Universidade do Minho
Escola de Engenharia

Mineração de Dados

Mini Teste 2 - Individual
MiEI - 4º Ano - 1º Semestre

A85308 Filipe Miguel Teixeira Freitas Guimarães

Braga,
5 de dezembro de 2020

1 Apresente uma análise ao desempenho dos algoritmos NaiveBayes, BayesNet, J48 e AdaBoost sobre NB, Bagging sobre J48 no dataset ionosphere. Considere as seguintes alienas:

1.1 Usando a decomposição bias-variance estude o evoluir da complexidade da árvore j48 (por aplicação de diferentes níveis de pruning) e o efeito no erro e suas componentes.

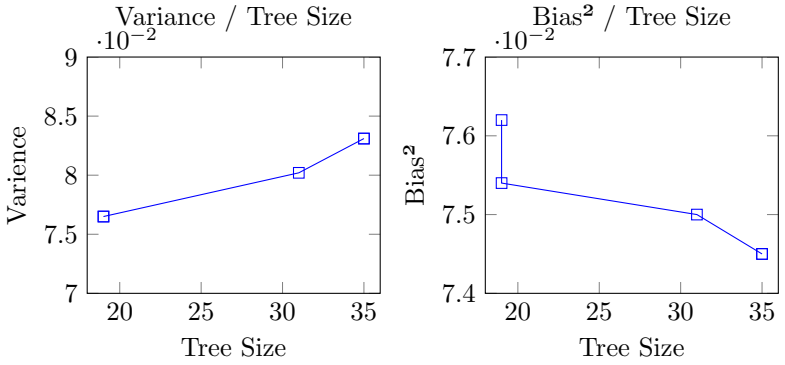
Para facilitar a análise desta árvore decidi compilar os dados numa tabela. Obtive os dados recorrendo ao comando dentro da pasta de instalação do *weka*

```
$ java -cp weka.jar weka.classifiers.BVDecompose -t data/ionosphere.arff
-W weka.classifiers.trees.J48 -- -C *Confidence Factor* -M 2
```

alterando os valores de confiança bem como o *Weka*.

Correctly Classified Instances	**Confidence Factor**	**Error**	**Variance**	**Bias²**	**Tree Size**
90.5983	0.01	0.1543	0.0765	0.0762	19
90.5983	0.1	0.1555	0.0785	0.0754	19
91.1681	0.2	0.1568	0.0802	0.075	31
91.423	0.25	0.1593	0.0831	0.0745	35
91.423	0.5	0.1593	0.0831	0.0745	35

Fiz também dois gráficos para ver o comportamento da *Variance* e *Bias²* em função do crescimento da árvore.



Com estes dados consigo verificar que, com o aumento do fator de confiança, o *Bias²* diminui enquanto a *Variance* aumenta. Com um aumento do fator de confiança consegue-se ver também um acrescimo no tamanho da àrvore.

1.2 Apresente os resultados do erro por validação cruzadas dos cinco modelos referidos. Elabore uma justificação para cada resultado obtido.

Recorrendo ao *Weka* para analisar os algoritmos obtive os seguintes resultados:

Algorithm	**Mean Absolute Error**	**Root Mean Squared Error**	**Relative Absolute Error**	**Root Relative Squared Error**
NaiveBayes	0.1736	0.3935	37.70%	82.02%
BaysNet	0.1069	0.3179	23.22%	66.26%
J48	0.0938	0.2901	20.36%	60.46%
AdaBoost sobre NB	0.1043	0.2611	22.64%	54.42%
Bagging sobre J48	0.1109	0.2431	24.08%	50.66%

Analisando a tabela verifico que o algoritmo que apresenta maior erro é o *Naive-Bayes* que é superior ao mesmo recorrendo ao *AdaBoost*.Boosting é normalmente usado para reduzir a taxa de erros e por isso era o esperado. Quanto ao J48 é um pouco diferente existe uma certa variação nos diferentes dados recorrendo ou não o Bangging, tal como era esperado por ser usado para reduzir a variabilidade dos modelos individuais.

1.3 Para a classe b e considerando os cinco modelos referidos:

- **Qual seria o melhor modelo a recuperar os casos desta classe?**
- **E o modelo com melhor qualidade de previsão nesta classe?**

Justifique as suas respostas.

