



Dottorato di Ricerca in Ingegneria dell'Informazione

Data Mining and Soft Computing

Francisco Herrera

**Research Group on Soft Computing and
Information Intelligent Systems (SCI²S)**

Dept. of Computer Science and A.I.

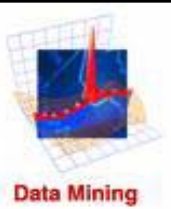
University of Granada, Spain

Email: herrera@decsai.ugr.es

<http://sci2s.ugr.es>

<http://decsai.ugr.es/~herrera>





Data Mining and Soft Computing

Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Data Mining - From the Top 10 Algorithms to the New Challenges
5. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
6. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
7. Genetic Fuzzy Systems: State of the Art and New Trends
8. Some Advanced Topics I: Classification with Imbalanced Data Sets
9. Some Advanced Topics II: Subgroup Discovery
10. Some advanced Topics III: Data Complexity
11. Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.

Design of Experiments in Data Mining/Computational Intelligence

In this talk

We focus on the use of statistical test for analyzing the results obtained in a design of experiments within the fields of Data Mining and Computational Intelligence.

Design of Experiments in Data Mining/Computational Intelligence

Motivation

The experimental analysis on the performance of a new method is a crucial and necessary task to carry out in a research on Data Mining, Computational Intelligence techniques.

Deciding when an algorithm is better than other one may not be a trivial task.

Design of Experiments in Data Mining/Computational Intelligence

Motivation

Deciding when an algorithm is better than other one may not be a trivial task.

Example for classification

Large Variations in Accuracies of Different Classifiers

	Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5	Alg. 6	Alg. 7
aud	25.3	76.0	68.4	69.6	79.0	81.2	57.7
aus	55.5	81.9	85.4	77.5	85.2	83.3	85.7
bal	45.0	76.2	87.2	90.4	78.5	81.9	79.8
bpa	58.0	63.5	60.6	54.3	65.8	65.8	68.2
bps	51.6	83.2	82.8	78.6	80.1	79.0	83.3
bre	65.5	96.0	96.7	96.0	95.4	95.3	96.0
cmc	42.7	44.4	46.8	50.6	52.1	49.8	52.3
gls	34.6	66.3	66.4	47.6	65.8	69.0	72.6
h-c	54.5	77.4	83.2	83.6	73.6	77.9	79.9
hep	79.3	79.9	80.8	83.2	78.9	80.0	83.2
irs	33.3	95.3	95.3	94.7	95.3	95.3	94.7
krk	52.2	89.4	94.9	87.0	98.3	98.4	98.6
lab	65.4	81.1	92.1	95.2	73.3	73.9	75.4
led	10.5	62.4	75.0	74.9	74.9	75.1	74.8
lym	55.0	83.3	83.6	85.6	77.0	71.5	79.0
mmg	56.0	63.0	65.3	64.7	64.8	61.9	63.4
mus	51.8	100.0	100.0	96.4	100.0	100.0	99.8
mux	49.9	78.6	99.8	61.9	99.9	100.0	100.0
pmi	65.1	70.3	73.9	75.4	73.1	72.6	76.0
prt	24.9	34.5	42.5	50.8	41.6	39.8	43.7
seg	14.3	97.4	96.1	80.1	97.2	96.8	96.1
sick	93.8	96.1	96.3	93.3	98.4	97.0	96.7
soyb	13.5	89.5	90.3	92.8	91.4	90.3	76.2
tao	49.8	96.1	96.0	80.8	95.1	93.6	88.4
thy	19.5	68.1	65.1	80.6	92.1	92.1	86.3
veh	25.1	69.4	69.7	46.2	73.6	72.6	72.2
vote	61.4	92.4	92.6	90.1	96.3	96.5	95.4
vow	9.1	99.1	96.6	65.3	80.7	78.3	87.6
wne	39.8	95.6	96.8	97.8	94.6	92.9	96.3
zoo	41.7	94.6	92.5	95.4	91.6	92.5	92.6
Avg	44.8	80.0	82.4	78.0	82.1	81.8	81.7

