split there were some different observations. Below are the observations for the test case percentage split:

S. No	Split Percentage (%)	Correctly classified instances (%)	Unclassified instances out of total 14 instances
1	33	55.5556	2 out of 9
2	50	71.4286	0 out of 7
3	66	60	0 out of 9

The best results were observed when the dataset was split 50 percent where the percent of correctly classified instances was approximately 71 percent with no instance left as unclassified. Amongst the two test cases, Cross Validation performed better than Percentage Split.

Classification of breast - cancer dataset using PRISM

- The heuristic I used to deal with the missing values in the breast-cancer dataset in the previous assignment was: Set the missing values by comparing the rows. The row which has missing value was compared with all the other rows. The missing value was then set to a value of the row which was maximum similar to the missing value row. In the previous assignment I implemented my heuristic in 2 runs. First by checking for the similarity of 5 attributes and then for

the left over missing values similarity of 3 attributes was checked. However I modified this a bit in this experiment. This time I implemented the heuristic in two steps, in the first step I compared the rows where at max 5 attributes are similar. In the second step I set the set of the missing values to the value which has the maximum frequency. I then tested this breast-cancer dataset which had no missing values with the PRISM classifier. To carry out mu experiment I tested 2 types of breast-cancer dataset with the PRISM classifier. First, the dataset which I got after implementing my heuristic. And second, the dataset which I got after implementing the 'ReplaceMissingValue' filter on the original breast-cancer dataset. Below were the observations noticed:

S. No.	Breast-Cancer dataset (ReplaceMissingValue filter)			Breast-Cancer dataset (heuristic implemented)		
	No. of folds	Unclassified instances (out of 286)	Correctly classified instances (%)	No. of folds	Unclassified instances (out of 286)	Correctly classified instances (%)
1	10	39	58.7413	10	42	54.5455
2	15	29	60.8392	15	33	56.2937
3	20	36	58.3916	20	36	56.2937
4	25	27	60.8392	25	33	55.9441
5	30	34	59.7902	30	31	56.2937

There was no trend noticed for the correctly classified instances with respect to the number of folds. However the best results were observed when the 'ReplaceMissingValue' filter was applied for 15 and 25 folds.

Classification of soybean dataset using PRISM

- To explore the ability of the PRISM classifier to generalize, a series of tests were carried out on the soybean dataset. The soybean dataset was divided into training sets and test sets. However before dividing the data the missing values in the soybean dataset were handled as the PRISM classifier does not work for missing values. To handle the missing values, 'ReplaceMissingValue' filter was used. Once the dataset with no missing values was obtained, it was

then divided into different percent of test and training data. The experiment was carried out for the percentages 60-40, 70-30 and 80-20, where 60, 70 and 80 being the training percentage and 40, 30 and 20 being the test percentage of the total dataset. The results produced by the PRISM classifier were compared to those of ID3. Below were the observations for the PRISM and ID3 classifier:

S. No	Test Percent	Training Percent	PRISM Classifier		ID3 Classifier	
			Unclassified Instances	Correctly Classified Instances (%)	Unclassified Instances	Correctly Classified Instances (%)
1	40	60	38	77.6557	6	86.0806
2	30	70	14	86.3415	1	89.2683
3	20	80	8	92.7007	1	91.9708

I could clearly observe that there was not much difference between the performance of ID3 and PRISM classifier. However even if there was a minute difference, the ID3 classifier proved to be better as compared to PRISM. Also the ID3 classifier classifies all the instances unlike PRISM which leaves some instances as unclassified which are outliers and does not lie under some particular defined rule. The result produced by the 60-40