I Workshop on Knowledge Extraction based on Evolutionary Learning
15-16 May, 2008

KEEL

A study of statistical techniques and performance of GBML

Julián Luengo

Dpt. Computer Science and Artificial Intelligence
University of Granada
18071 – Spain

In collaboration with
Francisco Herrera, Salvador García and Alberto Fernández

DECSAI
Universidad de Granada

A study of statistical techniques and performance of GBML

- Genetic-Based Machine Learning
- Inferential Statistics
- Analyzing parametric tests
- Comparing two algorithms: From parametric to non-parametric tests
- Multiple comparisons: Non-parametric tests
- Lessons learned

A study of statistical techniques and performance of GBML

- **Genetic-Based Machine Learning**
- **Inferential Statistics**
- **Analyzing parametric tests**
- **Comparing two algorithms: From parametric to non-parametric tests**
- **Múltiple comparisons: Non-parametric tests**
- **Lessons learned**

Genetic-Based Machine Learning

- They are Evolutionary Rule-based systems
  - So called Genetic-Based Machine Learning (GBML)
- Claimed advantages:
  - Interpretable models
  - No assumption of prior relationships among attributes
  - Possibility of obtaining compact and precise rule sets
- Some proposed GBMLs: GABIL, SIA, XCS, DOGMA and JoinGA, G-Net, UCS, GASSIST, OCEC and HIDER.

Genetic-Based Machine Learning

- We have chosen four Genetic Interval Rule Based Algorithms:
  - *Pittsburgh Genetic Interval Rule Learning Algorithm.*
  - *XCS Algorithm.*
  - *GASSIST Algorithm.*
  - *HIDER Algorithm.*
- GBML will be analyzed by two performance measures: *Accuracy* and *Cohen's kappa*.

- **How we state which is the best?**

Genetic-Based Machine Learning

Experimental Study

- We have selected 14 data sets from UCI repository.

Data set	#Ex.	#Atts.	#C.
bupa (bup)	345	6	2
cleveland (cle)	297	13	5
ecoli (eco)	336	7	8
glass (gla)	214	9	7
haberman (hab)	306	3	2
iris (iri)	150	4	3
monk-2 (mon)	432	6	2
new-Thyroid (new)	215	5	3
pima (pim)	768	8	2
vehicle (veh)	846	18	4
vowel (vow)	988	13	11
wine (win)	178	13	3
wisconsin (wis)	683	9	2
yeast (yea)	1484	8	10

Experiments Design in Data Mining: Using non-parametric tests

- Genetic-Based Machine Learning
- **Inferential Statistics**
- Analyzing parametric tests
- Comparing two algorithms: From parametric to non-parametric tests
- Múltiple comparisons: Non-parametric tests
- Lessons learned

Inferential Statistics

Inferential Statistics - provide measures of how well your data support your hypothesis and if your data are generalizable beyond what was tested (*significance tests*)

In our case: Comparing two or various sets of experiments with the GBMLs

Parametric versus Nonparametric Statistics – When to use them and which is more powerful?

Inferential Statistics

Parametric Assumptions

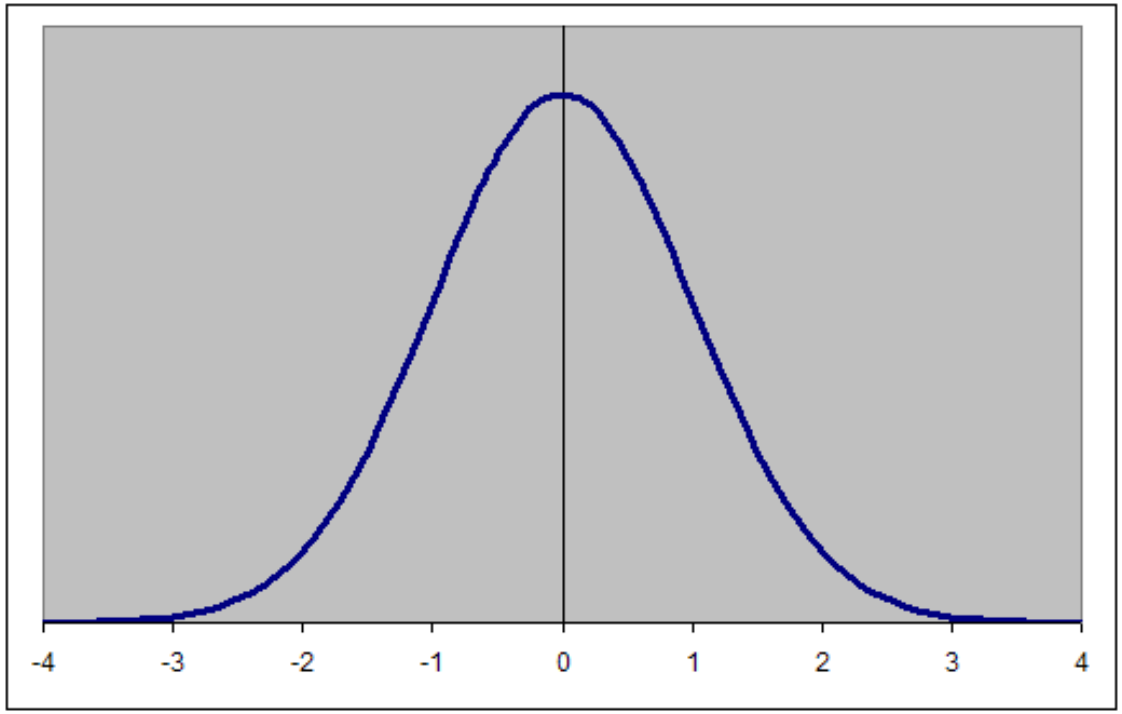
- The observations must be **independent**
- **Normality**: The observations must be drawn from normally distributed populations

