On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach

Steven L. Salzberg, 1997

Presented by Andres Tiko, March 18, 2008

Classifiers

- A Classifier is a mapping from a feature space X to a discrete set of labels Y.
- Many types of algorithms:
  - Nearest neigbour methods
  - Decision trees
  - Error back propagation
  - Reinforcement learning
  - Rule learning
  - etc.

How to choose the algorithm?

- Public databases:
  - UC Irvine repository of machine learning databases http://archive.ics.uci.edu/ml/
- Easy to compare algorithms
- Overreliance on public datasets makes it difficult to produce major new results.
- Many statistically invalid comparisons

Comparing algorithms

- Studies have shown that comparisons are usually not good enough
  - 29% of nearly 200 new algorithms were not evaluated on any real problem
  - Only 8% were compared to more than one alternative on real data
  - Only 3 out of 43 studies in leading journals used a separate data set for parameter tuning
- Many of the reported results might be overly optimistic

Statistical Validity

- p-value is the probability that a result occurred by chance, under the null hypothesis.
- The multiplicity effect:
  - 154 variations in a study were compared to a default classifier, differences reported as significant if p-value < 0.05
  - But expected number of "significant" results is 154 * 0.05 = 7.7

- Let α be the probability that if no differences exist among our algorithms, we will make at least one mistake.
- For each experiment, let the nominal significance level be α*. Then the chance of making the right conclusion for one experiment is 1 – α*.
- For n independent experiments $(1 – α^*)^n$
- A set of different algorithms compared on the same test data are clearly not independent

- Suppose that no real differences exist among the algorithms being tested. Then the chance that we will make at least one mistake is $\alpha = 1 - (1 - \alpha^*)^n$
- For $\alpha^* = 0.05$ $\alpha=0.9996$!
- The Bonferroni adjustment:
  - $\alpha = 1 - (1 - \alpha^*)^{154} \leq 0.05 \rightarrow \alpha^* \leq 0.0003$
- Many researchers still use a simple t-test, which is simply the wrong test for such a comparison (assumes that the test sets are independent)
- The whole framework of using alpha levels and p-values has been questioned when more than two hypotheses are under consideration

Alternative Statistical Tests

- The experimental design cited above only considers overall accuracy.
- A simple (but better) test could compare the percentage of times A > B vs B > A.
- Then use a binomial test for the comparison (with possible Bonferroni adjustment).
- Alternatively use random distinct samples for each algorithm and compare with ANOVA.

A Simple Example

Community Experiments

- The problem is even worse.
- Substantial danger that published results will be mere accidents of chance.
- 100 people study A and B which in fact have the same accuracy.
- We expect 5 of them to get results that are statistically significant at the $p \leq 0.05$ level.
- Duplication of results requires new data, but benchmark databases are normally static.

Repeated tuning

- Many researchers tune their algorithms repeatedly to make them perform optimally on at least some datasets.
- Every adjustment should be considered a separate experiment.
- 10 different combinations of parameters → p-values would have to be 10 times smaller
- Problems with already tuned algorithms.
- The use of cross validation allows to perform virtually unlimited tuning – tune before testing.

Generalizing results

- Common approach is to pick several datasets from the UCI repository.
- Cannot necessarily make more general claims, because UCI is not a representative sample of all problems
- Actually quite limited with many easily classifiable problems
- Algorithms designed with the datasets in mind

A Recommended Approach

- Choose other algorithms to include in the comparison. Make sure to include the algorithm that is most similar to the new algorithm.
- Choose a benchmark data set that illustrates the strength of new algorithm.
- Divide the data into k subsets for cross validation. A typical k = 10. For a small set, it may be better to choose a larger k.

A Recommended Approach

- Run a cross-validation as follows:
  - For each of the k subsets of the data set D, create a training set T = D – k.
  - Divide each training set into two smaller subsets, T1 and T2. T1 will be used for training, and T2 for tuning.
  - Once the parameters are optimized, re-run training on the larger set T.
  - Finally, measure accuracy on k.
  - Overall accuracy is averaged across all k partitions.

A Recommended approach

- To compare algorithms, use binomial test described before or the McNemar variant on that test.

- The above procedure applies to a single dataset. Use Bonferroni adjustment for multiple data sets.

Conclusion

- No single classification algorithm is the best for all problems.
- Comparative studies must be very careful about their methods and their claims.
- But when done correctly, they can be very powerful.
- This was not a criticism of work intended to introduce creative new ideas :)