Data Mining Assignment 2 B

Abstract:

The main objective of this assignment is to implement instance based learning algorithms IB1, IBK with possible values for N equals to 2,3. While, performing all these test we have to find out which algorithm performs better on given dataset and why.

Introduction:

Instance-based learning works in a similar manner to that of nearest neighbor algorithm except it process instances incrementally and normalizes the attribute ranges. However, the major drawback with IB1 is it stores every instances it sees. To solve this issue, IB2 was developed. IB2 only store those instances which are near to boundary line. As per IB2 implementation only instances which lies near boundary are enough to produce an accurate approximation. It is based on the approach that, instances which lies away from boundary does not contribute much in correct classification. Even though this approach saves the memory storage space but it tends to accumulate noisy instances, which leads in reducing an accuracy of correctly classified instances. IB3 algorithm only collects those instances which tends to performed well

during classification. It keeps a record of instances that performs well and applies significance test to determine noisy instances. This records not only stores instances which performs well but also predicts their future performance[2][5].

1 IB2:

1.1 Results:

1.1.1 Iris Data:

Table 1: Accuracy comparison of IB1 and IBK with K=2 on the iris data

Dataset	IB1	IBK (with K=2)
iris	95.33%	94.66%

1.1.2 LED data:

Table 2: Accuracy comparison of IBK with 2NN and J48 on LED data

LED Data			
Noise Percent	IB1	IBK (K=2)	J48
3	64.1%	72.6%	91.8%
4	61%	70.2%	88.4%
5	56.9%	66.3%	84.7%
6	54.7%	63.9%	81.7%
7	53%	61%	80.2%
8	49.9%	57.3%	76.1%
9	47.1%	54.4%	72.7%
10	44.5%	50.9%	71.3%
11	41.2%	48.6%	67.5%
12	38.7%	46.3%	65.8%
Average	51.11	59.15	78.02
Standard Deviation	8.36	9.13	8.80

2 IB3:

2.1 RemovePercentage:

This filter removes the given percentage of dataset. If I simply used entire dataset for test purpose then it is very easy to achieve perfect accuracy by simply learning an entire dataset. The nearest neighbor classifier IBK exactly does the same. Instead when we divide the dataset into training and testing then, it is possible to get new unseen instances and find the correct accuracy. In this case, I am dealing with just one data set which is LED data. So I have divided it into two parts i.e training and test part. The simplest way of doing it is to use RemovePercentage filter [4]. I have divided training data with size equals to 200 (by applying the filter RemovePercentage and setting the percentage field to 80%) for training data and in the similar manner for test data with size equals to 500(set percentage to 50%). I chose 500 and 200 because, Aha used the same number for testing the LED data.

2.2 Results:

2.2.1 LED data:

Table 3: Accuracy comparison of IB1, IBK with 2NN and J48 on LED data

LED Training and Test Data					
Noise Percentage	Dataset	IB1	IBK with K=2	IBK with K=3	J48
3	Test	61.6%	68%	72.2%	91.8%
	Training	61%	66.5%	68%	89.5%
4	Test	57.6%	64.4%	68%	88.2%
	Training	57%	60.5%	62.5%	87%
5	Test	54.6%	62.2%	64.8%	85.2%
	Training	53.5%	58%	60.5%	86%
6	Test	51.2%	59.8%	61.8%	82.8%
	Training	51%	58.5%	62.5%	83.5%
7	Test	48.6%	56.6%	58.8%	79.8%
	Training	48%	55%	59.5%	78.5%
8	Test	45.6%	53.8%	56%	78.2%
	Training	46%	52.5%	57.5%	73.5%
9	Test	43.4%	52.8%	55.2%	75.6%
	Training	44%	48%	53.5%	71.5%
10	Test	43.4%	48.4%	51.8%	74%
	Training	43.5%	47.5%	52%	68%
11	Test	41.8%	47.2%	50%	67.6%
	Training	40.5%	43.5%	49%	65.5%
12	Test	40.2%	44.6%	48.6%	65.8%
	Training	39.5%	42%	46%	62.5%
Average	Test	48.8	55.78	58.72	78.9
	Training	48.4	53.2	57.1	76.55
Standard Deviation	Test	7.24	7.79	7.89	8.45
	Training	7.12	7.91	6.85	9.69

2.2.2 Glass Data:

Table 4: Accuracy comparison of IB1, IBK with 2NN and 3NN, J48 on Glass data

Glass Data				
Test criteria	IB1	IBK with K=2	IBK with K=3	J48
10 fold cross validation	70.56%	67.75%	71.96%	66.82%
With 70% split	60.93%	56.25%	54.68%	62.5%
With 30% split	71.33%	70.66%	70.66%	72.66%
With 80% split	55.81%	58.13%	58.13%	67.44%
With 20% split	71.92%	68.42%	70.17%	72.51%
Average	66.11	64.24	65.12	68.38
Standard Deviation	7.30	6.56	8.07	4.27

Discussions/Conclusions:

From all the experiments performed above, I have noticed that IB2's classification decreases as the noise percentage increases. However, there is not much difference with IB1's classification accuracy with increasing noise percentage. This happens because noisy instances are most of the time miss-classified. IB2 saves small number of non-noisy instances and its saved noisy instances are used for generating classification decisions. Further, IB2 significantly reduces IB1's storage requirements. While testing IB1 and IB2's accuracy on non-noisy domains such as iris.arff, I observed that their classification accuracies are almost the same (Table 1). IB2's accuracy is slightly less than that of IB1's while dealing with noise-free dataset as seen in Table 4, I believe this is mainly because, IB2 tends to save most of the noisy instances. On the other hand, IB3 has higher classification accuracy than IB1 and IB2 when dealing with noisy dataset such as LED as seen from Table 3. In the above experiments, my results are less than those were presented in Aha's report. But, I believe this is partly because I performed the experiment only on 10 trials unlike the 50 that Aha used[1][2][5].

Video Presentation:

Video presentation URL:

https://cdnapisec.kaltura.com/index.php/extwidget/preview/partner_id/1371761/uiconf_id/31473632/entry_id/1_wrj0ey9v/embed/dynamic

References:

[1] Aha, David W. 1992. Training noisy, irrelevant and novel attributes in instance-based learning algorithms. International Journal of Man-Machine studies 36(2):267-287
[2] Aha, David W., Dennis Kibler and Marc Albert. 1991. Instance-based learning algorithms. Machine Learning 6:37-66.
[3] "Data Mining: Practical Machine Learning Tool And Techniques" - Third Edition by Ian H. Witten, Eibe Frank, Mark A Hall
[4] https://weka.wikispaces.com/How+do+I+divide+a+dataset+into+training+and+test+set
[5] https://pdfs.semanticscholar.org/395b/2e9bd23e56394b61b8deab2e0cbc9718f7fd.pdf

---

Last updated: March 13, 2017
Typeset using TEXshop