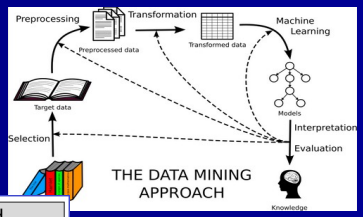
ACCURACY ESTIMATION AND STATISTICAL TESTS FOR COMPARISON

	Predicted	
	Positive	Negative
Actual True	TP	FN
Actual False	FP	TN

Iñaki Inza
Intelligent Systems Group, www.sc.ehu.es/isg
Computer Science Faculty
University of the Basque Country, Donostia - San Sebastian

OUTLINE

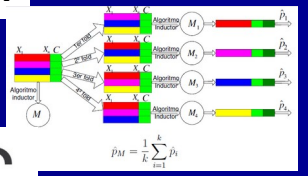
- Scenario – metrics: vocabulary on accuracy estimation
- Accuracy estimation techniques: hold-out, cross-validation, bootstrap
- Comparing classification schemes: the need of statistical tests



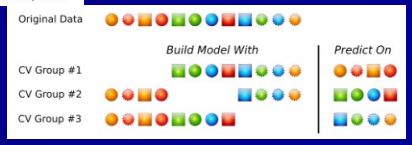
	Predicted	
	Positive	Negative
Actual True	TP	FN
Actual False	FP	TN

Dataset	NB	SVM	Adaboost	Rand Forest
Anneal	96.43	99.44	83.63	99.55
Audiology	73.42	81.34	46.46	79.15
Balance Scale	72.30	91.51	72.31	80.97
Breast Cancer	71.70	66.16	70.28	69.99
Contact Lenses	71.67	71.67	71.67	71.67
Pima Diabetes	74.36	77.08	74.35	74.88
	70.63	62.21	44.91	79.87
	83.21	80.63	82.54	84.58

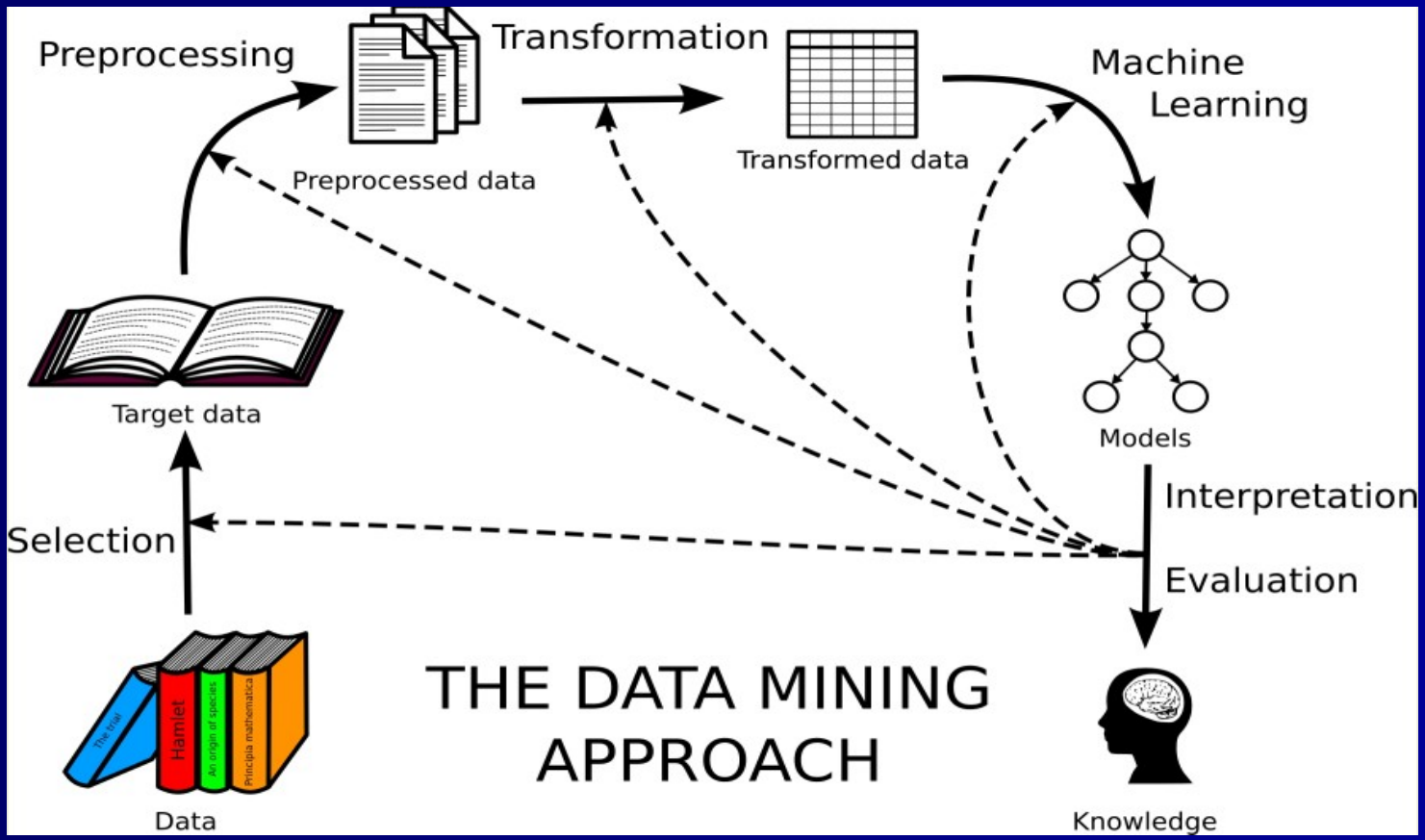
$$t = \frac{m_d}{\sqrt{\sigma_d^2 / k}}$$



STAC
Statistical Tests for Algorithms Comparison

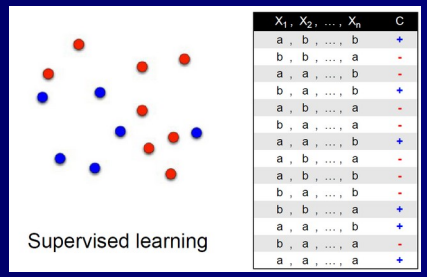


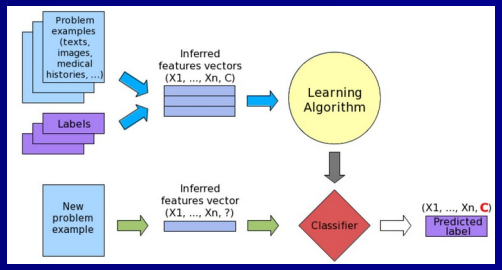
DATA MINING PIPELINE



SCENARIO - PROBLEM

- Fix the learning scenario: supervised classification

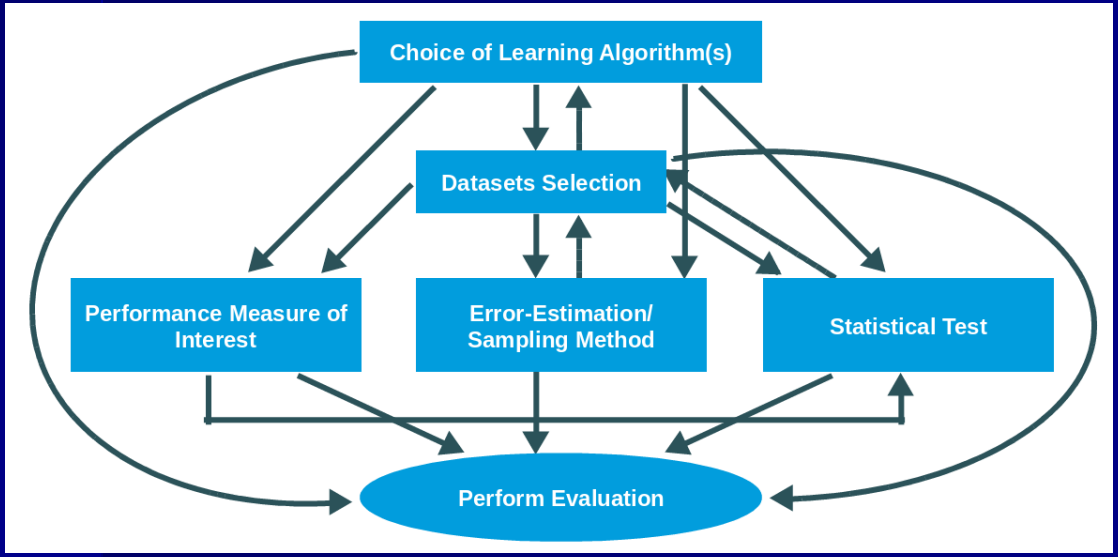


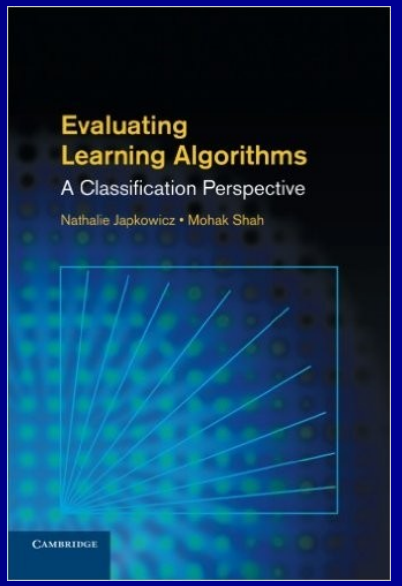


- Fix the problem:
- Estimate the goodness of the learned model over future samples
- Estimate its “generalization ability”
- In “production time”: prediction is a “bet”

