Data Mining Assignment 2

Ankita Kulkarni and 672369586

March 12, 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: akulk2 Date: March 12, 2016

Abstract:

Weka helps implementing the machine learning algorithms which classifies the data. It contains some predefined classifier. It is basically a computation learning. The instances in the data are classified. The classification can be correct as well as incorrect. Weka helps in classifying the

Introduction

The aim of this report is to classify the Iris and the Glass dataset for the algorithms IB1, IB2 and IB3. The classifier IB1 utilizes a basic separation measure to discover the preparation case nearest to the given test example, and predicts the same class as this training instance. It compares the instance with its nearest neighbour and hence decides its class to which it belongs. However the classifier IB2 compares the instances to 2 nearest neighbour. I further improves the performance of the IB1 algorithm. In IB2 algorithm only the instances which are important are stored, it thus improves and reduces the storage requirement. The IB3 algorithm compares the instances with the 3 nearest neighbour. It further improves the performance of the IB2 algorithm by handling the noisy data. I tested the Iris and Glass dataset for these 3 algorithms.

Methodology

1. I tested the Iris dataset with the IB1, IB2, and IB3 classifier for the test case cross validation and percentage split. To my observation there was not a major difference in the accuracy of the classifier. While running the cross validation test case for the values 5,7,10, and 15, I could observe that the correctly classified instances for all the three algorithms was fluctuating between 94% and 95%. To my thought this is because the iris dataset is a small one. For the IB1 algorithm the value for the folds 10, 5 and 7 was 95.3333%, 94% and 94.6667% respectively. When I tested for the folds more than 10, the correctly classified instances were always 95.3333%. Same was the case with the algorithm IB2. The correctly classified instances for the folds 5, 10 and 15 were 94.6667%, 94.6667% and 96 %. The value for the correctly classified instances was constant on increasing the number folds above 15. For the IB3 algorithm the maximum correctly classified instances was observed for 20 folds which was 96%. And again its value was constant on increasing the number of folds. On using the percentage split test case the correctly classified instances always remained constant and did not change with change in the percentage.

However while classifying the data for the baseline algorithm that is the ZeroR algorithm the correctly classified instances were very less. For 10 folds the correctly classified instances were 33.3333% and the value kept on decreasing on increasing the number of folds.

2. On testing the glass dataset I could notice that the algorithm IB2 performed badly as compared to the algorithms IB1 and IB3. However the correctly classified instances for the IB1 and IB3 algorithm were nearly same but the IB3 algorithm performed the best with the correctly classified instances of 71.9626% for 10 folds. This value decreased when the number of folds were increased. The algorithm IB2 could produce the correctly classified instances maximum to 67.757% which was for 10 folds. The algorithm IB1 could correctly classify instances to 70.5607% which was for 10 folds again. The performance of IB2 was bad for the glass data set because it contained noise which was removed by the IB3 algorithm.
   I also tested the glass dataset for the baseline or the ZeroR algorithm and it performed badly for the glass dataset. It could correctly classify the instance to a percent of 35.514. The correctly classified instances in all the three algorithms were nearly same and this is because the dataset is very small.

3. The next implementation that I performed was to create an .arff file for the LED program and classify that dataset with the classifiers J48, IB1, IB2 and IB3. These classifiers were run for the dataset by implementing a Remove Percentage filter for 80% training and remaining 20% as test. The testing was done by initializing the invertSection parameter to both true and false one by one. The .arff file was created by introducing a noise. This was done for the 10 trials where the introduced noise was 1, 2, 3, 4, 5, 6, 7, 8, 9, and 15. Below are the maximum value for the correctly classified instance that were observed when the invertSection parameter was initialized to “false” that is testing for 80% training data and the cross validation test case was implemented.

S. No	Noise % introduced	J48 Correctly Classified Instances	IB1 Correctly Classified Instances	IB2 Correctly Classified Instances	IB3 Correctly Classified Instances
1	1	97.8	99.5	99	99
2	2	94.9	96.4	96.7	96.8
3	3	90.9	90.7	90.4	92.9
4	4	87.8	85.3	86.2	86.4
5	5	83.7	82.6	82.5	83.4
6	6	80.5	77.8	75.5	78.5
7	7	76.6	72.7	74.7	74.9
8	8	71.5	67.8	71.2	72.3
9	9	68.8	60.2	69.6	69.9
10	15	55.7	50.5	59.6	58

When testing the LED data set for the classifiers it was observed that as the noise percentage increased, there was a decrease in the percentage of the correctly classified instances. Also the classifier IB3 performed the best amongst the 4 classifiers. For the same noise percent introduced, the dataset was also tested keeping the invertSection parameter in the filter to "true" that is testing for 20% test data. Below are the observations.

S. No	Noise Introduced	J48 Correctly Classified Instances	IB1 Correctly Classified Instances	IB2 Correctly Classified Instances	IB3 Correctly Classified Instances
1	1	96.5	95.1	96.7	97
2	2	92.9	92.1	94.6	94.3
3	3	91.9	88.9	92.4	92.5
4	4	90.8	86.3	90.4	90.6
5	5	88.7	82.7	89.5	88.9
6	6	86.5	80.4	87.3	86
7	7	85.6	79	85.6	85.2
8	8	80	74.3	78.5	79.5
9	9	76.8	69.8	73.9	72.8
10	15	60.5	47.25	59.3	59.6

After classifying the data by dividing it into test data and training data for the classifiers J48, IB1, IB2, and IB3, I could clearly observe that the classifier IB3 performs the best amongst the 4.

Result

After performing the above research and testing the performance of the classifiers J48, IB1, IB2 and IB3 for the data sets Iris, Glass and LED it was clearly observe that the classifier IB3 performs the best for all these data set.

Conclusion

This research clearly displayed that the performance of the classifiers also depends on the size of the data set as well as the type and number of attributes. Also there is a difference in the correctly classified instances when the data is dived into training data and test data. In my research I have divided the data into 80-20%, where 80% is the training data and 20% is the test data.

References:

1. http://cs.uccs.edu/~jkalita/work/cs586/2013/InstanceBasedLearning.pdf
2. Practical machine learning tools and techniques (3rd edition)
3. https://www.acs.org.au/__data/assets/pdf_file/0016/15550/JRPIT37.2.211.pdf