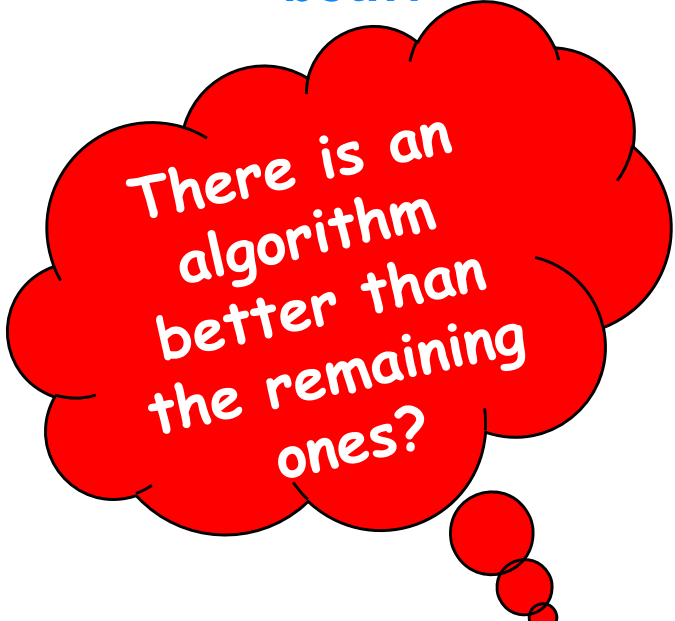
Design of Experiments in Data Mining/Computational Intelligence

Motivation

Alg. 4 is the winner in 8 problems with average 78.0

Alg. 2 is the winner for 4 problems with average 80.0

What is the best between both?



There is an algorithm better than the remaining ones?

	Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5	Alg. 6	Alg. 7
aud	25.3	76.0	68.4	69.6	79.0	81.2	57.7
aus	55.5	81.9	85.4	77.5	85.2	83.3	85.7
bal	45.0	76.2	87.2	90.4	78.5	81.9	79.8
bpa	58.0	63.5	60.6	54.3	65.8	65.8	68.2
bps	51.6	83.2	82.8	78.6	80.1	79.0	83.3
bre	65.5	96.0	96.7	96.0	95.4	95.3	96.0
cmc	42.7	44.4	46.8	50.6	52.1	49.8	52.3
gls	34.6	66.3	66.4	47.6	65.8	69.0	72.6
h-c	54.5	77.4	83.2	83.6	73.6	77.9	79.9
hep	79.3	79.9	80.8	83.2	78.9	80.0	83.2
irs	33.3	95.3	95.3	94.7	95.3	95.3	94.7
krk	52.2	89.4	94.9	87.0	98.3	98.4	98.6
lab	65.4	81.1	92.1	95.2	73.3	73.9	75.4
led	10.5	62.4	75.0	74.9	74.9	75.1	74.8
lym	55.0	83.3	83.6	85.6	77.0	71.5	79.0
mmg	56.0	63.0	65.3	64.7	64.8	61.9	63.4
mus	51.8	100.0	100.0	96.4	100.0	100.0	99.8
mux	49.9	78.6	99.8	61.9	99.9	100.0	100.0
pmi	65.1	70.3	73.9	75.4	73.1	72.6	76.0
prt	24.9	34.5	42.5	50.8	41.6	39.8	43.7
seg	14.3	97.4	96.1	80.1	97.2	96.8	96.1
sick	93.8	96.1	96.3	93.3	98.4	97.0	96.7
soyb	13.5	89.5	90.3	92.8	91.4	90.3	76.2
tao	49.8	96.1	96.0	80.8	95.1	93.6	88.4
thy	19.5	68.1	65.1	80.6	92.1	92.1	86.3
veh	25.1	69.4	69.7	46.2	73.6	72.6	72.2
vote	61.4	92.4	92.6	90.1	96.3	96.5	95.4
vow	9.1	99.1	96.6	65.3	80.7	78.3	87.6
wne	39.8	95.6	96.8	97.8	94.6	92.9	96.3
zoo	41.7	94.6	92.5	95.4	91.6	92.5	92.6
Avg	44.8	80.0	82.4	78.0	82.1	81.8	81.7

Design of Experiments in Data Mining/Computational Intelligence

Motivation

We must use
statistical tests for
comparing the
algorithms.

