HOMEWORK 2 – Data Mining

S309164 – Bruno Palermo

Introduction

The purpose of this homework is to exploit data mining classification algorithms to analyze the breast dataset which contains medical records about patients that have contracted breast cancer. Having ascertained that, the recurrence property is considered as a label in RapidMiner. By doing so, the label acts as a target for the following learning operators discerning consequently the patient's tumor is "recurrent or not" basing this analysis on the proposed classification algorithms:

-Decision Tree

-Bayesian classifier

-distance-based classifier (K-NN)

<u>Answering the proposed questions.</u>

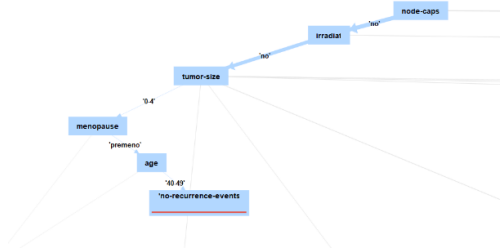
Decision Tree

1a. The most discriminative attribute appearing among the others is "node_caps" because, as shown in the picture, it is the represented root in the decision tree which means that is the one better splitting data so that the gain among the split partitions is maximized.

1b. The height of the Decision Tree generated is 7 provided that the root node is counted in the computation of the total height.

1c. One of the pure parts in the Decision Tree found is the one with the values of the following attributes:

- node-caps: 'no'
- irradiat: 'no'
- tumor-size: '0-4'
- menopause: 'premeno'
- age: '40-49'

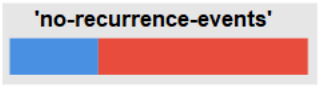


It is pure because all the tuples in it belong to the same class.

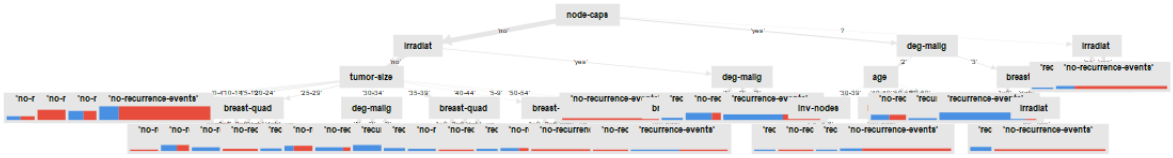
Decision Tree: parameter configuration

2a. Here follows the 5 screenshots showing the Decision Tree with different settings:

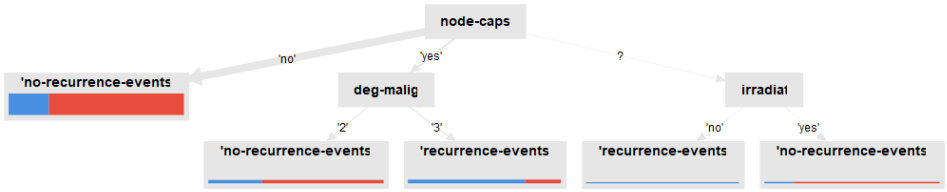
1. Minimal gain: 0.001 Maximal Depth: 5



2. Minimal gain: 0.1 Maximal Depth: 5



3. Minimal gain: 0.01 Maximal Depth: 5



4. Minimal gain: 0.001 Maximal Depth: 3

5. Minimal gain: 0.0001 Maximal Depth: 10

The Decision Tree tries to minimize the entropy of each split it performs yielding a large number of pure small partitions. To do so it chooses the split that maximizes the gain which is the difference between the parent node and the split partitions.

As can be shown in the pictures, by changing the minimal gain and maximal depth, modifications of the final generated decision tree are displayed.

In particular, a bigger minimal gain shows less number of splits since each of them may not reach the requested lower bound resulting in a smaller tree. Instead, the lower it is the more splits on attributes that satisfy the constraint and the more articulated the decision tree is. For example, comparing images 1 and 2 the number of partitions shown is way different.

Secondly, a bigger maximal depth allows displaying of a higher decision tree. For example, comparing images 3 and 5 by changing the maximal depth the difference is clear.

In conclusion, image 2 shows a particular case in which by setting a very high minimal gain any split is shown and each record is labeled as “no-recurrence-events” with a high level of impurity.

10-fold Stratified Cross-Validation

3. The 10-fold Stratified Cross-Validation, performing cross-validation to estimate the performance of the decision tree model, is used to analyze the average accuracy achieved by the Decision Tree using the settings used in the former analysis.

Considerations must be done seeing the variation of the minimal gain. A too-low minimal gain could lead to a high number of small leaf nodes that may amplify the so-called phenomena of overfitting decreasing the accuracy. This means that the learned model fits more against its training data resulting in low accuracy with the testing data. In our case, the low minimal gain of tree number 1 compared to one number 3 shows a slightly lower accuracy because of that.

Secondly, if the minimal gain is too high the accuracy may be badly affected since the opposite phenomenon is experienced: underfitting. So a tradeoff between the two must be found.