- Key ideas:
  - final model is learned with the whole dataset
  - goodness estimation is performed with the whole dataset

MAIN STEPS OF EVALUATION





Artif Intell Rev (2015) 44:467–508
DOI 10.1007/s10462-015-9433-y

CrossMark

Dealing with the evaluation of supervised classification algorithms

Guzman Santafe[1] · Iñaki Inza[2] · Jose A. Lozano[2]

ERROR COUNTS

- Starting from the popular confusion matrix:

	Predicted	
	Positive	Negative
Actual True	**TP**	**FN**
Actual False	**FP**	**TN**

$$\text{Sensitivity} = \text{Recall} = \text{TP-rate} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

- How are the errors counted in the final evaluation?
  - Commonly, Accuracy:
    - under 0/1 loss, "the winner class takes all"
  - Cost-sensitive classification
  - Probabilistic error (e.g. Brier score)

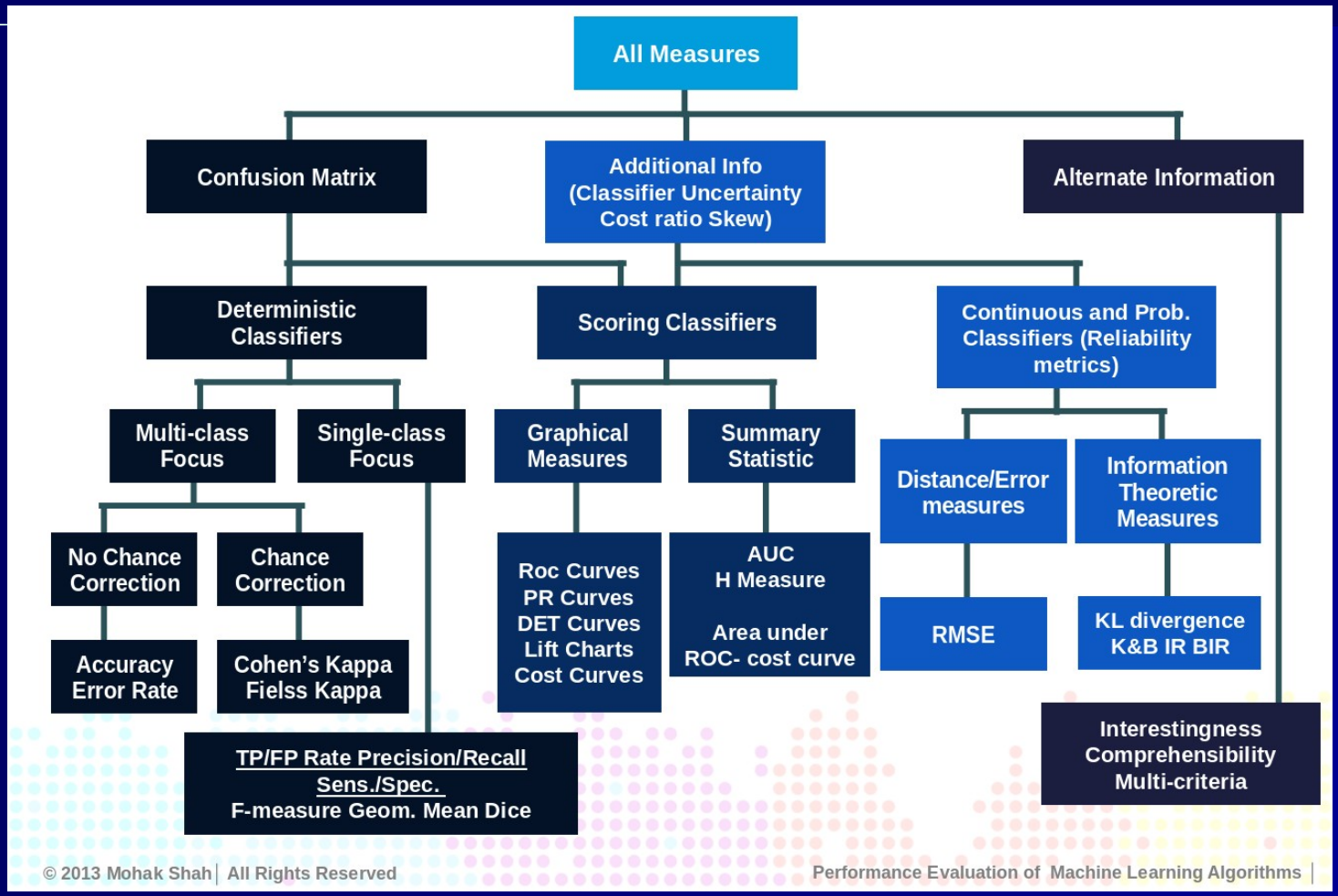
ERROR COUNTS
BRIER SCORE – CALIBRATION

	X_1	...	X_n	C	$p(C_M = 0\|\boldsymbol{x})$	$p(C_M = 1\|\boldsymbol{x})$
$(\boldsymbol{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	...	$x_n^{(1)}$	1	0.18	0.82
$(\boldsymbol{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	...	$x_n^{(2)}$	0	0.51	0.49
...		...		...	...	
$(\boldsymbol{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	...	$x_n^{(N)}$	1	0.55	0.45

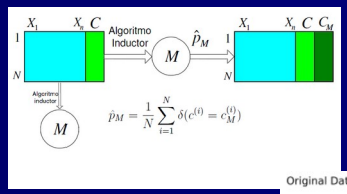
$$B = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2}[p(C_M = c|\boldsymbol{x}^{(i)}) - \delta(c^{(i)}, c_M^{(i)})]^2$$

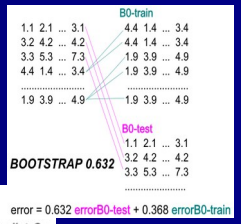
$$B = \tfrac{1}{N}[(0.18-0)^2 + (0.82-1)^2 + (0.51-1)^2 + (0.49-0)^2 + \\ \ldots + (0.55-0)^2 + (0.45-1)^2]$$

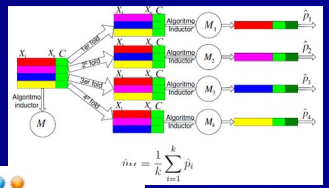
OVERVIEW OF PERFORMANCE MEASURES

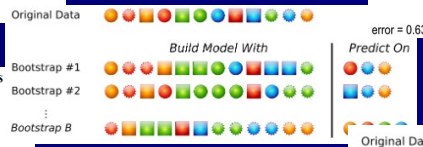


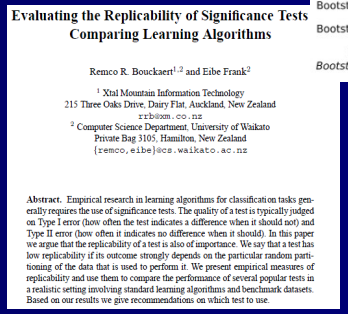
RESHUFFLING THE DATASET FOR ERROR ESTIMATION









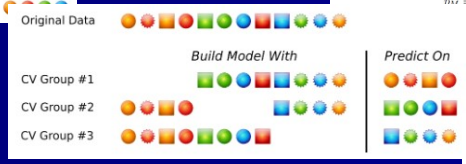


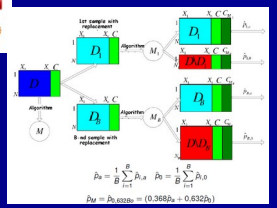
Evaluating the Replicability of Significance Tests Comparing Learning Algorithms

Remco R. Bouckaert[1,2] and Eibe Frank[2]

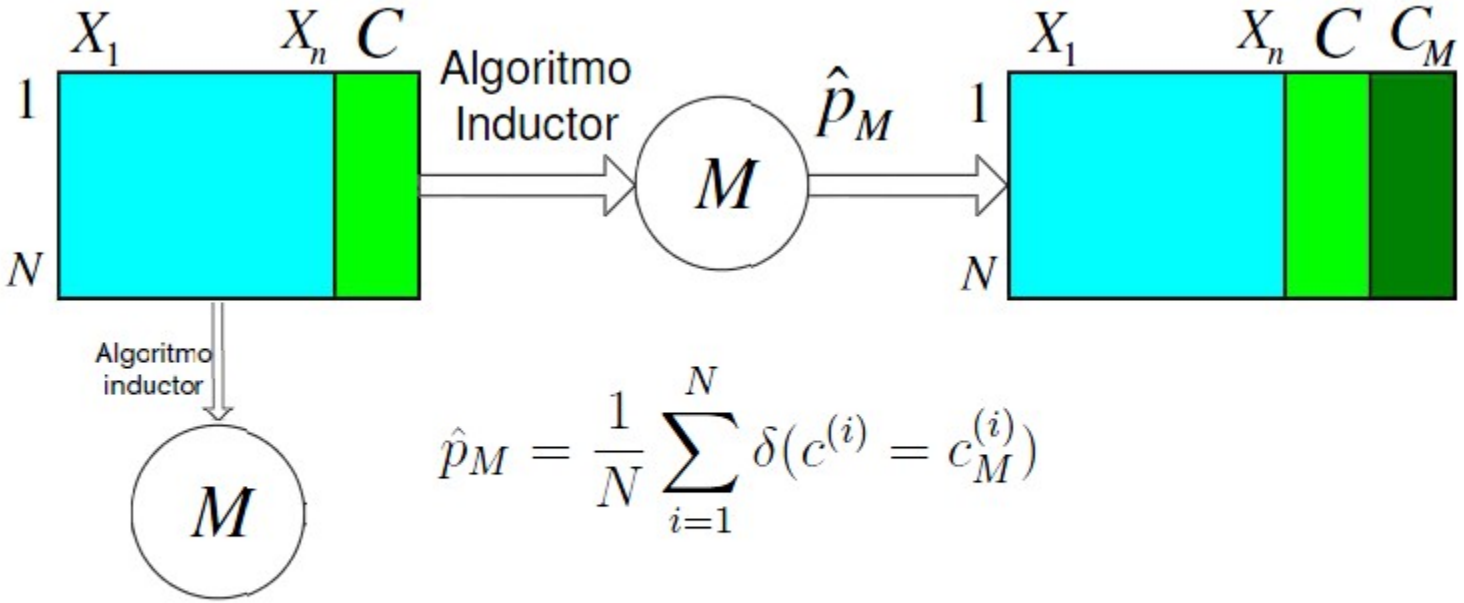
[1] Xtal Mountain Information Technology
215 Three Oaks Drive, Dairy Flat, Auckland, New Zealand
rrb@xtm.co.nz
[2] Computer Science Department, University of Waikato
Private Bag 3105, Hamilton, New Zealand
{remco,eibe}@cs.waikato.ac.nz