  (Tests: Kolmogorov-Smirnov, Shapiro-Wilk, D'Agostino-Pearson)
- **Homoscedasticity**: These populations must have the same variances

  (Levene's Test)

Inferential Statistics

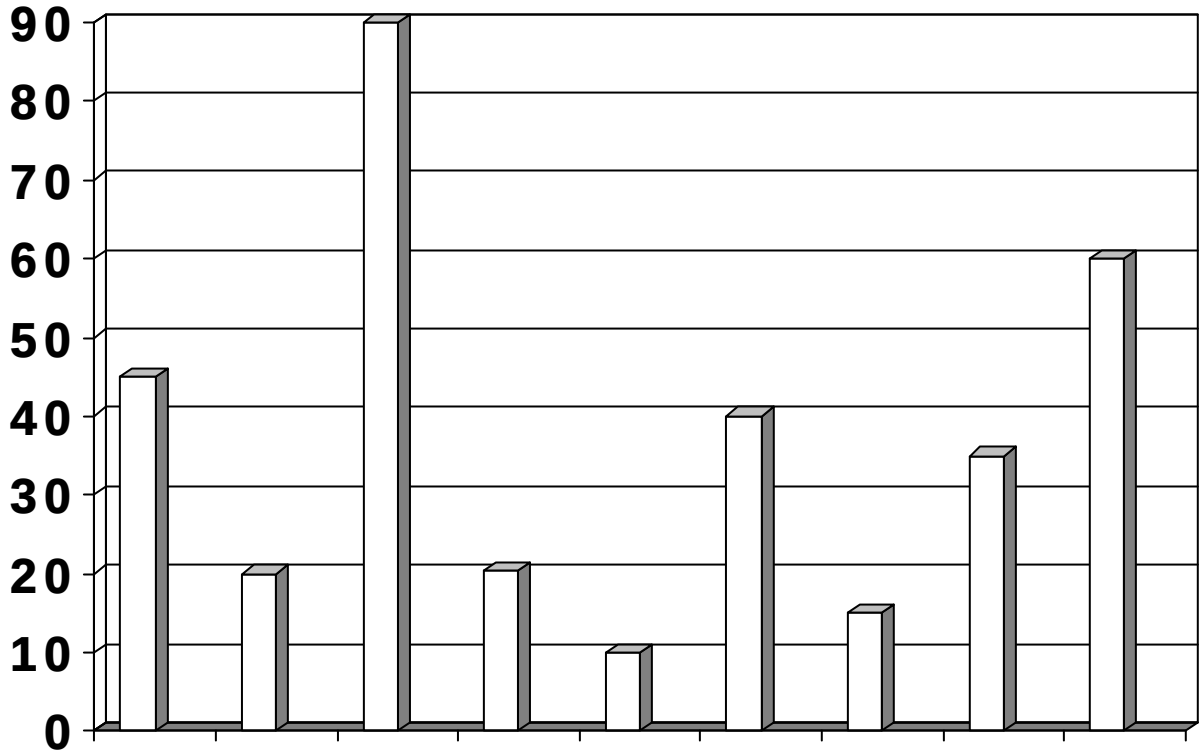
If your data looks like this, you can do a parametric test!



The graph has a mean of zero and a standard deviation of 1, i.e., (μ=0, σ=1).

Inferential Statistics

If your data looks like this, don't do a parametric test!



Histogram

Inferential Statistics

Nonparametric Assumptions

- Observations are independent (unpaired test) or not (paired test)
- Data represented in an ordinal way of ranking.

How do nonparametric tests work?

- Most nonparametric tests use ***ranks*** instead of raw data for their hypothesis testing.
- The cases of test are used for getting the average rank.

Inferential Statistics (basic concepts)

p-values

- p-values for all tests tell us whether or not to reject the null hypothesis (and with what confidence)
- In linguistic research, a confidence level of 95% is often sufficient, some use 99%
- This decision is up to you. Note that the more stringent your confidence level, the more likely is a type II error (you don't find a difference that is actually there)

Inferential Statistics (basic concepts)

p-values

- If you decide for a p-value of 0.05 (95% certainty that there indeed is a significant difference), then a value smaller than 0.05 indicates that you can reject the **null-hypothesis**
- **Remember:** the null-hypothesis generally predicts that there is no difference
- So, in a test, if you have p = 0.07 means that you **cannot reject** the null hypothesis that "there is no difference"
  ⇨ **there is no significant difference between the two groups**

Inferential Statistics (basic concepts)

There is at least one nonparametric test equivalent to a parametric test

- Compare two variables (i.e. two GBMLs)
- If more than two variables (i.e. a group of GBMLs)

