Christopher Morgan

CS478 : Brother Christophe

Oct 3, 2013

Thoughtfully answer the following questions.

1. **Using your knowledge of how ID3 selects its root node, describe how ID3 could be used to design a simple attribute/feature selection mechanism.**

ID3 selects a root node that causes a split the data that results in a division with the least amount of confusion versus any other attribute that could have been chosen. Thus ID3 could be used to select the most telling attribute for a particular data set, telling meaning the one that reduces the confusion among classes the most. This idea could also be followed to select the second best, third, and so on.

1. **Suppose that a potential large customer *C* has placed your company *BetterSoft* in competition with another company *GoodSoft* to test your relative abilities to develop good software. The task is to design an algorithm to solve a class *P* of problems. Your company produces algorithm *A*, whilst the other company produces algorithm *B*. Both you and your competitor are asked to run your own batch of 350 tests and report how often your algorithm gave acceptable solutions (as defined by *C*). *GoodSoft* comes out on top with a score of 83% against only 78% for your algorithm. Just as *C* is about to award its lucrative contract to *GoodSoft*, you realize that the problems in class *P* are not all of the same complexity. In fact, it turns out that there are two clear levels of difficulty: simple and complex. You ask *C* to collect more detailed data from *GoodSoft*and yourself. The results, when complexity is factored in, are as follows.**

| **Simple** | | **Complex** | |
| --- | --- | --- | --- |
| **Alg A** | **Alg B** | **Alg A** | **Alg B** |
| 81 out of 87 | 234 out of 270 | 192 out of 263 | 55 out of 80 |
| 93% | 86% | 73% | 68% |

* + **May this additional information change *C*'s decision as to which company to hire? If so,how?**

Yes it should change C decision. There was a confounding variable in the aggregate accuracy number, this variable being the different types of complexity found in the training data. Therefore, when our company is compared to the other company with this variable in mind we actually out performed them in both categories. This implies that we should get the deal.

* + **A variable like complexity above is known as a confounding (or lurking) variable, because it interacts with the calculated outcome in a way that may easily be overlooked, but may have an adverse effect on the conclusions reached. Briefly describe another situation based on the real world where confounding effects would be at play.**

There was a study done about colleges and there practices of safe sex. BYU actually rated rather poorly in the study. However, upon inspection the study was performed by Trojan and their criteria for safe sex was the use of condemns. The confounding variable was the practice of abstinence. BYU students simply don’t have sex unless they are married. The practice of abstinence completely undermined the studies findings.

1. Using Weka and the [Waveform](http://dml.cs.byu.edu/~cgc/docs/mldm_tools/Assignments/Datasets/waveform-5000.arff) dataset, experiment with CFS (known as CfsSubsetEval in Weka) and PCA (known as PrincipalComponents in Weka). Compare the results obtained with each method. What is the main difference between these two methods?

CFS: Selected attributes: 4,5,6,7,8,9,10,11,12,13,14,15,16,17,18

PC: Selected attributes: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,

21,22,23,24,25,26,27,28,29,30,31,32,33,34

The results for CFS narrowed the data’s attributes to a selection of only 15 of 40 attributes, where PCA narrowed it to 34 attributes. CFS began its ordering at attribute 4 and continued sequentially to attribute 18, where as PCA also ranked its attributes sequentially but started at 1 and ended at 34.

Using only the CFS selected attributes the J48 classifier improved accuracy from 75% to 77%. However, using only the PCA selected attributes the J48 classifier showed only a 0.8% accuracy improvement. From this it is interesting to note that not only can attribute selection decrease the number of attributes that is being considered but the decrease in attributes can also help improve accuracy depending on the data set and classifying algorithm.

Also, the main difference between these two algorithms is the way in which they try to narrow attributes. CFS is a filter-based algorithm, which tries to filter down attributes to the most representative set. PCA, however, is a feature extraction algorithm that tries to generate new attributes from the existing attributes that will represent the data set.