Para avaliar a escolha do melhor modelo, recuperar casos e melhorar a qualidade de precisão decidi compilar os resultados de Recall, Precision e F-Measure na seguinte tabela:

Algorithm	**Precision**	**Recall**	**F-Measure**
NaiveBayes	0.712	0.865	0.781
BaysNet	0.874	0.825	0.849
J48	0.929	0.825	0.874
AdaBoost sobre NB	0.883	0.897	0.890
Bagging sobre J48	0.857	0.857	0.896

Pode-se verificar que o melhor modelo para recuperar os casos desta classe seria o Adaboost sobre *NaiveBays* já que é o que apresenta o maior Recall.
O melhor com melhor qualidade de previsão seria o Bagging sobre J48 já que tem o valor de *precision* mais elevado.

2 Que tipo de benefícios esperaria da aplicação de Bagging sobre Naive Bayes num dataset específico, sabendo que o resultado do modelo individual Naive Bayes nesse dataset é erro = 0.005. Justifique.

O Bangging faz reduzir a variância do erro do classificador. Esta técnica funciona para algoritmos que envolvem árvores de decisão, ou seja, para algoritmos instáveis. Usando esta técnica em algoritmos mais estáveis, como o Naive Bayes, pode fazer aumentar a variância, já que é um algoritmo com uma variância baixa.
Sendo este um erro muito pequeno, a utilização de *Bangging* neste caso não seria benéfico, pode ainda aumentar o erro nas instâncias que foram corretamente classificadas.

3 Para um determinado conjunto de teste com 10 exemplos, a seguinte tabela representa as previsões obtidos com os modelos M1 e M2 para a classe A. Os modelos são classificadores binários por definição de threshold. O valor de threshold usado é 0.9. A coluna Class indica a classe efetiva de cada caso de teste.

M1			M2		
#	Score	Class	#	Score	Class
1	0.996	A	1	0.999	A
2	0.995	A	2	0.998	A
3	0.977	A	3	0.997	B
4	0.951	A	4	0.979	A
5	0.915	B	5	0.931	A
6	0.895	B	6	0.920	A
7	0.881	B	7	0.915	B
8	0.795	B	8	0.812	B
9	0.786	A	9	0.775	B
10	0.675	B	10	0.771	B

		Actual Class	
	M1	A	B
Pedicted Class	A	4	1
	B	1	4

Matriz de confusão para M1

		Actual Class	
	M2	A	B
Pedicted Class	A	5	2
	B	0	3

Matriz de confusão para M2

(a) **Apresente o valor de rácio de erro para os dois modelos.**

Legenda:
AA = Prevê A e é A
AB = Prevê A mas é B
BA = Prevê B mas é A
BB = Prevê B e é B

$$\textbf{M1} : erro = \frac{(AB + BA)}{(AA + AB + BB + BA)} = \frac{(1+1)}{(4+1+4+1)} = \frac{1}{5} \tag{1}$$

$$\textbf{M2} : erro = \frac{(AB + BA)}{(AA + AB + BB + BA)} = \frac{(2+0)}{(5+2+0+3)} = \frac{1}{5} \tag{2}$$

(b) **Qual devia ser o modelo escolhido para esta classe? Justifique.**

$$\textbf{M1} : AUC = \frac{(2+7+8+10) - (5 * \frac{6}{2})}{(5*5)} = 0.84 \tag{3}$$

$$\textbf{M2} : AUC = \frac{(5+6+7+9+10) - (5 * \frac{6}{2})}{(5*5)} = 0.88 \tag{4}$$

Pode-se usar a *Area Under the ROC* , que mede qualidade das previsões do modelo, para escolher o modelo para esta classe. O AUC varia entre 0 e 1, ou seja, todas das previsões erradas ou todas as previsões certas, respetivamente. Como os resultados que obtem-se pode-se dizer que o **modelo 2** deveria ser o escolhido por te uma maior **AUC**.

(c) **Sabendo que** $precision = TP/(TP + FP)$ **e** $FPR = FP/(FP + TN)$ **calcule precision(M1) e FPR(M2) para a classe A e interprete os valores obtidos (nota: FPR = false positive rate).**

$$\textbf{M1} : precision = \frac{AA}{AA + AB} = \frac{4}{4+1} = \frac{4}{5} = 80\% \tag{5}$$

$$\textbf{M2} : FPR = \frac{AB}{AB + BB} = \frac{2}{2+3} = \frac{2}{5} = 40\% \tag{6}$$

Existe uma boa precisão no modelo 1 visto que temos 4 em 5 (80%) de instâncias corretamente classificadas.
Para o **FPR** é melhor quanto menor for o mesmo. Se houver uma maior tolerância a falsos negativos estes 40% serão aceitaveis. Caso contrário este valor é alto.