Parametric	Nonparametric
t-test	Sign test
	Wilcoxon's signed rank test
ANOVA	Friedman's test Iman and Davenports' test
Turkey, Tamhane, ...	Bonferroni-Dunn's test Holm's method

Inferential Statistics (basic concepts)

Advantages of Nonparametric Tests

- Can treat data which are inherently in ranks as well as data whose seemingly numerical scores have the strength in ranks
- Easier to learn and apply than parametric tests

  (only one run for all cases of test)

If sample sizes as small as N=6 are used, there is no alternative to using a nonparametric test

Inferential Statistics (basic concepts)

Criticisms of Nonparametric Procedures

- Losing precision/wasteful of data
- Low power
- False sense of security
- Lack of software
- Testing distributions only
- Higher-ordered interactions not dealt with

Inferential Statistics (basic concepts)

What happens if I use a nonparametric test when the data is normal?

- It will work, but a parametric test would be more powerful, i.e., give a lower p value.
- If the data is not normal, then the nonparametric test is usually more powerful
- Always look at the data first, then decide what test to use. Don't choose the one that gives the answer you want. Integrity is essential

Inferential Statistics (basic concepts)

- We distinguish between two types of analysis: *single data set* analysis and *multiple data set analysis.*
- Central Limit Theorem for classification performance is rarely held
  - It depends on the case of the problem studied and the number of runs of the algorithm.
  - an excessive number of runs affects negatively
- Thus, we use single data set analysis.

A study of statistical techniques and performance of GBML

- Genetic-Based Machine Learning
- Inferential Statistics
- **Analyzing parametric tests in our study**
- Comparing two algorithms: From parametric to non-parametric tests
- Múltiple comparisons: Non-parametric tests
- Lessons learned

Analyzing parametric tests

- Parametric statistical test are well suited for the GBMLs?
  - Remember: we use the Accuracy and Cohen's Kappa
- We study the needed parametric tests conditions with the results samples obtained running the algorithms several times.
  - We use a 10-fcv
  - Each partition is executed 5 times with different seeds
  - We obtain 50 results of both performance measures.

Analyzing parametric tests

In order to use the parametric tests, is necessary to check the following conditions:

Independence: In statistics, two events are independent when the fact that one occurs does not modify the probability of the other one occurring.

- **When we compare two optimization algorithms they are usually independent.**
- **When we compare two machine learning methods, it depends on the partition:**
  - **The independency is not truly verified in 10-fcv (a portion of samples is used either for training and testing in different partitions.**
  - **Hold out partitions can be safely take as independent, since training and test partitions do not overlap.**

Analyzing parametric tests

Parametric tests assume that the data are taken from normal distributions

Normality: An observation is normal when its behaviour follows a normal or Gauss distribution with a certain value of average μ and variance σ. A normality test applied over a sample can indicate the presence or absence of this condition in observed data.

- **Kolmogorov-Smirnov**
- **Shapiro-Wilk**
- **D'Agostino-Pearson**

GBML Case of Study

TABLE I. Normality condition in accuracy

Shapiro-Wilk														
	bup	cle	eco	gla	hab	iri	mon	new	pim	veh	vow	win	wis	yea
Pitts-GIRLA	* (.02)	* (.00)	* (.00)	(.73)	* (.00)	* (.00)	* (.00)	* (.01)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)
XCS	(.25)	* (.03)	(.23)	* (.00)	* (.02)	* (.00)	* (.00)	* (.00)	* (.03)	(.17)	(.30)	* (.00)	* (.00)	(.45)
GASSIST	(.39)	(.21)	(.07)	(.19)	* (.04)	* (.00)	(.07)	* (.00)	(.12)	(.81)	(.51)	* (.00)	* (.00)	(.83)
HIDER	(.11)	(.42)	(.22)	* (.00)	* (.01)	* (.00)	(.06)	* (.00)	* (.00)	(.25)	(.15)	* (.00)	* (.00)	(.23)

D'Agostino-Pearson														
	bup	cle	eco	gla	hab	iri	mon	new	pim	veh	vow	win	wis	yea
Pitts-GIRLA	(.13)	(.10)	* (.00)	(.69)	* (.00)	(.11)	* (.00)	(.71)	* (.00)	* (.02)	* (.00)	* (.00)	* (.00)	* (.00)
XCS	(.44)	(.09)	(.61)	(.06)	(.22)	(.06)	* (.00)	* (.00)	(.24)	(.33)	(.40)	* (.00)	* (.03)	(.48)
GASSIST	(.55)	(.75)	(.59)	(.42)	(.79)	(.19)	(.89)	(.89)	(.25)	(.65)	(.18)	* (.03)	* (.03)	(.95)
HIDER	(.07)	(.52)	(.42)	(.05)	(.78)	* (.00)	(.19)	* (.00)	* (.00)	(.43)	(.37)	* (.00)	* (.02)	(.18)

a value smaller than 0.05 indicates that you can reject the **null-hypothesis** (i.e. the normality condition is not satisfied) and it is noted with "*"