Abstract. Empirical research in learning algorithms for classification tasks generally requires the use of significance tests. The quality of a test is typically judged on Type I error (how often the test indicates a difference when it should not) and Type II error (how often it indicates no difference when it should). In this paper we argue that the replicability of a test is also of importance. We say that a test has low replicability if its outcome strongly depends on the particular random partitioning of the data that is used to perform it. We present empirical measures of replicability and use them to compare the performance of several popular tests in a realistic setting involving standard learning algorithms and benchmark datasets. Based on our results we give recommendations on which test to use.





RESUBSTITUTION ERROR
OPTIMISTIC, NO HONEST

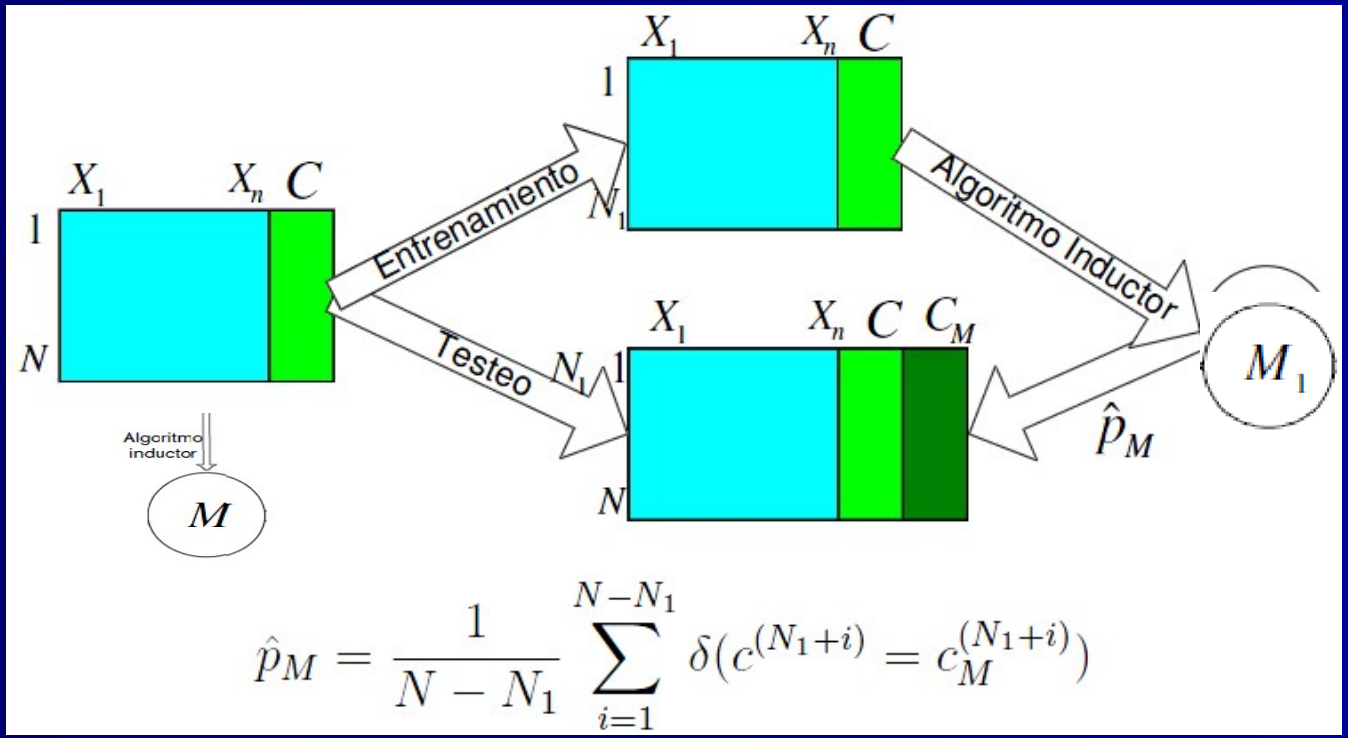


$$\hat{p}_M = \frac{1}{N} \sum_{i=1}^{N} \delta(c^{(i)} = c_M^{(i)})$$

PERCENTAGE SPLIT
HOLD-OUT

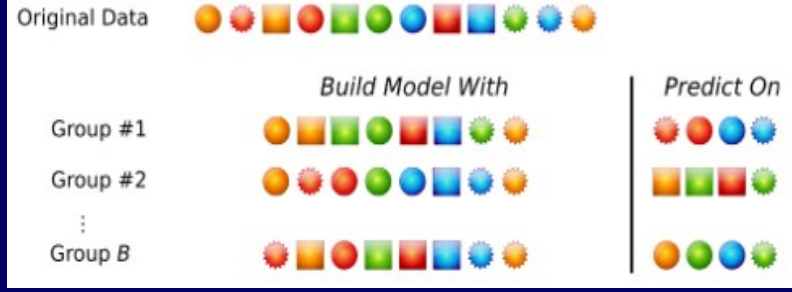
- Randomly split data into training and test sets (usually 2/3 for train, 1/3 for test)

- Build a classifier using the *train* set and evaluate it using the *test* set

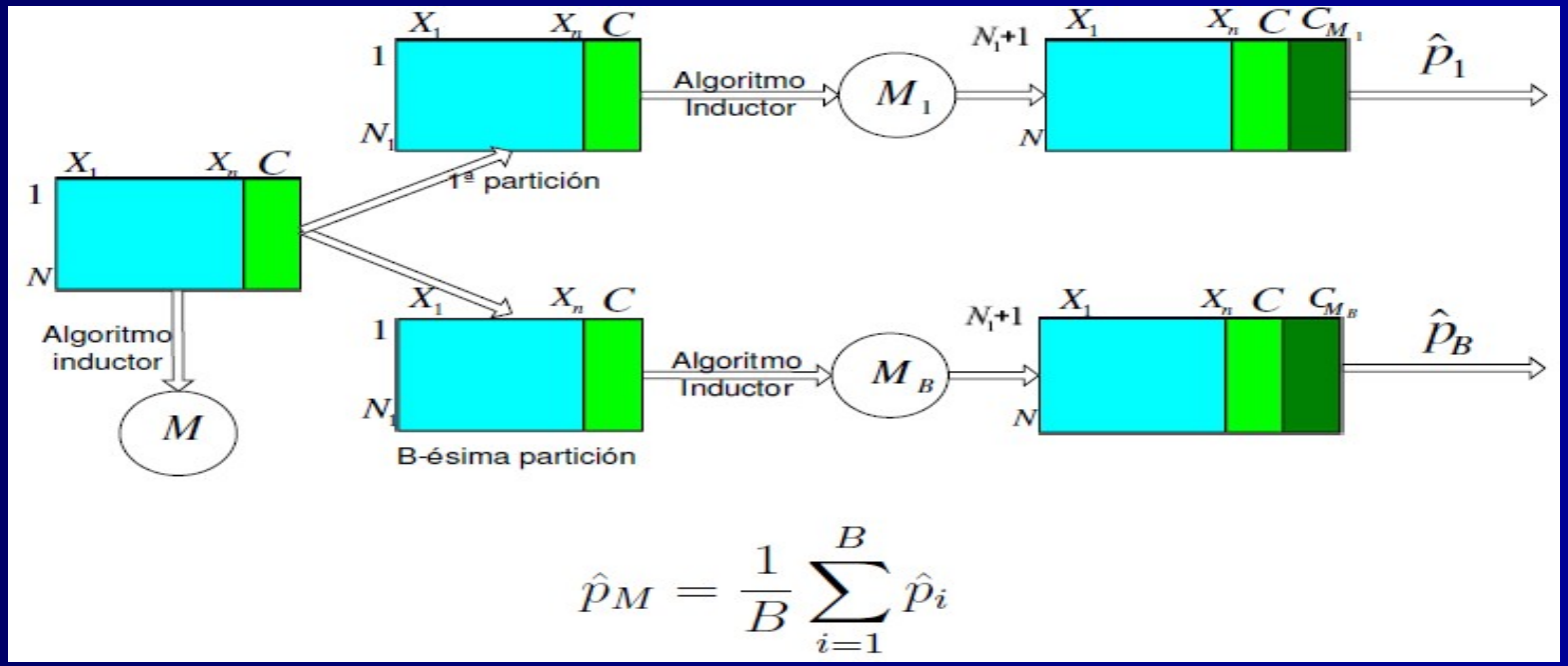


$$\hat{p}_M = \frac{1}{N - N_1} \sum_{i=1}^{N-N_1} \delta(c^{(N_1+i)} = c_M^{(N_1+i)})$$