The problem:

How must I do the
statistical
experimental
study?

What tests must I
use?

	Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5	Alg. 6	Alg. 7
aud	25.3	76.0	68.4	69.6	79.0	81.2	57.7
aus	55.5	81.9	85.4	77.5	85.2	83.3	85.7
bal	45.0	76.2	87.2	90.4	78.5	81.9	79.8
bpa	58.0	63.5	60.6	54.3	65.8	65.8	68.2
bps	51.6	83.2	82.8	78.6	80.1	79.0	83.3
bre	65.5	96.0	96.7	96.0	95.4	95.3	96.0
cmc	42.7	44.4	46.8	50.6	52.1	49.8	52.3
gls	34.6	66.3	66.4	47.6	65.8	69.0	72.6
h-c	54.5	77.4	83.2	83.6	73.6	77.9	79.9
hep	79.3	79.9	80.8	83.2	78.9	80.0	83.2
irs	33.3	95.3	95.3	94.7	95.3	95.3	94.7
krk	52.2	89.4	94.9	87.0	98.3	98.4	98.6
lab	65.4	81.1	92.1	95.2	73.3	73.9	75.4
led	10.5	62.4	75.0	74.9	74.9	75.1	74.8
lym	55.0	83.3	83.6	85.6	77.0	71.5	79.0
mmg	56.0	63.0	65.3	64.7	64.8	61.9	63.4
mus	51.8	100.0	100.0	96.4	100.0	100.0	99.8
mux	49.9	78.6	99.8	61.9	99.9	100.0	100.0
pmi	65.1	70.3	73.9	75.4	73.1	72.6	76.0
prt	24.9	34.5	42.5	50.8	41.6	39.8	43.7
seg	14.3	97.4	96.1	80.1	97.2	96.8	96.1
sick	93.8	96.1	96.3	93.3	98.4	97.0	96.7
soyb	13.5	89.5	90.3	92.8	91.4	90.3	76.2
tao	49.8	96.1	96.0	80.8	95.1	93.6	88.4
thy	19.5	68.1	65.1	80.6	92.1	92.1	86.3
veh	25.1	69.4	69.7	46.2	73.6	72.6	72.2
vote	61.4	92.4	92.6	90.1	96.3	96.5	95.4
vow	9.1	99.1	96.6	65.3	80.7	78.3	87.6
wne	39.8	95.6	96.8	97.8	94.6	92.9	96.3
zoo	41.7	94.6	92.5	95.4	91.6	92.5	92.6
Avg	44.8	80.0	82.4	78.0	82.1	81.8	81.7

Design of Experiments in Data Mining/Computational Intelligence

Objective

To show some results on the use of statistical tests for comparing algorithms in the field of DM/CI.

We will not discuss the performance measures that can be used neither the choice on the set of benchmarks.

For classification, for example, the following references is a study for performance measures.