GBML Case of Study

TABLE II. Normality condition in Cohen's Kappa

Shapiro-Wilk														
	bup	cle	eco	gla	hab	iri	mon	new	pim	veh	vow	win	wis	yea
Pitts-GIRLA	* (.00)	* (.02)	* (.00)	(.79)	* (.00)	* (.00)	* (.00)	* (.04)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)
XCS	(.65)	(.11)	(.37)	* (.00)	* (.01)	* (.00)	* (.00)	* (.00)	* (.04)	(.17)	(.30)	* (.00)	* (.00)	(.51)
GASSIST	(.26)	(.45)	(.18)	(.15)	(.14)	* (.00)	* (.00)	* (.00)	(.17)	(.81)	(.16)	* (.00)	* (.00)	(.76)
HIDER	(.61)	(.42)	(.21)	* (.00)	* (.01)	* (.00)	* (.00)	* (.00)	* (.01)	(.23)	(.56)	* (.00)	* (.00)	(.20)

D'Agostino-Pearson														
	bup	cle	eco	gla	hab	iri	mon	new	pim	veh	vow	win	wis	yea
Pitts-GIRLA	* (.00)	(.49)	* (.00)	(.58)	* (.00)	(.11)	* (.00)	(.80)	* (.00)	* (.01)	* (.01)	* (.00)	* (.00)	* (.00)
XCS	(.54)	(.41)	(.72)	(.06)	* (.03)	(.06)	* (.00)	* (.00)	(.27)	(.32)	(.40)	* (.01)	* (.04)	(.35)
GASSIST	(.45)	(.59)	(.57)	(.40)	(.29)	(.19)	* (.02)	* (.01)	(.26)	(.62)	(.18)	* (.03)	* (.03)	(.88)
HIDER	(.33)	(.45)	(.43)	(.05)	(.21)	* (.00)	* (.00)	* (.02)	* (.00)	(.41)	(.38)	* (.00)	* (.01)	(.20)

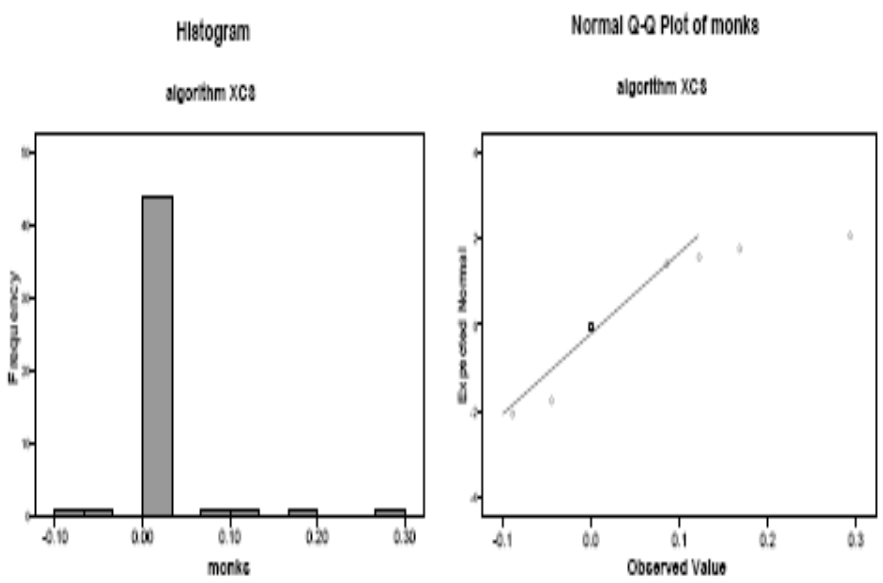
a value smaller than 0.05 indicates that you can reject the **null-hypothesis** (i.e. the normality condition is not satisfied) and it is noted with "*"

GBML Case of Study: some facts

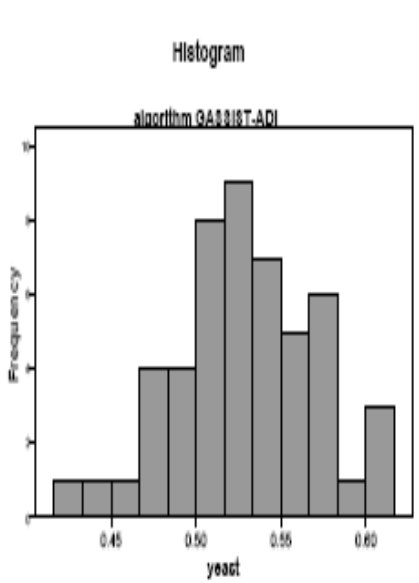
- Conditions needed for the application of parametric tests are not fulfilled in some cases.
  - The size of the sample should be enough (50)
- One main factor: the nature of the problem
- Graphically, we can use Q-Q graphics and histograms to see the normality

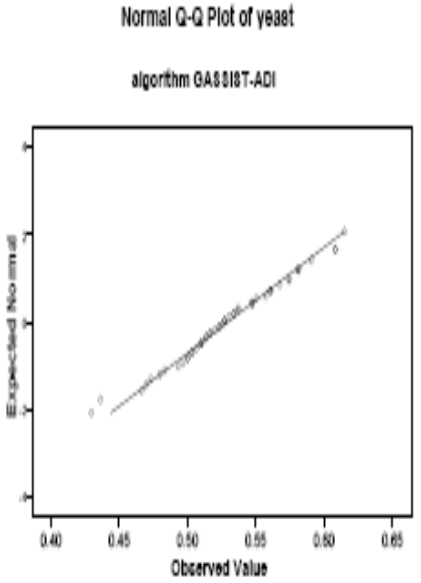
Analyzing parametric tests

Fig. 1. Breast problem: Histogram and Q-Q Graphic.