REPEATED HOLD-OUT

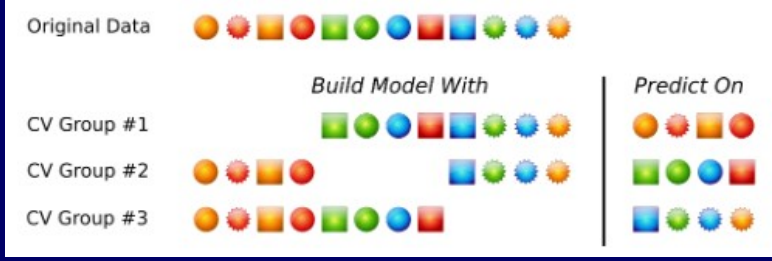


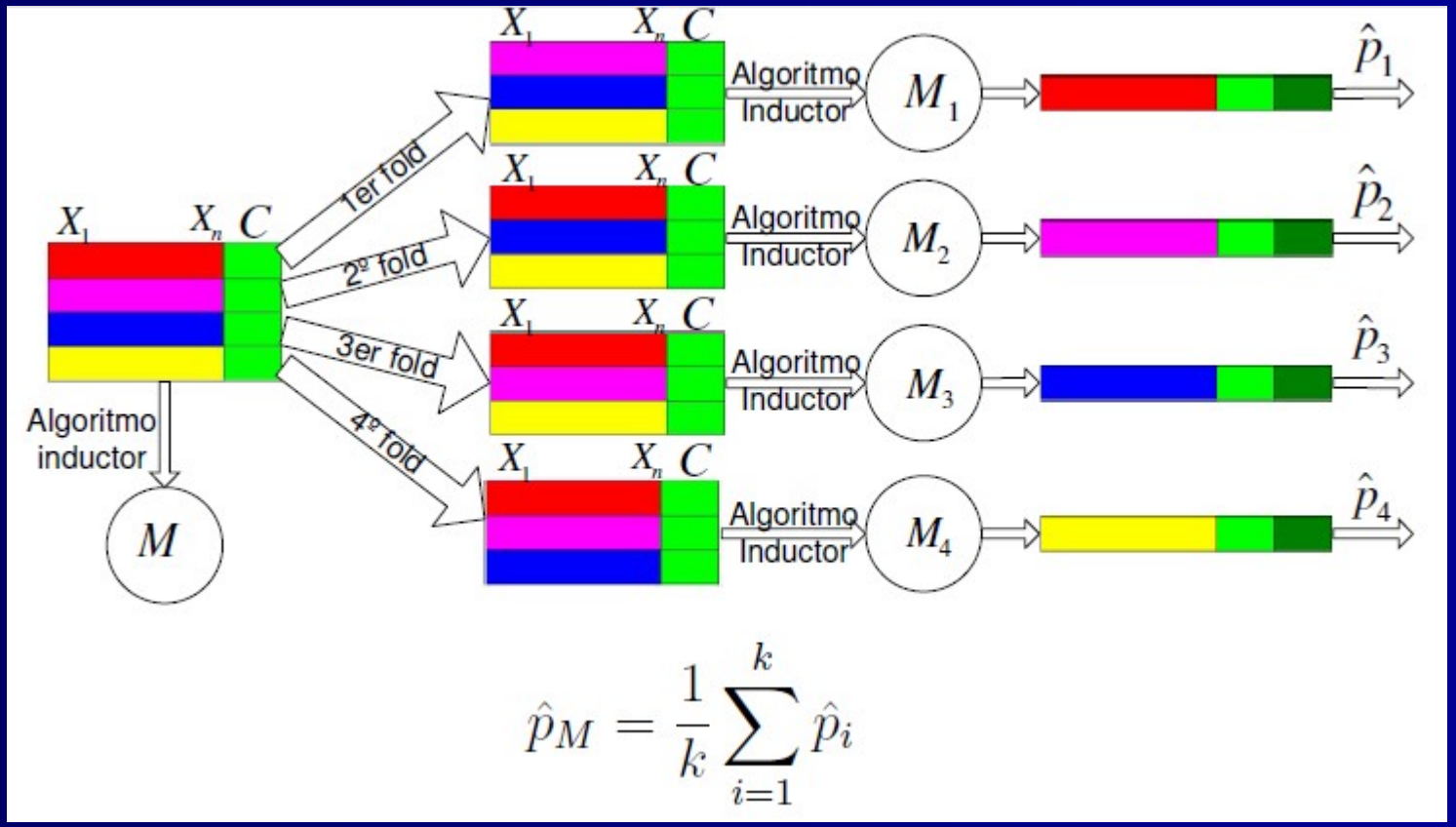
- Hold-out method can be improved, repeating the estimation process several times
- In this way, statistical comparisons between different classifiers can be performed



$$\hat{p}_M = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_i$$

CROSS-VALIDATION

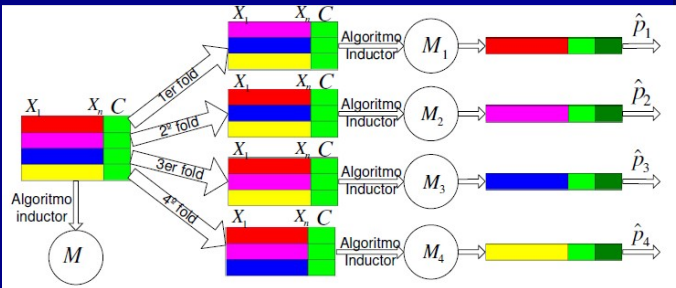




$$\hat{p}_M = \frac{1}{k} \sum_{i=1}^{k} \hat{p}_i$$

CROSS-VALIDATION

- Often the subsets are *stratified* before the cross-validation is performed
- When *k* (nº folds) equals the number of samples of the datasets → leave one out (LOOCV)
- Most popular validation technique (10-folds)



$$\hat{p}_M = \frac{1}{k}\sum_{i=1}^{k}\hat{p}_i$$

REPEATED CROSS-VALIDATION

- Even better: repeated stratified cross-validation:

  10-fold cross-validation is repeated 10 times and results are averaged (reduces the variance): 10x10cv

Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms

Remco R. Bouckaert[1,2] and Eibe Frank[2]

[1] Xtal Mountain Information Technology
215 Three Oaks Drive, Dairy Flat, Auckland, New Zealand
rrb@xm.co.nz
[2] Computer Science Department, University of Waikato
Private Bag 3105, Hamilton, New Zealand
{remco,eibe}@cs.waikato.ac.nz

Abstract. Empirical research in learning algorithms for classification tasks generally requires the use of significance tests. The quality of a test is typically judged on Type I error (how often the test indicates a difference when it should not) and Type II error (how often it indicates no difference when it should). In this paper we argue that the replicability of a test is also of importance. We say that a test has low replicability if its outcome strongly depends on the particular random partitioning of the data that is used to perform it. We present empirical measures of replicability and use them to compare the performance of several popular tests in a realistic setting involving standard learning algorithms and benchmark datasets. Based on our results we give recommendations on which test to use.

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,

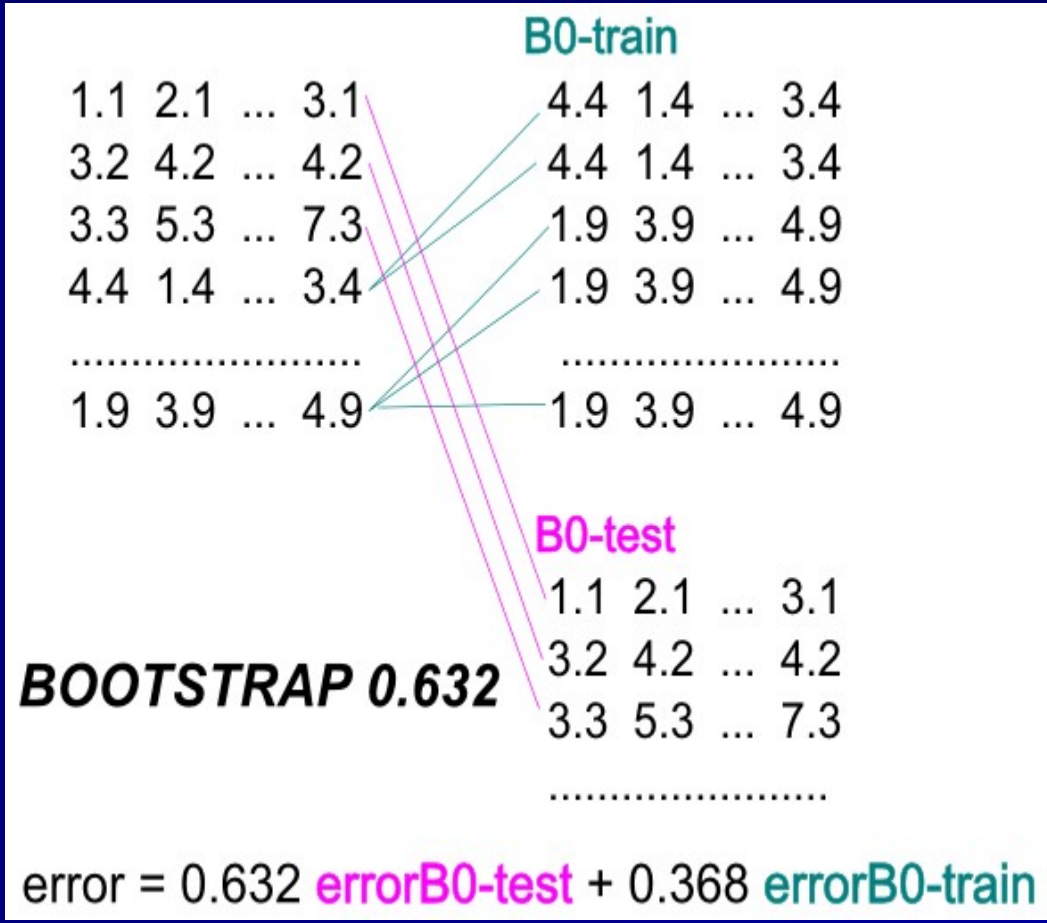
Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation

