Evaluating Hypothesis and Experimental Design

Patricia J Riddle

Computer Science 760

Evaluating Hypothesis

- Given observed accuracy of a hypothesis over a limited sample of data, how well does this estimate it's accuracy over additional examples?
- Given that one hypothesis outperforms another over some sample of data, how probable is it that this hypothesis is more accurate in general?
- When data is limited what is the best way to use this data to both learn a hypothesis and estimate its accuracy?

Estimating Hypothesis Accuracy

- Estimating the accuracy with which it will classify future instances - also probable error of this accuracy estimate!!!
- A space of possible instances X.
- Different instances in X may be encountered with different frequencies which is modeled by some unknown probability distribution D.
- Notice D says nothing about whether x is a positive or negative instance.

Learning Task

- The learning task is to learn the target concept, f, by considering a space H of possible hypothesis.
- Training examples of the target function f are provided to the learner by a trainer who draws each instance independently, according to the distribution D and who then forwards the instance x along with the correct target value f(x) to the learner.
- Are instances ever really drawn independently?

Sample error

- Are instances ever really drawn independently?
- Sample error - the fraction of instances in some sample S that it misclassifies

$$error_s(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x))$$

- Where n is the number of samples in S, and $\delta(f(x),h(x))$ is 1 if $f(x) \neq h(x)$ and 0 otherwise

True Error

- True error - probability it will misclassify a single randomly drawn instance from the distribution D

$$error_D(h) \equiv \Pr_{x \in D}[f(x) \neq h(x)]$$

- Where $\Pr_{x \in D}$ denotes that the probability is taken over the instance distribution D.

Sample error versus True error

- Really want $error_D(h)$ but can only get $error_S(h)$.

- How good an estimate of $error_D(h)$ is provided by $error_S(h)$?

Problems with Estimating Accuracy

- Bias in Estimate
- Variance in the Estimate

Bias in Estimate

- Observed accuracy of the learned hypothesis over the training examples is an optimistically biased estimate of hypothesis accuracy over future examples.
- Especially likely when the learner considers a very rich hypothesis space, enabling it to overfit the training examples.
- Typically we test the hypothesis on some set of test examples chosen independently of the training examples and the hypothesis.

Variance in Estimate

- Even if the hypothesis accuracy is measured over an unbiased set of test examples, the measured accuracy can still vary from true accuracy, depending on the makeup of the particular set of test examples.
- The smaller the set of test examples, the greater the expected variance.

Types of Bias

- Machine Learning Bias
- Systematic Error Bias
- “Statistical” Bias

Machine Learning Bias

- Every inductive learning algorithm must adopt a bias in order to generalize beyond the training data.
- This is good and bad!

Systematic Error Bias

- If there is systematic error in the training set, the learning algorithm cannot tell the difference between systematic error and real structure in the dataset.
- Therefore systematic error will also create a bias in the estimate.
- Systematic error example - pull-down menus

Statistical Bias

- Statistical Bias is the systematic error for a given sample size m.

- “statistical bias”is the notion that as the training set size gets smaller, then the systematic error will go up.

We can test for Bias

- But we can’t separate the 3 biases from each other.

- So this will include “statistical bias” and also the ML Bias and the Systematic Error Bias.

Bias Formula

- Bias(A,m,x) = f´(x) - f(x), where A is the learning algorithm, m is the training set size, x is a random example, and f´ is the expected value of f, where the expectation is taken over all possible training sets of fixed size m.

$$f'(x) = \lim_{l \to \infty} \frac{1}{l} \sum_{i=1}^{l} f_{s_i}(x)$$

Variance

- Variance(A,m,x) =E[$(f_S(x)-f'(x))^2$], where f_S is a particular hypothesis learned on training set S.
- Variance comes from variation in the training data, random noise in the training data, or random behavior in the learning algorithm itself.