Total lack of normality for XCS Kappa





Normality accepted by both tests in GASSIST Accuracy

* A Q-Q graphic represents a confrontation between the quartiles from data observed and those from the normal distributions.

Analyzing parametric tests

Heterocedasticity: This property indicates the existence of a violation of the hypothesis of equality of variances.

Levene's test is used for checking if k sampls present or not this homogeneity of variances (homocesdasticity).

GBML Case of Study

TABLE IV. Test of HETEROSCEDASTICITY OF LEVENE (BASED ON MEANS)

	bup	cle	eco	gla	hab	iri	mon	new	pim	veh	vow	win	wis	yea
Accuracy	(.13)	* (.00)	(.36)	(.34)	* (.01)	(.40)	* (.00)	(.26)	(.16)	* (.00)	* (.03)	* (.00)	* (.00)	* (.00)
Cohen's kappa	(.51)	(.05)	(.39)	(.25)	* (.04)	(.40)	* (.00)	(.40)	* (.00)	* (.00)	* (.03)	* (.00)	* (.00)	* (.00)

Table IV shows the results by applying Levene's tests, where the symbol "*" indicates that the variances of the distributions of the different algorithms for a certain function are not homogeneities (we reject the null hypothesis).

A study of statistical techniques and performance of GBML

- Genetic-Based Machine Learning
- Inferential Statistics
- Analyzing parametric tests
- **Comparing two algorithms: From parametric to non-parametric tests**
- Múltiple comparisons: Non-parametric tests
- Lessons learned

Two-Sample tests

Two-Sample Tests

When comparing means of two samples to make inferences about differences between two populations, there are 4 main tests that could be used:

	Unpaired data	Paired data
Parametric test	Independent-Samples T-Test	Paired-Samples T-Test
Non-parametric test	Mann-Whitney U test (or Wilcoxon rank-sum test)	Wilcoxon Signed-Ranks test (*Also, Sign test*)

Two-Sample tests

Wilcoxon Signed-Ranks Test for Paired Samples

The Wilcoxon Signed-Ranks test is used in exactly the same situations as the paired t-Test (i.e., where data from two samples are paired).

In general the Test asks:

H_o: The 2 samples come from populations with the same distributions.
Or, median of population 1 = median of population 2

H_1: The 2 samples come from populations with different distributions
Or, median of population 1 ≠ median of population 2

The test statistic is based on ranks of the differences between pairs of data.

NOTE: If you have ≤ 5 pairs of data points, the Wilcoxon Signed-Ranks test can never report a 2-tailed p-value < 0.05

Two-Sample tests

Procedure for the Wilcoxon Signed-Ranks Test

1. For each pair of data, calculate the difference. Keep track of the sign (+ve or –ve).

2. Temporarily ignoring the sign of the difference, rank the absolute values of the difference. When the differences have the same value, assign them the mean of the ranks involved in the tie.

3. Consider the sign of the differences again and ADD up the ranks of all the positive differences and all the negative differences (R^+, R^-). Ranks of difference equal to 0 are split evenly among the sums; if there is an odd number of them, one is ignored.

Two-Sample tests

Procedure for the Wilcoxon Signed-Ranks Test

4. Let T be the **smaller** of the sums of positive and negative differences. $T = \text{Min}\{R^+, R^-\}$.
Use an appropriate Statistical Table or computer to determine the test statistic, critical region or P-values.

5. Reject the H_o if test statistic $\leq$ critical value, or if $P \leq \alpha$ (alpha).

6. Report Test results.

Two-Sample tests

For $n \leq 30$: use T values (and refer to a Table B.12. Critical Values of the Wilcoxon T Distribution, Zar, App101)

For $n > 30$: use z-scores (z is distributed approximately normally).

(and refer to the z-Table, Table B.2. Zar – Proportions of the Normal Curve (One-tailed), App 17)

where,

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

With $\alpha = 0.05$, the null-hypothesis can be rejected if z is smaller than –1.96.

Two-Sample tests

Wilcoxon Signed-Ranks Test for Paired Samples

n	LEVEL OF SIGNIFICANCE FOR ONE-TAILED TEST		
	0.025	0.01	0.005
	LEVEL OF SIGNIFICANCE FOR TWO-TAILED TEST		
	0.05	0.02	0.01
6	0	—	—
7	2	0	—
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Critical value for T for N up to 25.