Juan Diego Rodríguez, Aritz Pérez, and
Jose Antonio Lozano, *Member*, *IEEE*

Abstract—In the machine learning field, the performance of a classifier is usually measured in terms of prediction error. In most real-world problems, the error cannot be exactly calculated and it must be estimated. Therefore, it is important to choose an appropriate estimator of the error. This paper analyzes the statistical properties, bias and variance, of the k-fold cross-validation classification error estimator (k-cv). Our main contribution is a novel theoretical decomposition of the variance of the k-cv considering its sources of variance: sensitivity to changes in the training set and sensitivity to changes in the folds. The paper also compares the bias and variance of the estimator for different values of k. The experimental study has been performed in artificial domains because they allow the exact computation of the implied quantities and we can rigorously specify the conditions of experimentation. The experimentation has been performed for two classifiers (naive Bayes and nearest neighbor), different numbers of folds, sample sizes, and training sets coming from assorted probability distributions. We conclude by including some practical recommendation on the use of k-fold cross validation.

BOOTSTRAPING

- Sampling with replacement to form the training set
- Sample dataset of *n* instances *n* times with replacement ~ form a new dataset of *n* instances
- Use this data for training set
- Use the instances from the original dataset that don't occur in the new training set for testing

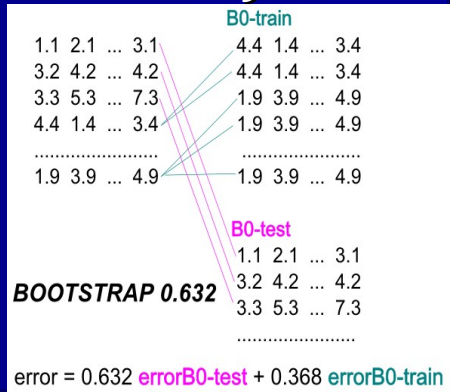


0.632 BOOTSTRAP

- A particular instance has a probability of (1−1/*n*) of NOT being picked
- Thus its probability of ending up in the test data is

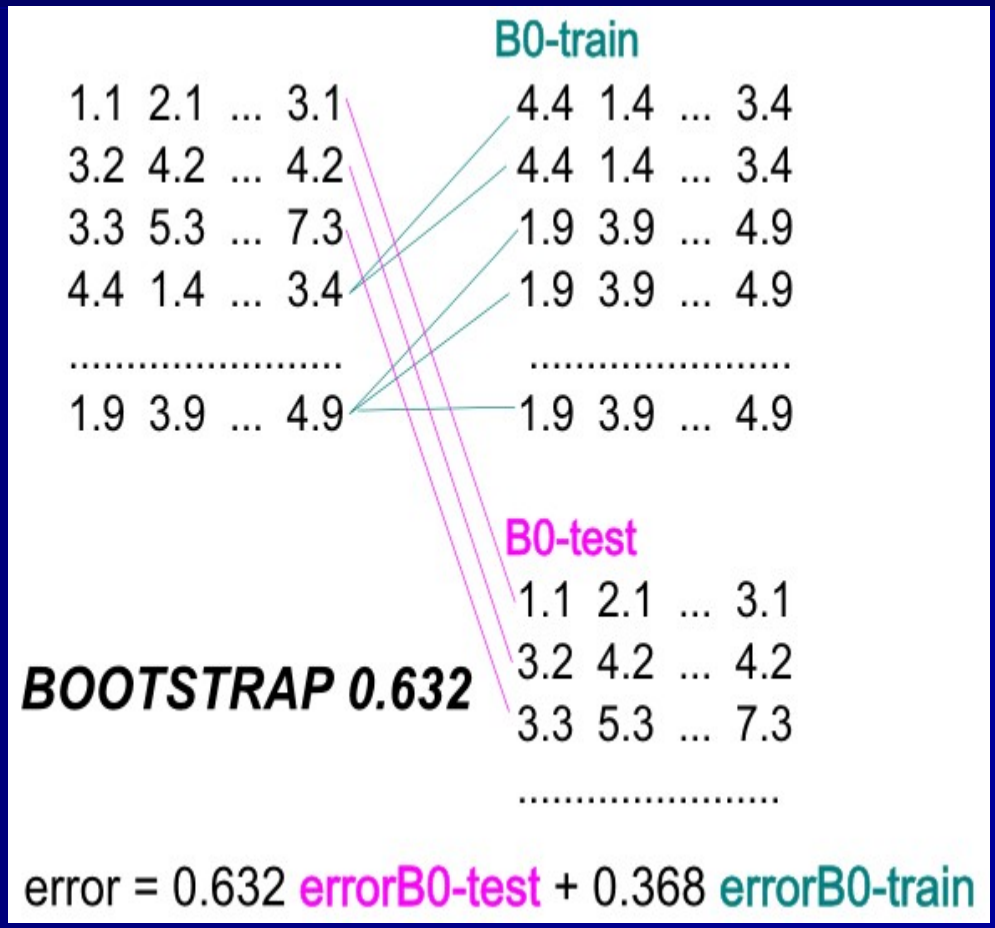
$$\left(1-\frac{1}{n}\right)^{n} \approx e^{-1} = 0.368$$

- This means the training data will contain approximately 63.2% of the instances

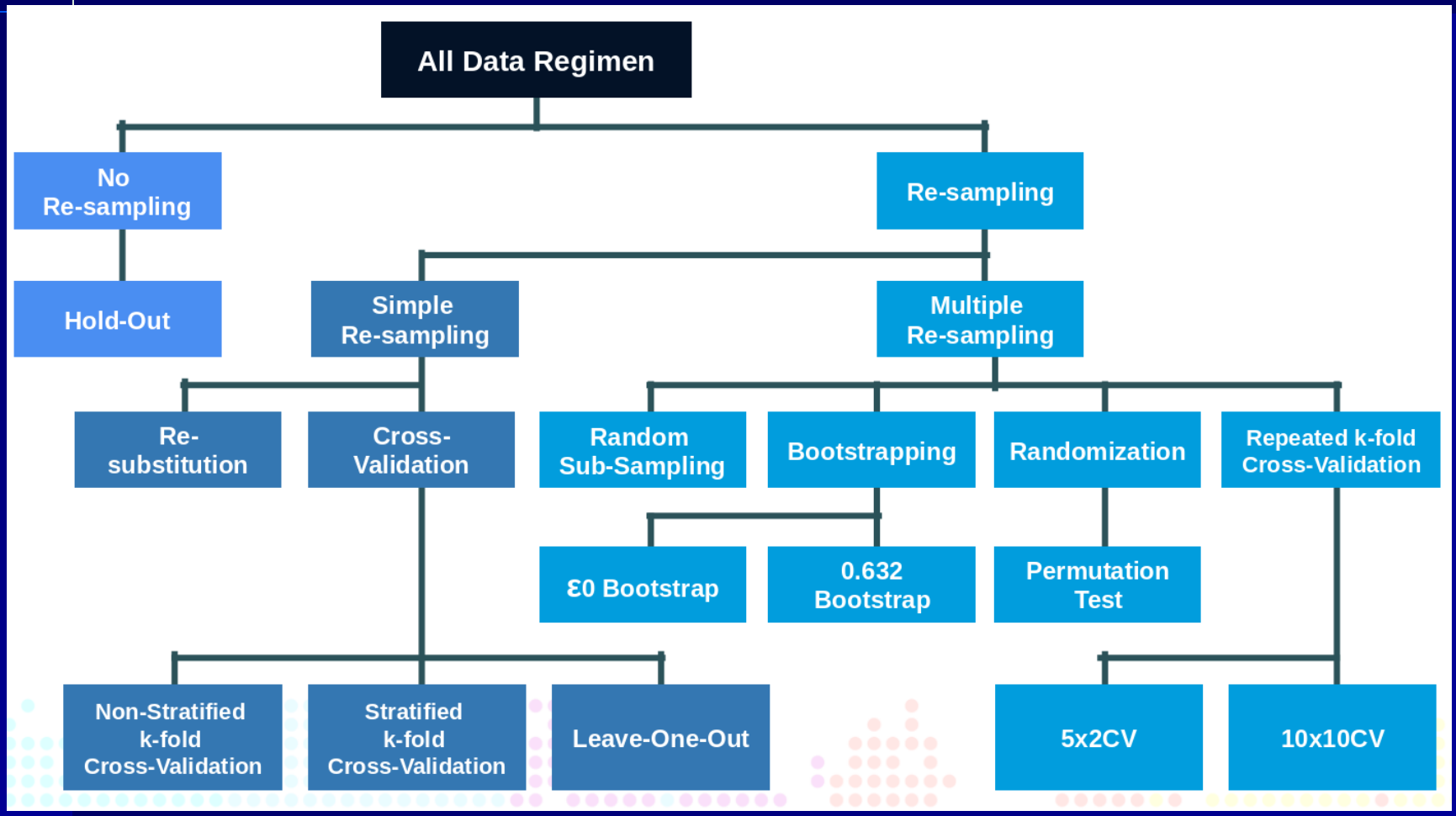


0.632 BOOTSTRAP

- The error estimate on the test data will be very pessimistic
- Trained on just ~63% of the instances
- Therefore, combine it with the resubstitution error
- Repeat process several times with different replacement samples; average the results