Error

- So error is just made up of Bias and Variance.
- $Error(A,m,x)=Bias(A,m,x)^2+Variance(A,m,x)$
- Remember that the Bias includes “statistical bias”, Machine Learning Bias, and Systematic Error Bias
- Also Bias is squared only because Variance is already squared

Absolute and Relative ML Bias

- Remember these definitions from decision tree lectures
- Absolute Bias - certain hypothesis are entirely eliminated from the hypothesis space - also called restriction of language bias
- Relative Bias - certain hypothesis are preferred over others - also called preference or search bias

Effect of ML Bias on Stat Bias and Variance

ML Bias Absolute	Relative	Statistical Bias	Variance
appropriate	too strong	high	low
appropriate	ok	low	low
appropriate	too weak	low	high
inappropriate	too strong	high	low
inappropriate	ok	high	moderate
inappropriate	too weak	high	high

Four Important Sources of Error

- Random variation in the selection of the test data - got today right
- Random variation in the selection of the training data - stock newsletters
- Randomness in the learning algorithm (e.g., initial weights) - trying 2000 seeds and only one works well
- Random classification error - guys on the line entering data

Dealing with Error

- Good statistical test should not be fooled by these sources of variation.
- To account for test-data variation and the possibility of random classification error, the statistical procedure must consider the size of the test set and the consequences of changes in the test set.
- To account for training-data variation and internal randomness, the statistical procedure must execute the learning algorithm multiple times and measure the variation in accuracy of the resulting classifiers.

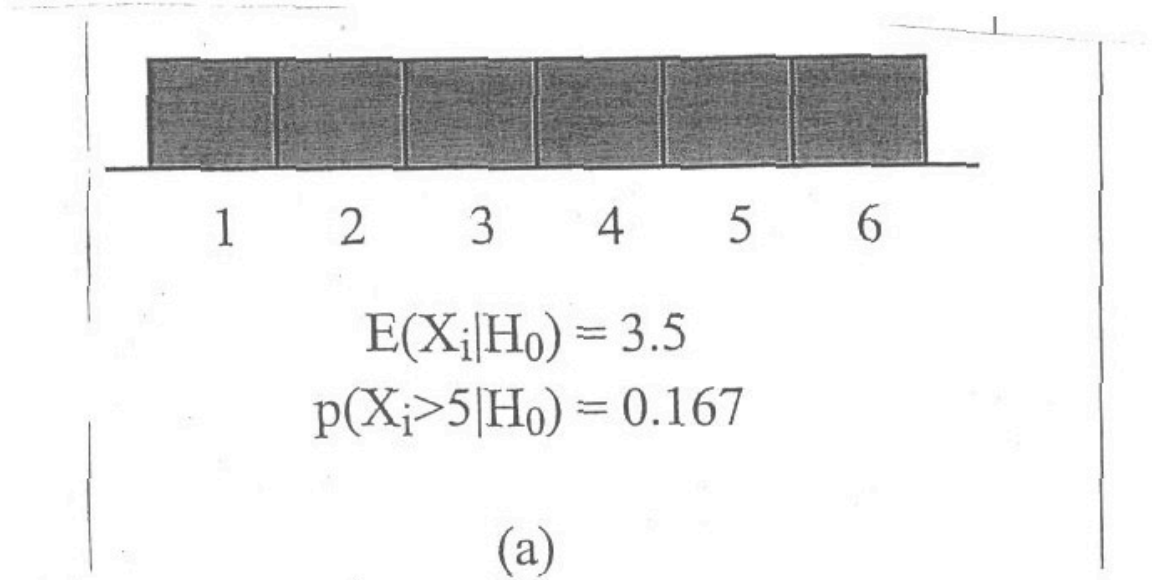
What is Overfitting

- Given a hypothesis space H, a hypothesis h ∈H is said to **overfit** the training data if there exists some alternative hypothesis h´∈ H, such that h has a smaller error than h´ over the training examples, but h´ has a smaller error that h over the entire distribution of instances.
- Not a very useful definition!