It T <= (table-value) then Reject the H_o

GBML Case of Study

Wilcoxon Signed-Ranks Test for Paired Samples

Wilcoxon's test applied over the all possible comparisons between the 4 algorithms in accuracy

	Classification rate		
Comparison	R^+	R^-	p-value
Pitts-GIRLA - **XCS**	0.5	104.5	0.001
Pitts-GIRLA - **GASSIST-ADI**	2	103	0.002
Pitts-GIRLA - **HIDER**	1	104	0.001
Pitts-GIRLA - **CN2**	6	99	0.004
XCS - GASSIST-ADI	81	24	0.074
XCS - HIDER	53	52	0.975
XCS - CN2	78	27	0.109
GASSIST-ADI - **HIDER**	13	92	0.013
GASSIST-ADI - CN2	57	48	0.778
HIDER - CN2	100	5	0.003

We stress in **bold** the winner algorithm in each row when the *p-value associated is below 0.05*

GBML Case of Study

Wilcoxon Signed-Ranks Test for Paired Samples

Wilcoxon's test applied over the all possible comparisons between the 4 algorithms in kappa

	Cohen's kappa		
Comparison	R^+	R^-	p-value
Pitts-GIRLA - **XCS**	0.5	104.5	0.001
Pitts-GIRLA - **GASSIST-ADI**	0	105	0.001
Pitts-GIRLA - **HIDER**	0	105	0.001
Pitts-GIRLA - **CN2**	10	95	0.008
XCS - GASSIST-ADI	78	27	0.109
XCS - HIDER	51	54	0.925
XCS - CN2	78	27	0.109
GASSIST-ADI - HIDER	23	82	0.064
GASSIST-ADI - CN2	60	45	0.638
HIDER - CN2	96	9	0.006

We stress in **bold** the winner algorithm in each row when the *p-value associated is below 0.05*

Two-Sample tests

- We should **not** try to extract from previous tables a conclusion which involves more than one comparison
  - We are losing control on the family-wise error rate (FWER)
  - An associated error that grows agreeing with the number of comparisons done

Two-Sample tests

Using Wilcoxon test por comparing multiple pairs of algorithms.

Given that this test carries out comparisons of pairs of algorithms in an independent way, the overall significance level is not controlled. The family-wise error rate (FWER) incrase. The true statistical signification for combining pairwise comparison test is given by:

$$
\begin{aligned}
p = P(Reject\ H_0 | H_0\ true) = \\
= 1 - P(Accept\ H_0 | H_0\ true) = \\
= 1 - P(Accept\ A_k = A_i, i = 1, \ldots, k-1 | H_0\ true) = \\
= 1 - \prod_{i=1}^{k-1} P(Accept\_A_k = A_i | H_0\ true) = \\
= 1 - \prod_{i=1}^{k-1} [1 - P(Reject\ A_k = A_i | H_0\ true)] = \\
= 1 - \prod_{i=1}^{k-1} (1 - p_{H_i})
\end{aligned}
$$

A study of statistical techniques and performance of GBML

- Genetic-Based Machine Learning
- Inferential Statistics
- Analyzing parametric tests
- Comparing two algorithms: From parametric to non-parametric tests
- **Multiple comparisons: Non-parametric tests**
- Lessons learned

Multiple Comparisons: When?

- A new proposal of GBML should be compared with existing methods.
- Pairwise comparisons could be used, but experiment wise error cannot be controlled.
- Multiple Comparisons procedures are designed for allowing us to fix the FWER before performing the analysis.
  - They also take into account all the influences that can exist within the set of results for each algorithm

Multiple Comparisons

Parametric	Nonparametric
ANOVA	Friedman's test Iman-Davenport's test
Turkey, Dunnet, …	Bonferroni-Dunn's test Holm's method Hochberg's method

Multiple Comparisons

Friedman's test: It is a non-parametric equivalent of the test of repeated-measures ANOVA. It computes the ranking of the observed results for algorithm (r_j for the algorithm j with k algorithms) for each function/algorithm, asisignng to the best of them the ranking 1, and to the worst the ranking k. Under the null hypothesis, formed fromsupposing that the results of the algorithms are equivalent and, therefore, their rankings are also similar, the Friedman statistic

r_j

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

is distributed according to χ_F^2 with $k - 1$ degrees of freedom, being , $R_j = \frac{1}{N}\sum_i r_i^j$ and N the number of functions/algorithms. (N > 10, k > 5)
(Table B.1. Critical Values of the Chi-Square Distribution, App. 12, Zar).

Multiple Comparisons

Iman and Davenport's test: It is a metric derived from the Friedman's statistic given that this last metric produces a conservative undesirably effect. The statistic is:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

and it is distributed acording to a F distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom.
(Table B.4. Critical values of the F Distribution, App. 21, Zar).

GBML Case of Study

Results of applying Friedman's and Iman-Davenport's test with level of significance $\alpha \leq 0.05$ to the GBMLs

	Friedman Value	Value in χ^2	Iman-Davenport Value	Value in F_F
Accuracy	**28.286**	9.487	**13.268**	2.55
Cohen's kappa	**27.186**	9.487	**12.265**	2.55

- The statistics of Friedman and Iman-Davenport are clearly greater than their associated critical values
  - There are significant differences among the observed results
- Next step: apply **post-hoc** test and find what algorithms partners' average results are dissimilar

Multiple Comparisons

Bonferroni-Dunn's test: If the null hypothesis is rejected in any of the previous tests, we can proceed with a posteriori test. The test of Bonferroni-Dunn is similar to Dunnet's test for ANOVA and it is used when we want to compare a control algorithm opposite to the remainder. The quality of two algorithms is significantly different if the corresponding average of rankings is at least as great as its critical difference.

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

.

