Yang_Jinxin_1168646_homework-2

BY BRIANYANG1106@GMAIL.COM

October 14, 2012

1 Question 1

1.1 Preprocessing

I found that there are only 4 numeric attributes which really make sense, because some of them are just duplicates or another way to express the same attribute, these four plus class are:

- 3rd: age-at-heart-attack
- 6th: epss
- 7th: lvdd
- 9th: wall-motion-index
- 13th: alive-at-1

In Question-1, all the things have been done by the program itself without any mannul modification, from "*echocardiogram.data*" to generate the "*echocardiogram.arff*" and output the result.

For preprocessing, the basic step is:

- import one line of data
- find these 5 attributes and store temporarily
- if there is '?' in these 5 attributes, ignore this line
- else output these 5 attributes into standard *arff* file
- store these 5 attributes (at last, there are 61 examples in this mode)

After the function of preprocessing, we can use the arff file to test it in WEKA first. WEKA's Classifier output file is in the directory "*q1/fromWEKA/Classifier output*", and the decision tree is:

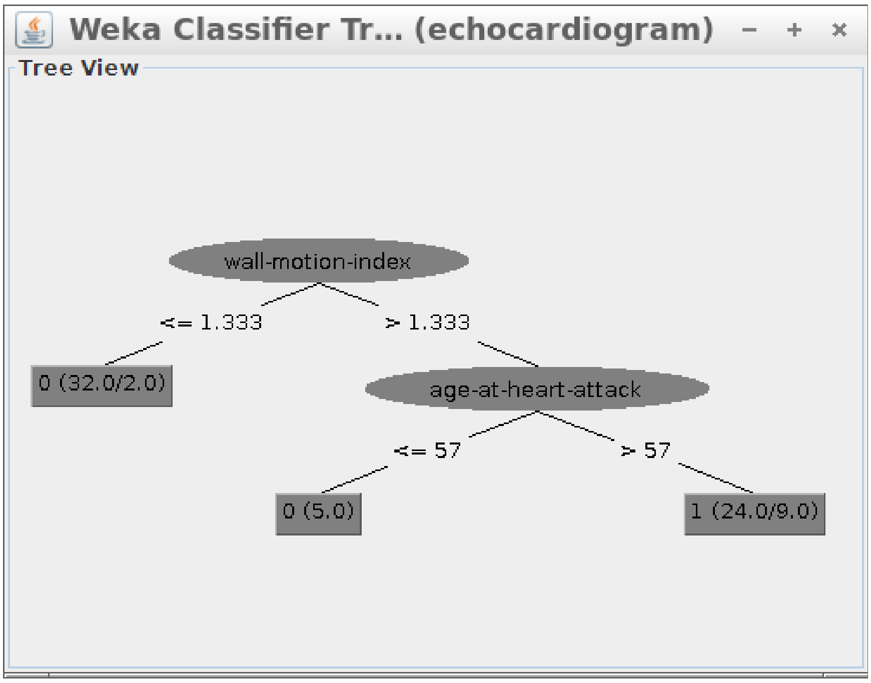


Figure 1. Decision Tree from WEKA J48

We may notice that it uses wall-motion-index as the root and the threshold is 1.333.

1.2 Processing

For 4 attributes, we should set 9 thresholds to test:

- find minimum and maximum values of each attribute
- divide range by 10 and get the size of each bin
- from minimum value, plus certain times' size

then use the information gain formula:

$$\mathrm{IG}(A) = H(S) - \sum \left(\frac{S_v}{S}\right) H(S_v)$$

where $H(X) = -\sum P_i \log P_i$

then we calculate each threshold's information gain and get this graph:

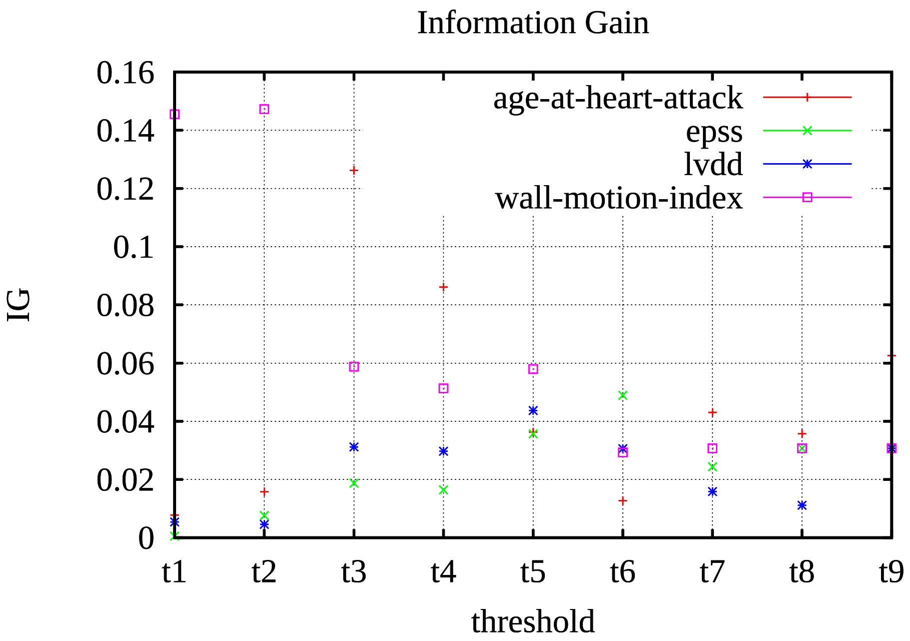
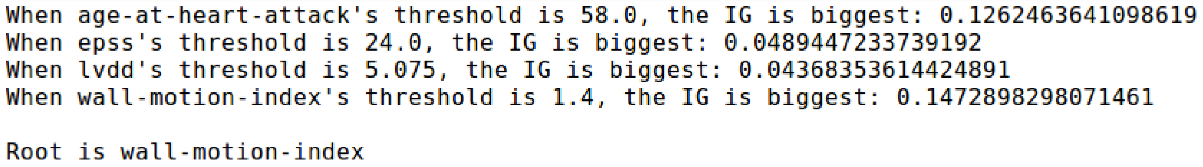


Figure 2. Information Gain of each round

We can find that in t2, the attribute *wall-motion-index* has a global maximum, it can also be seen in the output of program:



```
When age-at-heart-attack's threshold is 58.0, the IG is biggest: 0.1262463641098619
When epss's threshold is 24.0, the IG is biggest: 0.0489447233739192
When lvdd's threshold is 5.075, the IG is biggest: 0.04368353614424891
When wall-motion-index's threshold is 1.4, the IG is biggest: 0.1472898298071461

Root is wall-motion-index
```

Figure 3. Program output echo

1.3 Conclusion

We can find that in my program, the root is *wall-motion-index*, it is identical with WEKA's result, and also my threshold is 1.4 while WEKA's is 1.333, it doesn't mean my result is wrong, because the size in my program is range/10, but WEKA, I am sure that, uses the very small size of between each threshold. Also, we can see the right child, both WEKA and my result is *age-at-heart-attack*, the threshold of this attribute is close too (57 vs 58.0).

So, my program successfully verifies the information gain and attribute choice in this Qusetion.

1.4 By the way

- you can import the directory "*q1/src/IG*" as the "Existing Projects into Workspace" by Eclipse and run ig.java
- it works both on Windows and Linux Eclipse
- if there's something wrong with the import, please email me
- I use Gnuplot to plot the information gain and it should be linespoints, but I don't know why it just displays as points, sorry for that.

2 Quesion 2

Please see the other scaned pdf "*Question-2.pdf*". Thank you.