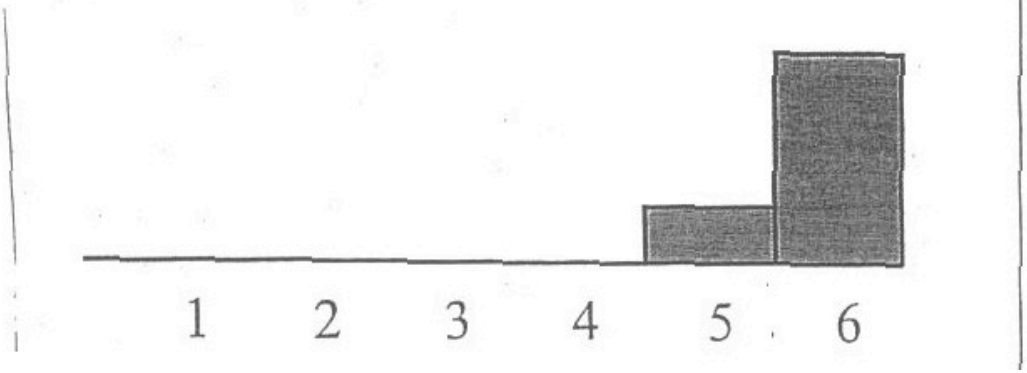
What causes Overfitting?

- Why would complexity cause overfitting???
- What about multiple comparisons?

Sampling Distributions for 1 die and 10 dice?



ampling distributions for one die and ten dice



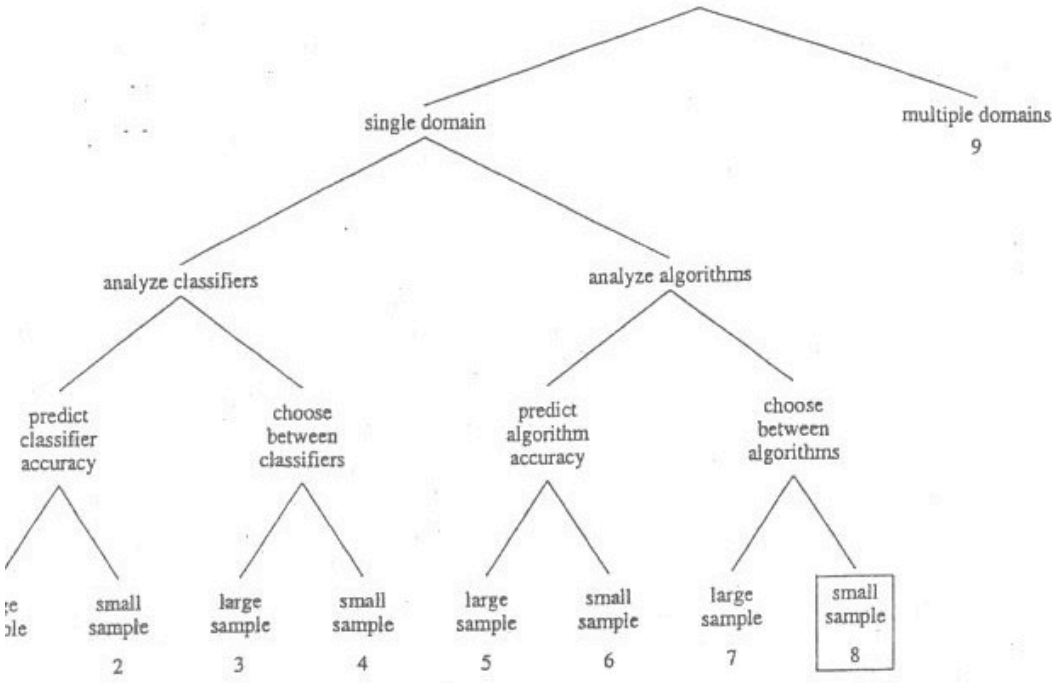
Multiple Comparisons

- Cause overfitting, oversearching, feature selection problems
- Solutions
  - New test data
  - Bonferroni & Sidak (mathematical adjustment, assumes independence)
  - Cross validation - biased if k is to large because then the training sets are virtually the same - leave one out
  - Randomization tests - my favorite - drawback is time complexity - but to estimate p-values between .1 and .01 usually requires no more than 100-1000 trials

Why does Pruning Decision Trees Work?