The value of q_α is the critical value of Q' for a multiple non-parmetric comparison with a control (Table B.16. Critical Values of Q' for Nonparametric Multiple Comparison Testing with a Control, App108, Zar and Table 5.b of the following slide).

Multiple Comparisons

#classifiers	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.241	2.394	2.498	2.576	2.638	2.690	2.724	2.773
$q_{0.10}$	1.645	1.960	2.128	2.241	2.326	2.394	2.450	2.498	2.539

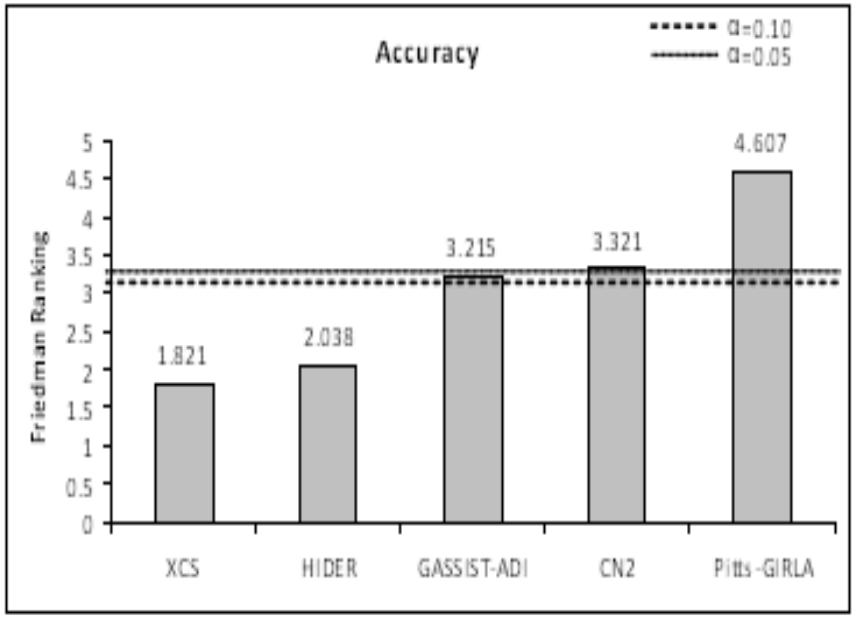
(b) Critical values for the two-tailed Bonferroni-Dunn test; the number of classifiers include the control classifier.

Table 5: Critical values for post-hoc tests after the Friedman test

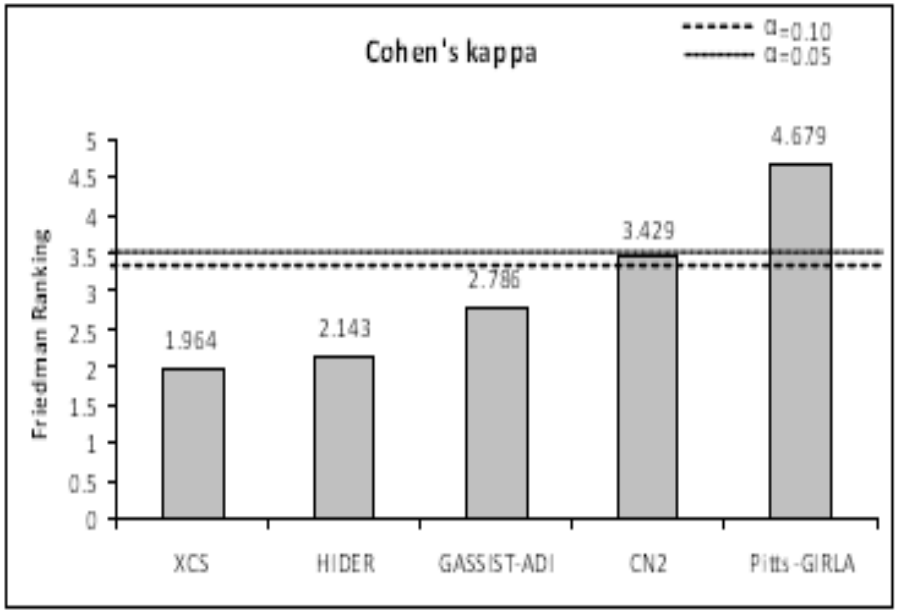
Source: Demsar, J., Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research. Vol. 7. pp. 1–30. 2006.

GBML Case of Study

Bonferroni-Dunn's test: CD = 1.493 and CD= 1.34 for *a = 0.05 and a = 0.10 respectively in the* two measures considered.



Bonferroni-Dunn's graphic for accuracy



Bonferroni-Dunn's graphic for kappa

Multiple Comparisons

Holm's method: For contrasting the procedure of Bonferroni-Dunn, we dispose of a test that sequentially checks the hypothesis ordered according to their significance. We will denote the p values ordered: $p_1 \leq p_2 \leq \ldots \leq p_{k-1}$.