Ferri, C., Hernández-Orallo, J., Modroiu, R.
An experimental comparison of performance measures for classification.
Pattern Recognition Letters, 2008, in press.
Doi: 10.1016/j.patrec.2008.08.010

Design of Experiments in Data Mining/Computational Intelligence

Outline

- **Introduction**
- **Conditions for the safe use of parametric tests**
- **Using non-parametric tests: Data Mining/
Computational Intelligence based case studies**
 - **Two sample tests/Multiple comparisons**
 - **Evolutionary Algorithms: CEC'05 Special Session on
parameter optimization**
 - **Neural network and genetic learning experiments**
- **Lessons learned**

Design of Experiments in Data Mining/Computational Intelligence

Outline

- **Introduction** (Inferential statistics, basic concepts)
- Conditions for the safe use of parametric tests
- Using non-parametric tests: Data Mining/
Computational Intelligence based case studies
 - Two sample tests/Multiple comparisons
 - Evolutionary Algorithms: CEC'05 Special Session on parameter optimization
 - Neural network and genetic learning experiments
- Lessons learned

Introduction

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

For example: Comparing two or various sets of experiments in a computational problem.

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

Inferential Statistics

(basic concepts)

Null-Hypothesis

H_0 : The 2 samples come from populations with the same distributions.

Or, median of population 1 = median of population 2
(generalization with n samples)

Significance level α

- Significance level for all tests tell us whether or not to reject the null hypothesis (and with what confidence).
- A significance level of 90% or 95% is often sufficient, some use 99%

Inferential Statistics

(basic concepts)

Significance level α

- If you decide for a significance level of 0.05 (95% certainty that there indeed is a significant difference), then a **p-value** (provided by the test) smaller than 0.05 indicates that you can reject the **null-hypothesis**
- **Remember:** the null-hypothesis generally predicts that the means are equal.
- So, in a test, if you have $p = 0.07$ means that you **cannot reject** the null hypothesis that "there is equal means" \Rightarrow **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
- If more than two variables

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

Inferential Statistics

Parametric Assumptions

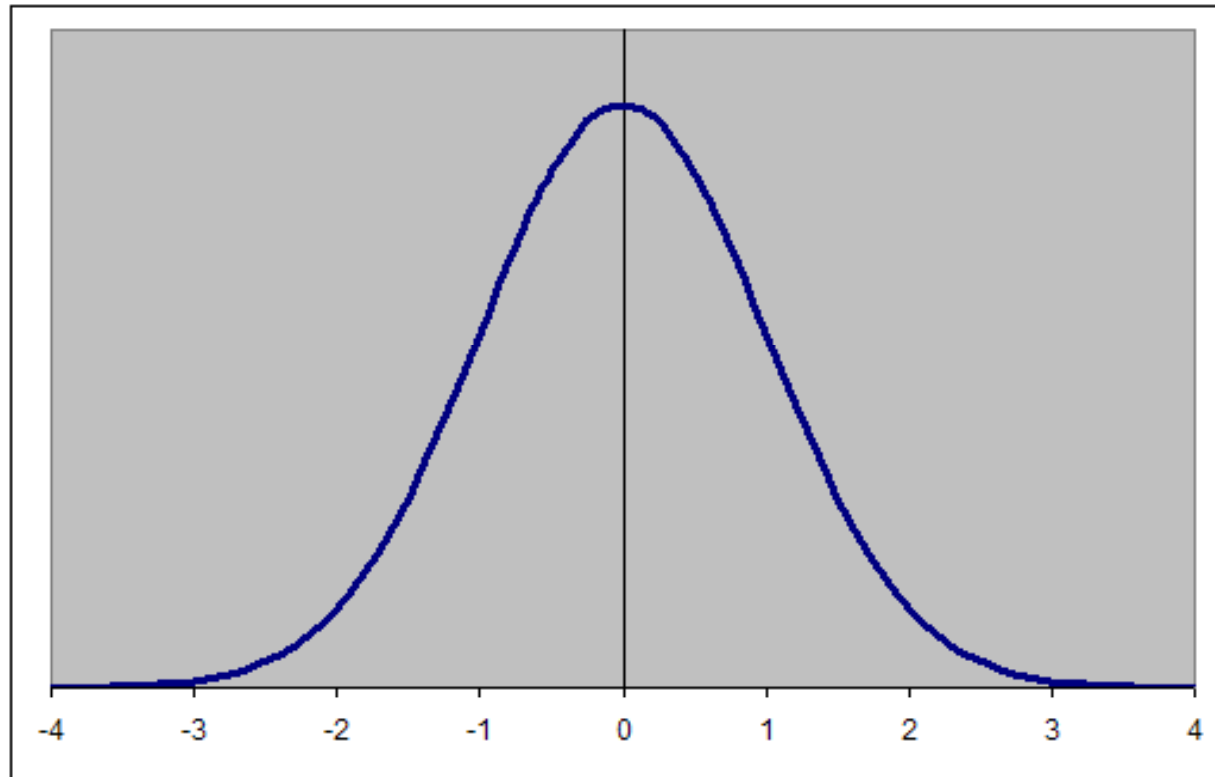
(t-test, ANOVA, ...)