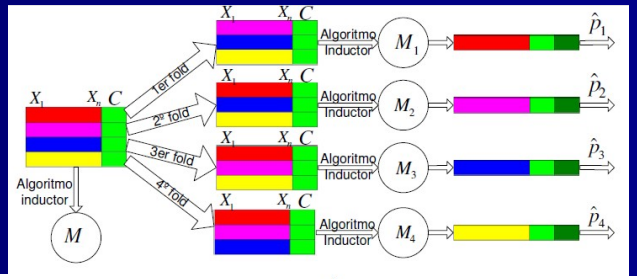


ONTOLOGY OF ERROR ESTIMATION TECHNIQUES



OVERFITTING RISK

- Data in test partitions is not used *in any way* to learn the classifier
  - Not tune with entire data
  - The test data can't be used for parameter tuning!
  - Neither for discretization, imputation, feature selection…



$$\hat{p}_M = \frac{1}{k}\sum_{i=1}^{k}\hat{p}_i$$

Journal of Machine Learning Research 3 (2003) 1371-1382 Submitted 5

Overfitting in Making Comparisons Between Variable Selection Methods

Juha Reunanen JUHA.REUNANEN@FI.ABB.COM
ABB, Web Imaging Systems
P.O. Box 94, 00381 Helsinki, Finland

TICS **LETTER TO THE EDITOR** Vol. 26 no. 3 2010, pages 440–443
doi:10.1093/bioinformatics/btp621

Pitfalls of supervised feature selection

Pawel Smialowski[1,2,*], Dmitrij Frishman[1,2] and Stefan Kramer[3]
[1]Department of Genome Oriented Bioinformatics, Technische Universität München Wissenschaftszentrum Weihenstephan, Am Forum 1, 85350 Freising, [2]Helmholtz Zentrum Munich, National Research Center for Environment and Health, Institute for Bioinformatics, Ingolstädter Landstraße 1, 85764 Neuherberg and [3]Institut für Informatik/I12, Technische Universität München, Boltzmannstr. 3, 85748 Garching b. München, Germany

Received and revised on October 7, 2009; accepted on October 26, 2009
Advance Access publication October 29, 2009
Associate Editor: Martin Bishop

BIAS + VARIANCE ERROR DECOMPOSITION

- Bias and variance decomposition of the estimated error
- Specially recommended for low number of training samples

To appear in Machine Learning: Proceedings of the Thirteenth International Conference, 1996

Bias Plus Variance Decomposition for Zero-One Loss Functions

Ron Kohavi
Data Mining and Visualization
Silicon Graphics, Inc.
2011 N. Shoreline Blvd
Mountain View, CA 94043-1389
ronnyk@sgi.com

David H. Wolpert
The Santa Fe Institute
1399 Hyde Park Rd.
Santa Fe, NM 87501
dhw@santafe.edu

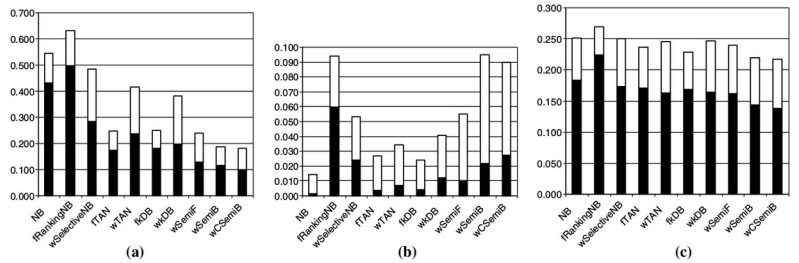


Fig. 4. Bias-variance decomposition examples. (a) *VEHICLE* data set. (b) *WINE* data set. (c) Average across all data sets.

STATISTICAL COMPARISON OF CLASSIFIERS

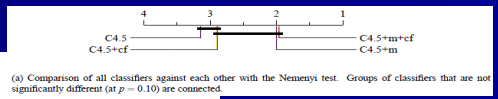
Dataset	NB	SVM	Adaboost	Rand Forest
Anneal	96.43	99.44	83.63	99.55
Audiology	73.42	81.34	46.46	79.15
Balance Scale	72.30	91.51	72.31	80.97
Breast Cancer	71.70	66.16	70.28	69.99
Contact Lenses	71.67	71.67	71.67	71.67
Pima Diabetes	74.36	77.08	74.35	74.88
Glass	70.63	62.21	44.91	79.87
Hepatitis	83.21	80.63	82.54	

$$t = \frac{m_I - m_{II}}{\sqrt{\frac{\sigma_I^2}{k} + \frac{\sigma_{II}^2}{j}}}$$

$$t = \frac{m_d}{\sqrt{\sigma_d^2 / k}}$$

STAC

Statistical Tests for Algorithms Comparison



(a) Comparison of all classifiers against each other with the Nemenyi test. Groups of classifiers that are not significantly different (at $p = 0.10$) are connected.

Journal of Machine Learning Research 7 (2006) 1–30

Statistical Comparisons of Classifiers over Multiple Data Sets

Janez Demšar
Faculty of Computer and Information Science
Tržaška 25
Ljubljana, Slovenia
JANEZ.DEMSAR@FRI.UNI-LJ.SI

Where there are tied groups, take the rank to be equal to

2. U is then given by:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

where n_1 is the sample size for sample 1, and R_1 is the sum of the ranks in sample 1.

Note that it doesn't matter which of the two samples is considered sample 1. An equally valid formula for U is

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}.$$

The smaller value of U_1 and U_2 is the one used when consulting significance tables. The sum of the two values is given by

$$U_1 + U_2 = R_1 - \frac{n_1(n_1+1)}{2} + R_2 - \frac{n_2(n_2+1)}{2},$$

Knowing that $R_1 + R_2 = N(N+1)/2$ and $N = n_1 + n_2$, and doing

$$U_1 + U_2 = n_1 n_2.$$

Journal of Machine Learning Research 9 (2008) 2677-2694 Submitted 10/07; Revised 4/08; Published 12/08

An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons

Salvador García SALVAGL@DECSAI.UGR.ES
Francisco Herrera HERRERA@DECSAI.UGR.ES
Department of Computer Science and Artificial Intelligence
University of Granada
Granada, 18071, Spain

COMPARING 2 CLASSIFIERS IN A DATASET

- Frequent situation: we want to know which one of two learning schemes performs better, has *better accuracy*

- Better accuracy ? *92.8%±3.4% vs 92.2%±2.8%*
  - Apparently? → YES...
  - Statistically? → ¿?¿?

- Key: are the differences *"statistically significant"*?

COMPARING 2 CLASSIFIERS IN A DATASET

- Compare the results of 10-fold CV estimates, or several runs of percentage-split (hold-out) estimates...
- "Compare two samples of numbers"
- Paired? Non-paired? → same partitions in both algorithms?

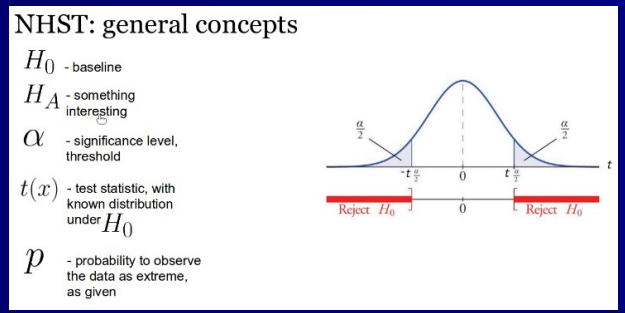
classifier	accuracy
naiveBayes	72.3
naiveBayes	73.4
naiveBayes	71.4
naiveBayes	73.1
naiveBayes	72.3
3-NN	70.8
3-NN	72.0
3-NN	72.4
3-NN	71.9
3-NN	73.4

TOOL
STATISTICAL SIGNIFICANCE TESTS

- Hypothesis significance tests, "simplifying":
  - *Null hypothesis (H_0)*: there is no "real" difference
  - *Alternative hypothesis (H_1)*: there is a difference
    (the algorithm with the best mean accuracy is significantly better)

- Measure → "difference degree"
- Measure → evidence to reject the null hypothesis?

- Test value → follows a known probability distribution



classifier	accuracy
naiveBayes	72.3
naiveBayes	73.4
naiveBayes	71.4
naiveBayes	73.1
naiveBayes	72.3
3-NN	70.8
3-NN	72.0
3-NN	72.4
3-NN	71.9
3-NN	73.4

FAMILIES STATISTICAL TESTS

- Paired *versus* non-paired samples
- Both sets of values obtained from the same randomizations of data?
- Defines the type of test → paired *versus* non-paired

- Defines the type of test → parametric *versus* non-parametric

classifier	accuracy
naiveBayes	72.3
naiveBayes	73.4
naiveBayes	71.4
naiveBayes	73.1
naiveBayes	72.3
3-NN	70.8
3-NN	72.0
3-NN	72.4
3-NN	71.9
3-NN	73.4

classifier	accuracy
naiveBayes	72.3
naiveBayes	73.4
naiveBayes	71.4
naiveBayes	73.1
naiveBayes	72.3
3-NN	70.8
3-NN	72.0
3-NN	72.4
3-NN	71.9
3-NN	73.4

STUDENT'S t-test

- Parametric tests → work with mean and standard deviation values of the accuracy samples
  - *Student's t-test* → means of two samples, m_I and m_{II} (sample I and sample II) are *significantly different ?*

paired

$$t = \frac{m_d}{\sqrt{\sigma_d^2 / k}}$$

non-paired

$$t = \frac{m_I - m_{II}}{\sqrt{\frac{\sigma_I^2}{k} + \frac{\sigma_{II}^2}{j}}}$$

in case of paired samples, m_d represents the mean of the differences between sample I and sample II

- Parametric assumptions (normality) should be fulfilled to apply the test... Kolmogorov-Smirnov test, Shapiro-Wilk, etc.