- By pruning decision trees we are making the hypothesis space smaller (only small decision trees are allowed) so the effect of the multiple comparison's problem is reduced.
- Do I believe this?

Statistical Questions in Machine Learning



ure 1: A taxonomy of statistical questions in machine learning. The boxed node (Question 8) he subject of this paper.

Question Assumptions

- We assume that all datapoints (examples) are drawn independently from a fixed probability distribution defined by the particular problem.
- This is almost never the case!!!

Question 1

- Given a large data set, S, suppose we apply learning algorithm A to S to construct classifier C_A; how accurately will C_A classify new examples?
- The classifier C may have been constructed using part of the data, but there is enough data remaining for a separate test set. Measure the accuracy of C on the test set and construct a binomial confidence interval. Note C need not have been produced by a learning algorithm.

Confidence Intervals for Discrete Value Hypotheses

- Assume S contains n examples drawn independently of each other and of h, according to the probability distribution D.
- Also assume n≥30 and h commits r errors over these n examples ($error_s(h)=r/n$)
- The most probable value of $error_D(h)$ is $error_S(h)$ and
- With 95% probability $error_D(h)$ lies in the interval

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1-error_S(h)}{n}}$$

- This can be done for any percent confidence level, all that must change is the constant 1.96. The smaller the percent the smaller the confidence interval.

Question 3

- Given two classifiers C_A and C_B and enough data for a separate test set, determine which classifier will be more accurate on the new test examples.
- Measure the accuracy of each classifier on the separate test set and apply McNemar's test.

Comparing Learning Algorithms

- Divide sample into training set, R, and test set, T
- Train both algorithms A and B on R yielding the classifiers f_A and f_B
- Test classifiers on T and construct contingency table

Confusion Matrix

n_{00} Number of examples misclassified by both f_A and f_B	n_{01} Number of examples misclassified by f_A but not by f_B
n_{10} Number of examples misclassified by f_B but not by f_A	n_{11} Number of examples misclassified by neither f_A nor f_B

McNemar's Test

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

- Null hypothesis is incorrect if greater than 3.81459

Problems with McNemar

- Did not measure variability due to choice of training set or internal randomness - only use if we believe these to be small
- It does not directly compare performance on training sets of size |S| but only |R| - we must assume the relative difference will still hold for training sets of size |S|

Question 6

- Given a learning algorithm A and a small dataset S, what is the accuracy of the classifiers produced by A when A is trained on new training sets of the same size as S?
- Kohavi shows that stratified 10-fold cross validation produces fairly good estimates.
- Note in any resampling approach we cannot train on training sets of exactly the same size. We train on smaller datasets (90% the size of S) and rely on the assumption that the performance of learning algorithms changes smoothly with changes in the size of the training data. This assumption is not always valid.

10-fold Cross Validation

- Break data into 10 sets of size n/10.
- Train on 9 datasets and test on 1.
- Repeat 10 times and take a mean accuracy.

Question 8

- Given two learning algorithms A and B and a small data set S, which algorithm will produce more accurate classifiers when trained on datasets of the same size as S?
- Because S is small it will be necessary to use holdout and resampling methods.
- This means we cannot answer the question directly without making the assumption that the performance of the two learning algorithms changes smoothly with changes in the size of the training set.
- The technique for this analysis is 5*2 cross validation