- 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

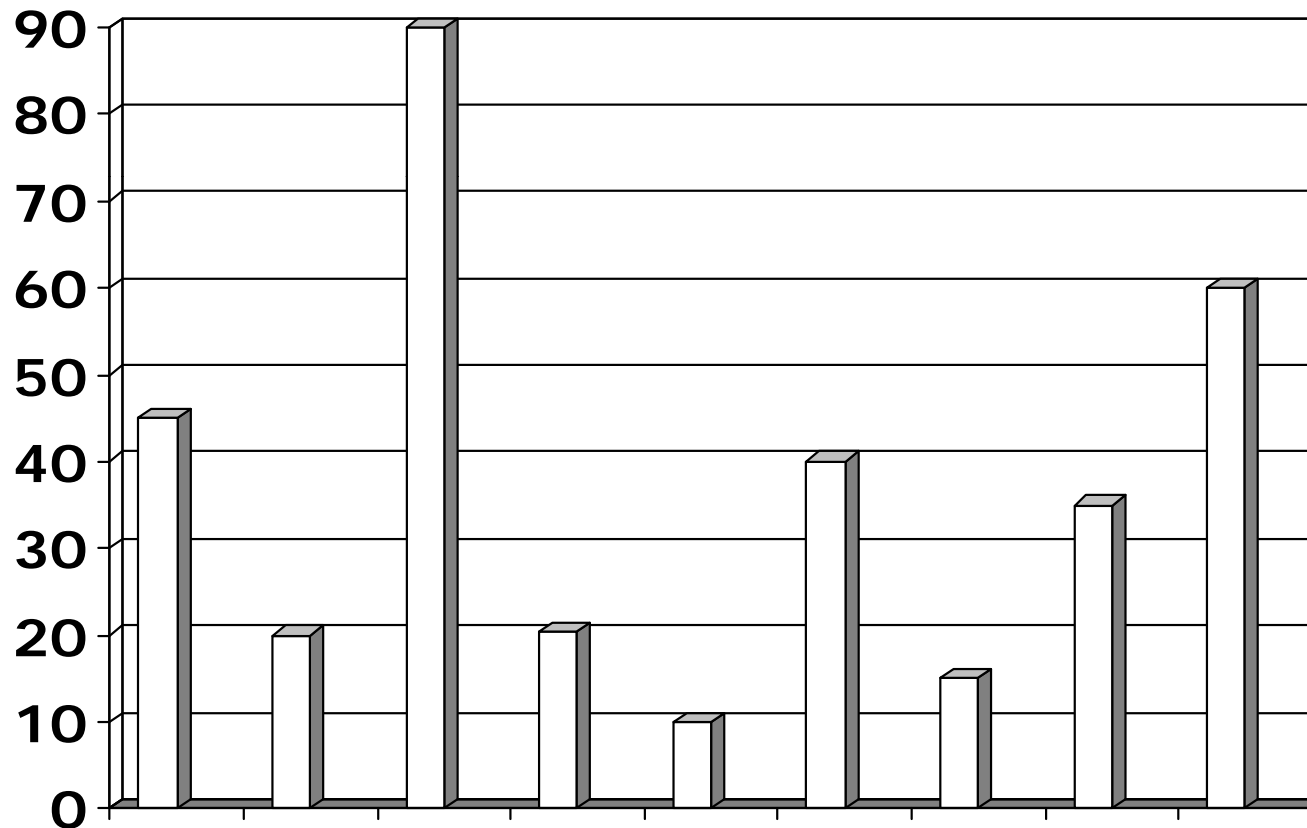
Normality

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



Inferential Statistics

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



Histogram

Inferential Statistics

Nonparametric Assumptions

- Observations are independent
- 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.

