Report 11-Multiple Comparisons in Induction Algorithms

Xinyu Chen

April 29, 2017

1 Summary

This paper emphasized the pitfall for three pathologies of induction algorithms. The authors thought the cause behind these pathologies was the misuses of evaluation functions that selected item with the maximum score. They called this a *multiple comparison procedure*. These three pathologies included attribute selection errors, overfitting and oversearching.

The paper claimed most induction algorithms made the same mistake in making statistical inferences in MOP,*multiple comparison procedure*. This procedure includes generating n items as n candidates, calculating a score x for each candidate using an evaluation function f on the data sample S and choosing the candidate with the highest score x_{max}. The authors questioned about this procedure. The distribution of sample data S reflected by the scores X is different from the distribution of the max score X_{max}. They used an example of choosing investment advisor to illustrate this idea. In this example, the authors proved that even the investment advisor that had the best score could achieve this by chance as the number of candidates increased. The same procedure were used in induction algorithms so the same mistake could lead to overfitting, attributes selection error and oversearching.

In the following sections, the paper illustrated the distribution of random variables X and X_{max} are different by using joint distribution table and simulations. They proved that x_{max} is a biased estimator of random variable X and the degree of bias increases as n, the number of items involved in the multiple comparison, increases. Some factors affect the difference between the distribution of X_{max} and X. They are independence, sample size and

expected value. When items are correlated, with large sample size, and large difference in expected values, the bias of X_{max} decreases. Based on these factors, the authors proposed solutions to the three pathologies. These included use new data sample, cross-validation, randomization and Boferroni adjustment to overcome the bias between X_{max} and X.

2 Key Takeaway

This paper carries an important idea that machine learning algorithms sometimes have simplifications in using statistics method. This could lead to errors. The authors use examples plus mathematical proofs to remind us to pay attention to such situations. This is a great idea and I should learn from that to look into original statistical methods when implementing machine learning algorithms. Use randomization, cross-validation and boot-strapping to avoid the pathologies they mentioned in this paper.

3 Discussions

The paper is interesting. The following questions are important.

- *Benchmark Dataset.* We have a lot of benchmark datasets in testing and validating learning algorithms. Could these datasets be biased? Could current machine learning algorithms be overfitted by these datasets?

- *Pathologies in algorithms.* The paper mentioned attribute selection error and overfitting. These errors are more relative in building decision tree models. Is there other models that related to such pathologies?

- *The joint distribution example.* In the example of joint distribution of X_1 and X_2, the author showed this distribution is different from X_{max}. But in some algorithms, we are not necessarily looking for joint distribution.

- *Bonferroni adjustment.* We have read about several *family wise error rate* adjustment methods in this class. The Bonferroni adjustment is the most conservative. Is this adjustment very useful in machine learning implementations?