Comparing Algorithms with Small Samples

- do 5*2 cross validation
- Each time train both algorithms and test them on each set - this gives 4 error estimates - $p_A^{(1)}$, $p_B^{(1)}$, $p_A^{(2)}$, $p_B^{(2)}$
- Use these measures in the augmented t test as follows
- $p_A^{(1)}$ is (n00+n01)/n for algorithm A and trained on S_1 and tested on S_2
- Subtract corresponding error estimates to get two estimated differences $p^{(1)}=p_A^{(1)}-p_B^{(1)}$ and likewise for $p^{(2)}$

Estimated Variance

- The estimated variance is (n is 2 in this case) $S^2 = (p^{(1)} - \bar{p})^2 + (p^{(2)} - \bar{p})^2$
- Where $\bar{p}$ is $\bar{p} = \frac{1}{n} \sum_{i=1}^{n} p^{(i)}$
- Do this for each cross validation so S_i^2 is the variance computed for the i^{th} repetition

5x2cv t statistic

- 5x2cv t statistic is:

$$t = \frac{p_1^{(1)}}{\sqrt{\frac{1}{5}\sum_{i=1}^{5} s_i^2}}$$

- Using 2-fold cv gives large test sets and disjoint training sets, but training sets are only 1/2 the size which may cause problems
- 5 replications shown to produce lowest Type I error (incorrectly detecting a difference when no difference exists)

Question 9

- Given two learning algorithms A and B and datasets from several domains, which algorithm will produce more accurate classifiers when trained on examples from new domains?
- This is perhaps the most fundamental and difficult question in machine learning
- We need to combine the results from several answers to question 8, where each answer has an associated uncertainty

Four Spurious Effects

- Ceiling Effect - Holte's 1R example

Data set	IR	LA	LY	MU	SE	SO	VO	VI
C4	93.8	77.2	77.5	100	97.7	97.5	95.6	89.4
1R	95.9	87.4	77.3	98.4	95	87	95.2	87.9
Max	95.9	87.4	77.5	100	97.7	97.5	95.6	89.4