Design of Experiments in Data Mining/Computational Intelligence

Outline

- Introduction
- **Conditions for the safe use of parametric tests**
- Using non-parametric tests: Data Mining/
Computational Intelligence based case studies
 - Two sample tests/Multiple comparisons
 - Evolutionary Algorithms: CEC'05 Special Session on parameter optimization
 - Neural network and genetic learning experiments
- Lessons learned

Conditions for the safe use of 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.

Conditions for the safe use of 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**

Conditions for the safe use of parametric tests

CASE OF STUDY:

Neural networks models: MLP, RBFN (10-cfv, 5 runs per partition)

TABLE I
DATA SETS USED FOR EXPERIMENTATION

Data set	# Instances	# Attributes	# Classes
breast	682	10	2
cleveland	303	13	5
crx	689	16	2
glass	214	9	7
iris	150	4	3
pima	768	8	2
wisconsin	699	10	2

Conditions for the safe use of parametric tests

TABLE II
RESULTS FOR ANNs USED

Using 10-fold cross validation								
Method	MLP Backprop.-1x25		MLP Backprop.-1x5		RBFN Decremental		RBFN	
Dataset	Mean	St. Desv.	Mean	St. Desv.	Mean	St. Desv.	Mean	St. Desv.
Breast	0.96	0.01	0.96	0.01	0.83	0.06	0.86	0.04
Cleveland	0.51	0.07	0.49	0.10	0.35	0.09	0.35	0.10
Crx	0.85	0.05	0.82	0.09	0.45	0.02	0.45	0.02
Glass	0.50	0.10	0.46	0.14	0.29	0.12	0.37	0.13
Iris	0.74	0.10	0.75	0.13	0.90	0.09	0.86	0.09
Pima	0.74	0.05	0.70	0.09	0.68	0.05	0.62	0.12
Wisconsin	0.97	0.02	0.96	0.05	0.84	0.09	0.86	0.07

Conditions for the safe use of parametric tests

TABLE I. Kolmogorov-Smirnov test

TABLE III
RESULTS FOR KOLMOGOROV-SMIRNOV TEST

	10-fold cross validation						
	Breast	cleveland	crx	glass	iris	pima	wisconsin
MLP backpropagation-1x25	* (.00)	* (.02)	* (.04)	(.20)	* (.00)	* (.00)	* (.00)
MLP backpropagation-1x5	* (.00)	(.20)	* (.00)	(.20)	* (.00)	* (.00)	* (.00)
RBFN Decremental	* (.00)	(.05)	* (.00)	(.08)	* (.00)	(.20)	* (.00)
RBFN	* (.00)	* (.04)	* (.00)	(.20)	* (.00)	* (.00)	* (.00)

a **p-value** smaller than 0.05 indicates that you can reject the **null-hypothesis**