NON-PARAMETRIC TESTS

No assumption about each sample distribution is assumed

Methods [edit]

Non-parametric (or **distribution-free**) **inferential statistical methods** are mathematical procedures for statistical hypothesis testing which, unlike parametric statistics, make no assumptions about the probability distributions of the variables being assessed. The most frequently used tests include

- Anderson–Darling test
- Cochran's Q
- Cohen's kappa
- Efron–Petrosian test
- Friedman two-way analysis of variance by ranks
- Kendall's tau
- Kendall's W
- Kolmogorov–Smirnov test
- Kruskal-Wallis one-way analysis of variance by ranks
- Kuiper's test
- Mann–Whitney U or Wilcoxon rank sum test
- median test
- Pitman's permutation test
- Rank products
- Siegel–Tukey test
- Spearman's rank correlation coefficient
- Student–Newman–Keuls (SNK) test
- Van Elteren stratified Wilcoxon rank sum test
- Wald–Wolfowitz runs test
- Wilcoxon signed-rank test.

NON-PARAMETRIC TESTS

- Non-parametric tests usually work with *ranks:*
  - After ordering both samples in the same list, a sum of the ranks of each sample in the general-list are assessed → this sum of ranks follows an specific probability distribution which can be used to calculate the significance of the differences

- Mann-Whitney U-test for non paired-samples
- Wilcoxon rank-signed test for paired samples

1. Add up the ranks for the observations which came from sample 1. Where there are tied groups, take the rank to be equal to $N(N+1)/2$ where N is the total number of observations.
2. U is then given by:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

where n_1 is the sample size for sample 1, and R_1 is the sum of the ranks in sample 1.

Note that it doesn't matter which of the two samples is considered sample 1. An equally valid formula for U is

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}.$$

The smaller value of U_1 and U_2 is the one used when consulting significance tables. The sum of the two values is given by

$$U_1 + U_2 = R_1 - \frac{n_1(n_1+1)}{2} + R_2 - \frac{n_2(n_2+1)}{2}.$$

Knowing that $R_1 + R_2 = N(N+1)/2$ and $N = n_1 + n_2$, and doing some algebra, we find that the sum is

$$U_1 + U_2 = n_1 n_2.$$

STATISTICAL SIGNIFICANCE OF A TEST: p-value

- Calculated test → follows a probability distribution
- Each test returns a *p-value* → interpret to asses the significance (degree) of the differences between compared classifiers:
  - *Null hypothesis (H_0)*: there is no "real" difference
  - *Alternative hypothesis (H_1)*: there is a difference
    (the algorithm with the best mean accuracy is significantly better)

p-value interpretation:

- *The risk to reject the null hypothesis (similarity of samples), when it is true*
- *The probability to observe the current differences , assuming that the null hypothesis is true*

p-value "informal" interpretation:

- *Degree of similarity between compared samples*
- *The influence of the chance to explain the observed differences between compared samples*

→ Fix a threshold in the p-value → maintain or discard the Null hypothesis
→ Commonly 0.05 or 0.10 → below this threshold → discard null hypothesis

McNEMAR test
– not based on accuracy –

- Previous tests → based on accuracy

- McNemar test
- Two classifiers' comparison in a dataset
- Single hold-out validation → single test set
- Popular in Deep Learning

- Contingency table from test set
- Different correct and incorrect predictions between both models

	Instance,	Classifier1 Correct,	Classifier2 Correct
2	1	Yes	No
3	2	No	No
4	3	No	Yes
5	4	No	No
6	5	Yes	Yes
7	6	Yes	Yes
8	7	Yes	Yes
9	8	No	No
10	9	Yes	No
11	10	Yes	Yes

	Classifier2 Correct,	Classifier2 Incorrect
Classifier1 Correct	4	2
Classifier1 Incorrect	1	3

McNEMAR test
– not based on accuracy –

- Null hypothesis
- Classifiers disagree in the same amount

- Alternative hypothesis
- Disagreements are "significantly" skewed

- McNemar test → non-parametric + paired

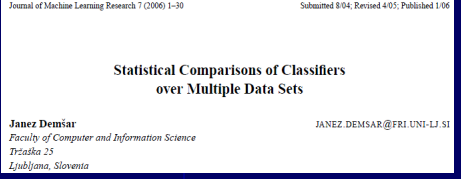
- ~~Similar accuracy of classifiers?~~

	Instance,	Classifier1 Correct,	Classifier2 Correct
1	1	Yes	No
2	2	No	No
3	3	No	Yes
4	4	No	No
5	5	Yes	Yes
6	6	Yes	Yes
7	7	Yes	Yes
8	8	No	No
9	9	Yes	No
10	10	Yes	Yes

	Classifier2 Correct,	Classifier2 Incorrect
Classifier1 Correct	4	2
Classifier1 Incorrect	1	3

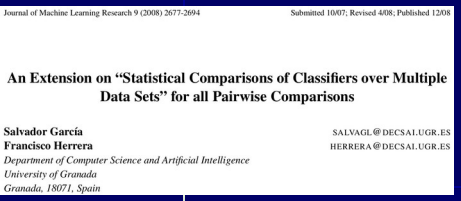
		Model 2		
		Correct	Incorrect	
Model 1	Correct	n_{11}	n_{12}	$n_{1\bullet}$
	Incorrect	n_{21}	n_{22}	$n_{2\bullet}$
		$n_{\bullet 1}$	$n_{\bullet 2}$	n_{test}

$$\chi^2 \sim \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}$$


Journal of Machine Learning Research 7 (2006) 1–30 Submitted 8/04; Revised 4/05; Published 1/06

Statistical Comparisons of Classifiers over Multiple Data Sets

Janez Demšar JANEZ.DEMSAR@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
Tržaška 25
Ljubljana, Slovenia


Journal of Machine Learning Research 9 (2008) 2677-2694 Submitted 10/07; Revised 4/08; Published 12/08

An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons

Salvador García SALVAGL@DECSAI.UGR.ES
Francisco Herrera HERRERA@DECSAI.UGR.ES
Department of Computer Science and Artificial Intelligence
University of Granada
Granada, 18071, Spain

COMPARING MULTIPLE ALGORITHMS IN MULTIPLE DOMAINS

- Until now → comparing two classifiers in a single problem
- A natural extension → comparing multiple classifiers on multiple domains

Dataset	NB	SVM	Adaboost
Anneal	96.43	99.44	83.63
Audiology	73.42	81.34	46.46
Balance Scale	72.30	91.51	72.31
Breast Cancer	71.70	66.16	70.28
Contact Lenses	71.67	71.67	71.67
Pima Diabetes	74.36	77.08	74.35
Glass	70.63	62.21	44.91
Hepatitis	83.21	80.63	82.54
Hypothyroid	98.22	93.58	93.21
Tic-Tac-Toe	69.62	99.90	72.54

COMPARING MULTIPLE ALGORITHMS IN MULTIPLE DOMAINS

How does the hypothesis test work?
- Algorithms **ranked on each domain**
- For each classifier → sum of its ranks in all domains is calculated

dataset	nB	SVM	AdaBoost
Anneal	100.00	99.44	83.63
Audiology	85.34	81.34	46.46
Balance	90.02	91.51	72.31
Tic-Tac	71.70	66.16	70.28
Hepatitis	72.67	71.67	71.67
Glass	79.36	77.08	74.35
Iris	70.63	62.21	44.91
Diabetes	83.21	80.63	82.54
Lenses	98.22	93.58	93.21
Contact	69.62	99.90	72.54

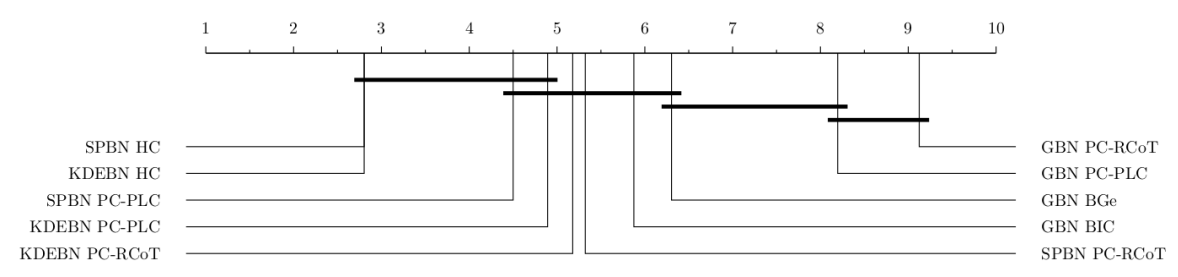
dataset	nB	SVM	AdaBoost
Anneal	1	2	3
Audiology	1	2	3
Balance	2	1	3
Tic-Tac	1	3	2
Hepatitis	1	2.5	2.5
Glass	1	2	3
Iris	1	2	3
Diabetes	1	3	2
Lenses	1	2	3
Contact	3	1	2
Ranks' Sum	13	18	24

COMPARING MULTIPLE ALGORITHMS IN MULTIPLE DOMAINS

- First: use of non-parametric Friedman's test:
  - Null hypothesis → all the classifiers perform similarly
  - Alternative hypothesis → there exists at least one pair of classifiers with significantly different performances

- In case of *rejection* of this null hypothesis →
  → "post-hoc bivariate tests" to *identify different pairs of classifiers*