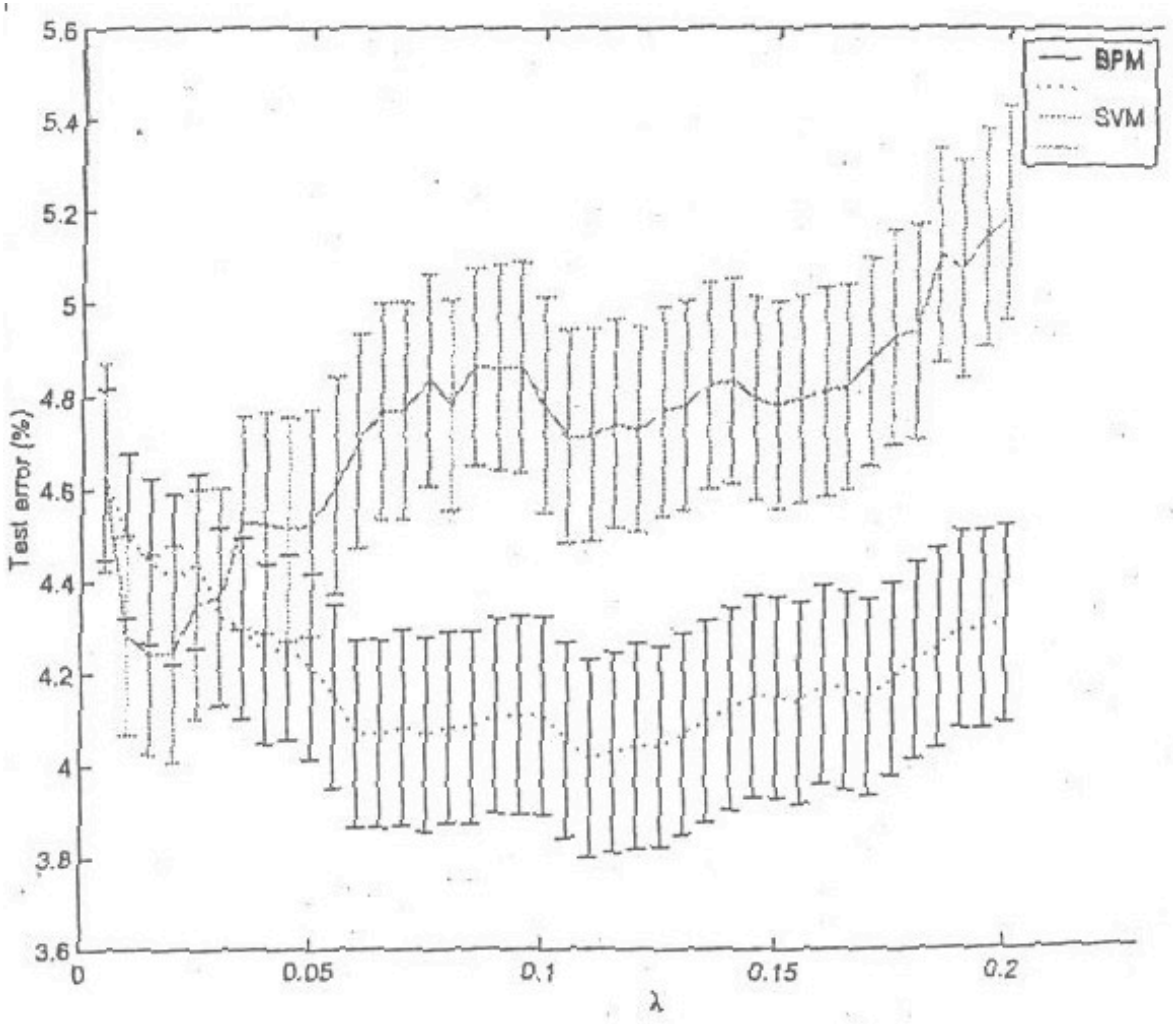
Other 3 Effects

- Regression Effects - if chance plays a role, then always run the same problems
- Order Effects - counter balancing or at least a few orders
- Sampling Bias - how data was collected is very important - the independent variable can change the location of the distribution but not its shape