On the other hand, a high maximal depth (see examples 1 and 4) and a low maximal depth show a low accuracy since in the last both training and test errors may be large.

Screenshots

accuracy: 69.59% +/- 8.18% (micro average: 69.58%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	35	37	48.61%
pred. 'no-recurrence-events'	50	164	76.64%
class recall	41.18%	81.59%	

1. Minimal gain: 0.001 Maximal Depth: 5

accuracy: 70.30% +/- 1.43% (micro average: 70.28%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	0	0	0.00%
pred. 'no-recurrence-events'	85	201	70.28%
class recall	0.00%	100.00%	

2. Minimal gain: 0.1 Maximal Depth: 5

accuracy: 70.28% +/- 7.75% (micro average: 70.28%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	35	35	50.00%
pred. 'no-recurrence-events'	50	166	76.85%
class recall	41.18%	82.59%	

3. Minimal gain: 0.01 Maximal Depth: 5

accuracy: 74.82% +/- 6.64% (micro average: 74.83%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	11	68.57%
pred. 'no-recurrence-events'	61	190	75.70%
class recall	28.24%	94.53%	

4. Minimal gain: 0.001 Maximal Depth: 3

accuracy: 66.44% +/- 7.66% (micro average: 66.43%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	37	48	43.53%
pred. 'no-recurrence-events'	48	153	76.12%
class recall	43.53%	76.12%	

5. Minimal gain: 0.0001 Maximal Depth: 10

K-Nearest Neighbor (K-NN)

Using this algorithm, the unknown record is labeled using the majority vote of class labels among the k-nearest training neighbors. Varying the k parameter that identifies the number of nearest neighbors I expect accuracy to vary. Setting a low k becomes too sensitive to noise points (example n.1) or if too large the newly included neighborhood may belong to other classes(example n.20) which leads to incorrect results thus a low accuracy rate.

accuracy: 66.44% +/- 7.28% (micro average: 66.43%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	30	41	42.25%
pred. 'no-recurrence-events'	55	160	74.42%
class recall	35.29%	79.60%	

1. k:1

accuracy: 73.77% +/- 5.98% (micro average: 73.78%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	16	61.90%
pred. 'no-recurrence-events'	59	185	75.82%
class recall	30.59%	92.04%	

2. k:5

accuracy: 75.20% +/- 5.43% (micro average: 75.17%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	25	11	69.44%
pred. 'no-recurrence-events'	60	190	76.00%
class recall	29.41%	94.53%	

3. k:10

accuracy: 73.79% +/- 5.61% (micro average: 73.78%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	17	7	70.83%
pred. 'no-recurrence-events'	68	194	74.05%
class recall	20.00%	96.52%	

4. k:20

accuracy: 74.51% +/- 5.02% (micro average: 74.48%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	12	66.67%
pred. 'no-recurrence-events'	61	189	75.60%
class recall	28.24%	94.03%	

5. k: 8

4.b Comparing the results provided by the K-NN classifier varying the k parameter and the Naive Bayes one it is clear that the K-NN is the one that on average is better.

accuracy: 72.45% +/- 7.70% (micro average: 72.38%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	41	35	53.95%
pred. 'no-recurrence-events'	44	166	79.05%
class recall	48.24%	82.59%	

Representation of the confusion matrix achieved by Naive Bayes

Correlation Matrix

Attribut...	age	menop...	tumor-s...	inv-nod...	node-c...	deg-ma...	breast	breast-...	irradiat
age	1	0.241	-0.045	-0.001	0.052	-0.043	0.067	-0.024	-0.011
menopa...	0.241	1	0.019	-0.011	0.130	-0.161	0.077	-0.096	-0.075
tumor-size	-0.045	0.019	1	-0.131	0.058	0.133	-0.022	-0.056	-0.022
inv-nodes	-0.001	-0.011	-0.131	1	-0.465	-0.213	0.040	0.063	0.399
node-caps	0.052	0.130	0.058	-0.465	1	0.098	0.024	-0.036	-0.197
deg-malig	-0.043	-0.161	0.133	-0.213	0.098	1	-0.073	0.018	-0.074
breast	0.067	0.077	-0.022	0.040	0.024	-0.073	1	0.175	-0.019
breast-q...	-0.024	-0.096	-0.056	0.063	-0.036	0.018	0.175	1	-0.005
irradiat	-0.011	-0.075	-0.022	0.399	-0.197	-0.074	-0.019	-0.005	1

5.a The Naive independence is the assumption made in the classification technique based on Bayes Theorem for which the presence of a particular feature in a class is unrelated to the presence of any other feature. Given this consideration and looking at the proposed attributes of the dataset, it doesn't seem there is a dipendence among the features thus the assumption hold. In addition, looking the values of the correlation among the attributes most of them have values close to 0. If there is indipendence then the correlation value must be 0.

5.b Having the highest module, the pair node-caps and inv-nodes is the most correlated among the others. In particular, since the correlation can assume a number between -1 and 1, the degree of association between the two attributes is negative.