- Web application STAC
- R package SCMAMP: [in R Journal] [authors' GitHub]
- R package PMCMRplus: [in RDocumentation]
- R package EXREPORT: [vignette] [testMultiplePairwise( ) function]



COMPARING MULTIPLE ALGORITHMS IN MULTIPLE DOMAINS

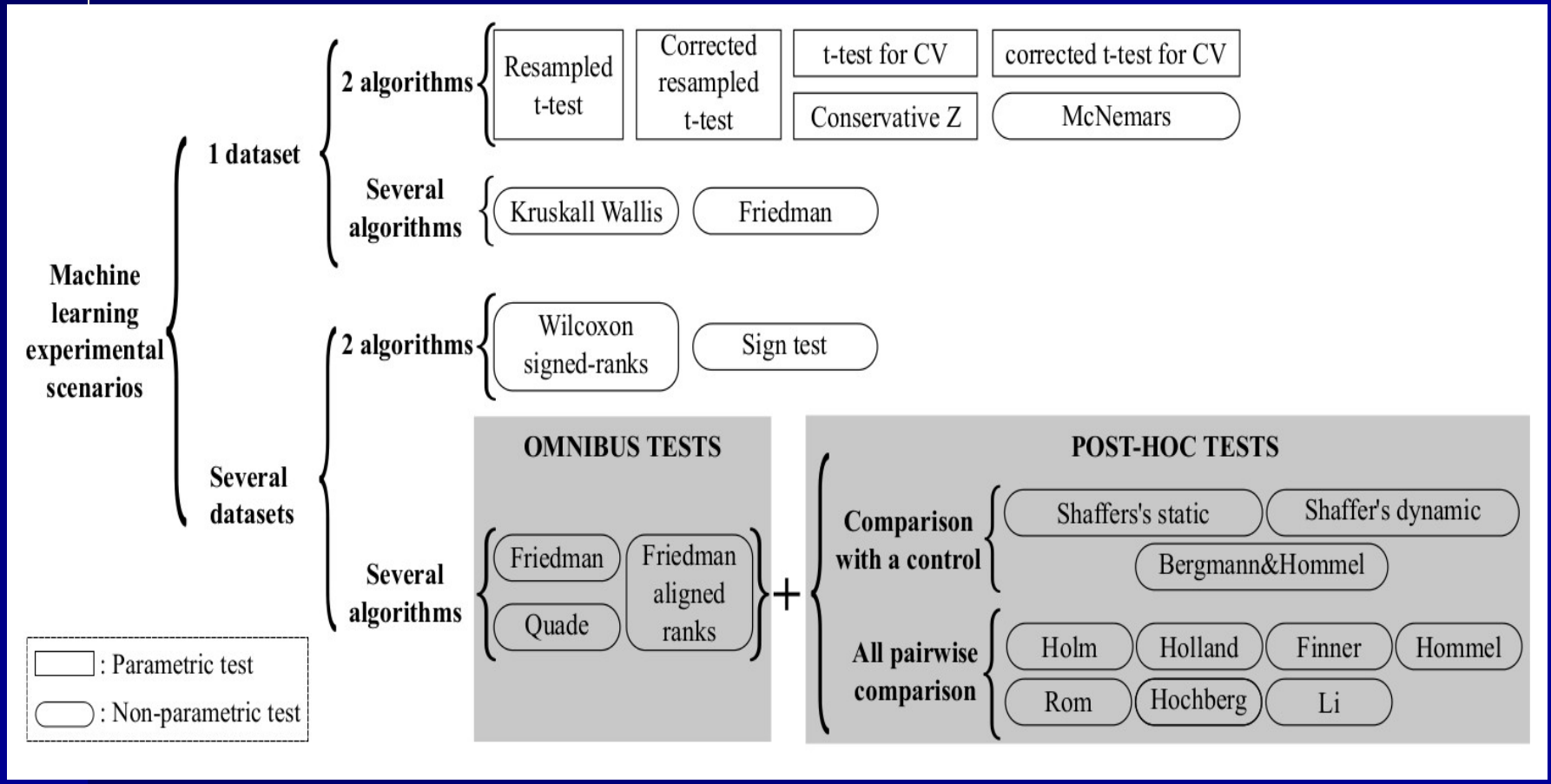
```
install.packages("PMCMRplus")
library(PMCMRplus)
classifiersResults <- matrix(c(100,85.34,90.02,71.70,72.67,79.36,70.63,83.21,98.22,69.62,
                               99.44,81.34,91.51,66.16,71.67,77.08,62.21,80.63,93.58,99.90,
                               83.63,46.46,72.31,70.28,71.67,74.35,44.91,82.54,93.21,72.54),
                             nrow=10, ncol=3,dimnames=list(1:10,c("nB","SVM","AdaBoost")))
print(classifiersResults)
friedman.test((classifiersResults))
frdAllPairsNemenyiTest(classifiersResults)
```

```
        Friedman rank sum test

data:  (classifiersResults)
Friedman chi-squared = 9.3846, df = 2, p-value = 0.009166
```

	nB	SVM
SVM	0.2140	-
AdaBoost	0.0072	0.3721

STATISTICAL TESTS' OVERVIEW



Artif Intell Rev (2015) 44:467–508
DOI 10.1007/s10462-015-9433-y

CrossMark

Dealing with the evaluation of supervised classification algorithms

Guzman Santafe[1] · Iñaki Inza[2] · Jose A. Lozano[2]

ARE WE DOING THINGS OK?

Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis

Alessio Benavoli[†] ALESSIO@IDSIA.CH
Giorgio Corani[†] GIORGIO@IDSIA.CH
Janez Demšar[♮] JANEZ.DEMSAR@FRI.UNI-LJ.SI
Marco Zaffalon[†] ZAFFALON@IDSIA.CH

[♮] *Faculty of Computer and Information Science, University of Ljubljana, Vecna pot 113, SI-1000 Ljubljana, Slovenia*
[†] *Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA) Galleria 2, 6928 Manno, Switzerland*

Editor: David Barber

Abstract

The machine learning community adopted the use of null hypothesis significance testing (NHST) in order to ensure the statistical validity of results. Many scientific fields however realized the shortcomings of frequentist reasoning and in the most radical cases even banned its use in publications. We should do the same: just as we have embraced the Bayesian paradigm in the development of new machine learning methods, so we should also use it in the analysis of our own results. We argue for abandonment of NHST by exposing its fallacies and, more importantly, offer better—more sound and useful—alternatives for it.

ARE WE DOING THINGS OK?

Durante los últimos años estamos ante un incipiente "run-run" sobre la utilidad de los tests estadísticos para la comparativa de modelos. Y mi olfato me dice que el siguiente trabajo, publicado en una prestigiosa revista de nuestro campo por autores altamente reconocidos en el campo, puede empezar a ofrecer otras alternativas: "Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis". Para los que os atraiga este campo de la comparativa de clasificadores, altamente recomendable: su estilo de escritura, como el título cita, es "tutorial".

Trataré de resumir sus principales items. Por defecto, estamos en un escenario de comparativa de dos modelos en un dataset, donde se dispone para cada modelo de los 100 accuracies resultantes de una validación de 10 repeticiones de 10-fold cross-validation:

- los tests de hipótesis no responden a la pregunta que tiene el experimentador en mente, esto es, si Model1 es mejor que Model2, o Model1 es peor que Model2, o Model1 y Model2 tienen un performance similar
- esto es, el *p*-value devuelto por el test de hipótesis representa la probabilidad de conseguir las diferencias observadas (o mayores), asumiendo que la hipótesis nula sobre la similitud en el comportamiento de los modelos comparados es cierta, informalmente redactado, p(diferencias_observadas|H_0). Y ésta no es la pregunta que queremos responder...
- los autores proponen, como ellos le llaman, "a colour thinking". Mediante un framework de análisis Bayesiano que incluye las distribuciones a priori de los parámetros implicados (de ahí el nombre de "Bayesian statistics" a los tests que proponen), llegan a formular un modelo para la distribución a-posteriori de la diferencia media entre ambos modelos comparados: ésta se puede representar gráficamente y se puede consultar e integrar.
- consultando esta distribución a-posteriori de la diferencia media entre ambos modelos se puede evaluar directamente la probabilidad de la hipótesis que tiene el experimentador en mente, esto es, p(Model1>Model2), p(Model2>Model1) y p(Model1=Model2).
- este último escenario de "igualdad" se debe matizar, y se debe definir a qué nos referimos por "prácticamente equivalentes". Los autores trabajan en el escenario de que el performance de "dos clasificadores es prácticamente equivalente" cuando su diferencia media es menor que el 1%. Este intervalo definirá la probabilidad de la "práctica equivalencia de dos modelos", p(Model1=Model2). A esta diferencia, clave para comprender la propuesta, los autores la denominan como "rope".
- los autores proponen una versión "Bayesian statistic" del popular t-test para muestras pareadas (i.e. los 100 accuracies de la 10x10foldCrossValidation parten de las mismas particiones de casos para ambos modelos) y de su compañero no-paramétrico, Wilcoxon signed-rank test. Para este segundo no se llega a una fórmula cerrada de las 3 distribuciones a computar, y se estiman por el método de simulación de Monte Carlo.
- los autores también proponen dos formulaciones de "Bayesian statistics" para la comparativa de dos algoritmos en varios datasets.
- los autores ofrecen el código que implementa los "Bayesian statistics" propuestos en varios lenguajes. https://github.com/BayesianTestsML/tutorial/
- el paquete de R "scmamp: statistical comparison of multiple algorithms in multiple problems", recoge la implementación de estos tests, así como unas ilustrativas viñetas que ayudan a comprenderlos, aparte de a usarlos.