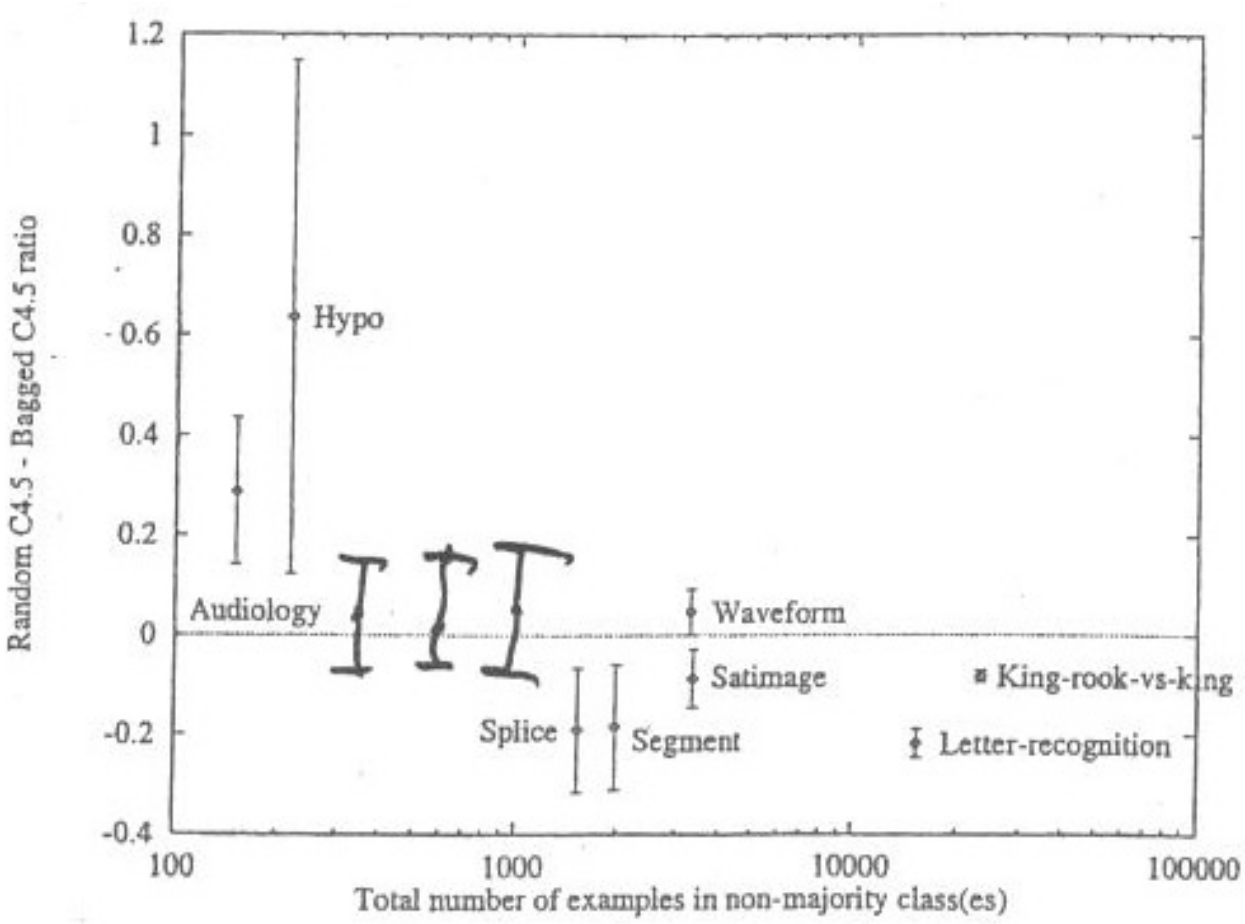
Experiments with Standard Deviation

ex	name	C4.5 P	C4.5 error rate	Randomized C4.5 P	Randomized C4.5 error rate	Bagged C4.5 P	Bagged C4.5 error rate	Adaboosted C4.5 P	Adaboosted C4.5 error rate
	sonar		0.3257±0.0637		0.2018±0.0545	*	0.2752±0.0607	*	0.1651±0.0505
	letter		0.1225±0.0045		0.0285±0.0023		0.0552±0.0032	*	0.0271±0.0023
	splice	*	0.0575±0.0081	*	0.0397±0.0068	*	0.0506±0.0076		0.0503±0.0076
	segment		0.0328±0.0073		0.0203±0.0058		0.0263±0.0065		0.0151±0.0050
	glass	*	0.3437±0.0636		0.2277±0.0562		0.2723±0.0596	*	0.2277±0.0562
	soybean		0.1262±0.0371	*	0.0852±0.0312	*	0.1009±0.0337	*	0.0757±0.0296
	autos		0.2326±0.0578	*	0.1581±0.0499		0.1814±0.0528		0.1814±0.0528
	satimage	*	0.1515±0.0157		0.0890±0.0125		0.1020±0.0133		0.0850±0.0122
	annealing	*	0.0132±0.0075		0.0088±0.0061		0.0099±0.0065		0.0055±0.0048
	krk		0.1887±0.0046		0.1309±0.0039		0.1463±0.0041	*	0.1026±0.0036
	heart-v	*	0.2762±0.0620	*	0.2429±0.0594		0.2619±0.0609	*	0.2810±0.0623
	heart-c	*	0.2396±0.0481	*	0.1853±0.0437	*	0.1981±0.0449		0.2045±0.0454
	breast-y	*	0.2601±0.0508	*	0.2500±0.0502	*	0.2635±0.0511	*	0.3142±0.0538
	phoneme	*	0.1661±0.0086		0.1437±0.0081		0.1509±0.0082	*	0.1464±0.0081
	voting	*	0.1146±0.0299	*	0.0921±0.0272	*	0.0966±0.0278	*	0.1034±0.0286
	vehicle		0.2944±0.0307		0.2477±0.0291		0.2570±0.0294		0.2196±0.0279
	lymph		0.1962±0.0640		0.1772±0.0615		0.1835±0.0624	*	0.1266±0.0536
	breast-w	*	0.0494±0.0161	*	0.0353±0.0137		0.0367±0.0139		0.0310±0.0128
	credit-g	*	0.2921±0.0282		0.2416±0.0265	*	0.2495±0.0268		0.2347±0.0263
	primary	*	0.5845±0.0525	*	0.5501±0.0530		0.5645±0.0528	*	0.5960±0.0522
	shuttle		0.0003±0.0003		0.0002±0.0002		0.0002±0.0002		0.0001±0.0002
	heart-s	*	0.0677±0.0444	*	0.0677±0.0444	*	0.0677±0.0444	*	0.0902±0.0506
	iris		0.0563±0.0369	*	0.0500±0.0349	*	0.0500±0.0349	*	0.0688±0.0405