Holm's method compares each p_i with $\alpha/(k-i)$ starting from the most significant p value. If p_1 Is below than $\alpha/(k-1)$, the corresponding hypothesis is rejected and it leaves us to compare p_2 with $\alpha/(k-2)$. If the second hypothesis is rejected, we continue with the process. As soon as a certain hypothesis cannot be rejected, all the remaining hypothesis are maintained as accepted. The statistic for comparing the *i* algorithm with the *j* algorithm is:

$$z = (R_i - R_j) \Big/ \sqrt{\frac{k(k+1)}{6N}} .$$

The value of z is used for finding the corresponding probability from the table of the nomal distribution, which is compared with the corresponding value of α .
(Table B.2. Zar – Proportions of the Normal Curve (One-tailed), App 17)

Multiple Comparisons

Hochberg's method: It is a step-up procedure that works in the opposite direction to Holm's method, comparing the largest *p* value with α, the next largest with $\alpha/2$ and so forth until it encounters a hypotheisis it can reject. All hypotheses with smaller p values are then rejected as well.

Hochberg's method is more powerful than Holm's although it may under some circumstances exceed the family-wise error.

GBML Case of Study

Adjusted *p*-values for the comparison of the control algorithm in each measure with the remaining algorithms

Accuracy (XCS is the control)					
i	algorithm	unadjusted p	p_{Bonf}	p_{Holm}	p_{Hoch}
1	**Pitts-GIRLA**	$3.141 \cdot 10^{-6}$	$1.256 \cdot 10^{-5}$	$1.256 \cdot 10^{-5}$	$1.256 \cdot 10^{-5}$
2	**CN2**	0.01207	0.04830	0.03622	0.03622
3	**GASSIST-ADI**	0.01977	0.07908	0.03954	0.03954
4	HIDER	0.71992	1.00000	0.71992	0.71992

Cohen's kappa (XCS is the control)					
i	algorithm	unadjusted p	p_{Bonf}	p_{Holm}	p_{Hoch}
1	**Pitts-GIRLA**	$5.576 \cdot 10^{-6}$	$2.230 \cdot 10^{-5}$	$2.230 \cdot 10^{-5}$	$2.230 \cdot 10^{-5}$
2	**CN2**	0.01428	0.05711	0.04283	0.04283
3	GASSIST-ADI	0.16928	0.67713	0.33857	0.33857
4	HIDER	0.76509	1.00000	0.76509	0.76509

- If the adjusted *p* for each method is lower than the desired level of confidence α (0.05 in our case), the algorithms are worse from bottom to top (stress in bold for 0.05)
- In practice, Hochberg's method is more powerful than Holm's one (but this difference is rather small), but in our study the results are the same.

A study of statistical techniques and performance of GBML

- Genetic-Based Machine Learning
- Inferential Statistics
- Analyzing parametric tests
- Comparing two algorithms: From parametric to non-parametric tests
- Múltiple comparisons: Non-parametric tests
- **Lessons learned**

Lessons learned

On the use of non-parametric tests: The need of using non-parametric tests given that the necessary conditions for using parametric tests are not verified in the use of GBMLs algorithms.
If we have a set of data sets, we must apply a parametric test for each data set. We only need to use a non-parametric test for comparing the algorithms on the whole set of data sets. Besides, the non-parametric test allows us to compare deterministic algorithms (like C4.5) and GBMLs.

- A multiple comparison must be carried out first by using a statistical method for testing the differences among the related samples means. (Iman-Davenport's test). If differences have been found, then use a post-hoc statistical procedures.

- Holm's procedure is a very good test. Hochberg's method can rejects more hypothesis than Holm's one (but it does not control the FWER error). It can be used with independent algorithms.

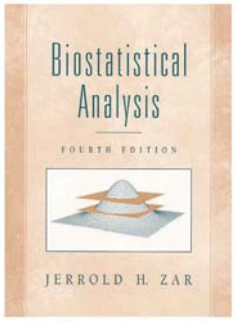
Lessons learned

- Wilcoxon's test computes a ranking based on differences between functions independently, whereas Friedman and derivative procedures compute the ranking between algorithms.

- Wilcoxon's test is highly influenced by the number of case of study (functions, data sets …). The N value determines the critical values to search in the statistical table. It is highly influenced by outliers when N is below or equal to 11.

- An appropriate number of algorithms in contracts with an appropriate number of case of study (functions, data sets ….) are need to be used in order to employ each type of test. As a rule of thumb, for k algorithms to compare, take at least $2 \cdot N$ cases of study.

Bibliography



J.H. Zar, Biostatistical Analyhsis, Prentice Hall, 1999.

D. Sheskin. Handbook of parametric and nonparametric statistical procedures. Chapman & Hall/CRC, 2003.

Demsar, J., Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research. Vol. 7. pp. 1–30. 2006.

S. García, A. Fernandez, A.D. Benítez, F. Herrera, Statistical Comparisons by Means of Non-Parametric Tests: A Case Study on Genetic Based Machine Learning. *Proceedings of the II Congreso Español de Informática (CEDI 2007). V Taller Nacional de Minería de Datos y Aprendizaje (TAMIDA 2007), Zaragoza (Spain), 95-104, 11-14 September 2007.*

S. García, D. Molina, M. Lozano, F. Herrera, A Study on the Use of Non-Parametric Tests for Analyzing the Evolutionary Algorithms' Behaviour: A Case Study on the CEC'2005 Special Session on Real Parameter Optimization. Journal of Heuristics, doi: 10.1007/s10732-008-9080-4, in press (2008)

http://keel.es

Thank you for your attention!

Questions?