	sick	*	0.0132±0.0036		0.0137±0.0037		0.0137±0.0037	*	0.0095±0.0031
	hepatitis		0.1758±0.0599		0.1636±0.0582		0.1636±0.0582	*	0.1636±0.0582
	credit-a	*	0.1614±0.0275	*	0.1400±0.0259		0.1371±0.0257	*	0.1300±0.0251
	waveform	*	0.2341±0.0117		0.1784±0.0106		0.1675±0.0104		0.1521±0.0100
	horse-colic	*	0.1561±0.0371		0.1561±0.0371		0.1481±0.0363	*	0.1825±0.0395
	heart-h	*	0.1645±0.0424	*	0.1809±0.0440	*	0.1579±0.0417		0.2039±0.0461
	labor		0.1493±0.0925	*	0.1493±0.0925		0.1194±0.0842	*	0.1194±0.0842
	krkp		0.0075±0.0030		0.0075±0.0030		0.0056±0.0026	*	0.0037±0.0021
	audiology		0.2203±0.0540	*	0.2458±0.0561		0.1822±0.0503	*	0.1525±0.0469
	hypo		0.0058±0.0024	*	0.0079±0.0028		0.0042±0.0021	*	0.0040±0.0020

Experiments with Learning Curves



Experiments with Difference in Performance Graph



Experiments with Pairwise Combination Chart

Table 3. All pairwise combinations of the four *methods for* four levels of noise and 9 domains. Each cell contains the number of wins, losses, and ties between the algorithm in that row and the algorithm in that column.

Noise = 0%	C4.5	Adaboost C4.5	Bagged C4.5
Random C4.5	5 – 0 – 4	1 – 6 – 2	3 – 3 – 3
Bagged C4.5	4 – 0 – 5	0 – 5 – 4	
Adaboost C4.5	6 – 0 – 3		

Noise = 5%	C4.5	Adaboost C4.5	Bagged C4.5
Random C4.5	5 – 2 – 2	3 – 2 – 4	1 – 5 – 3
Bagged C4.5	6 – 0 – 3	5 – 1 – 3	
Adaboost C4.5	3 – 3 – 3		

Noise = 10%	C4.5	Adaboost C4.5	Bagged C4.5
Random C4.5	4 – 1 – 4	5 – 1 – 3	1 – 6 – 2
Bagged C4.5	*5 – 0 – 4*	*6 – 1 – 2*	
Adaboost C4.5	2 – 3 – 4		

Noise = 20%	C4.5	Adaboost C4.5	Bagged C4.5
Random C4.5	5 – 2 – 2	5 – 0 – 4	0 – 2 – 7
Bagged C4.5	7 – 0 – 2	6 – 0 – 3	
Adaboost C4.5	3 – 6 – 0		

Summary

- What questions are we interested in asking?
- Binomial Confidence intervals, McNemar test, 5x2cv paired t test
- Problems to watch out for in experimental design
- Real cause of overfitting.