Conditions for the safe use of parametric tests

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

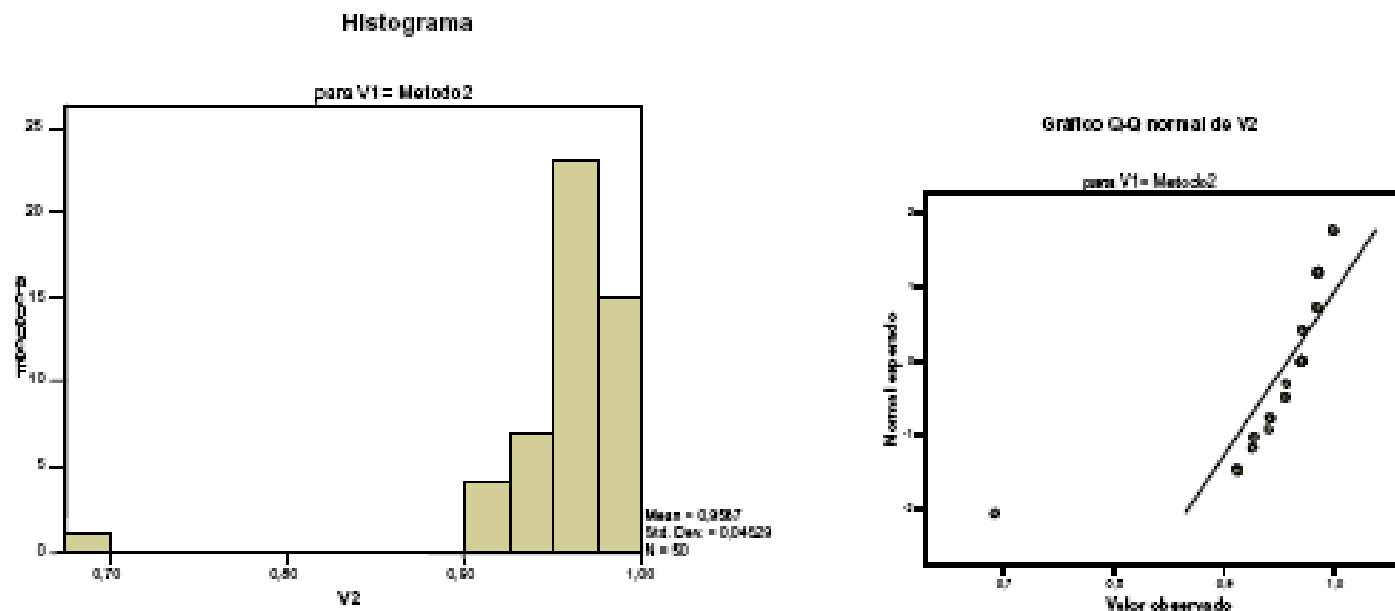


Fig. 1. MLP 1x5 in Breast with 10fcv

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

Conditions for the safe use of 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 samples present or not this homogeneity of variances (homoscedasticity).

Conditions for the safe use of parametric tests

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

TABLE V
RESULTS FOR LEVENE'S TEST

	Breast	cleveland	crx	glass	iris	pima	Wisconsin
Levene 10-fcv	* (.00)	(.15)	* (.00)	(.10)	(.16)	* (.00)	* (.00)
Levene Hold-out	* (.00)	* (.00)	(.13)	* (.01)	(.26)	* (.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 homogeneous (we reject the null hypothesis).

Design of Experiments in Data Mining/Computational Intelligence

Outline

- Introduction
- Conditions for the safe use of parametric tests
- **Using non-parametric tests: Data Mining/
Computational Intelligence based case studies**
 - Two sample tests
 - Multiple comparisons
 - Evolutionary Algorithms: CEC'05 Special Session on parameter optimization
 - Neural network and genetic learning experiments
- Lessons learned

Design of Experiments in Data Mining/Computational Intelligence

Using non-parametric tests: Data Mining/
Computational Intelligence based case studies

- Two sample tests
- Multiple comparisons
- Evolutionary Algorithms: CEC'05 Special Session on parameter optimization
- Neural network and genetic learning experiments

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 (<i>Also, Sign test</i>)

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_0 : The 2 samples come from populations with the same 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_0 if test statistic \leq critical value, or if $P \leq \alpha$ (alpha).

6. Report Test results.

Two-Sample tests

Wilcoxon Signed-Ranks Test for Paired Samples

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

	C4.5	C4.5+m	difference	rank	
adult (sample)	0.763	0.768	+0.005	3.5	
breast cancer	0.599	0.591	-0.008	7	
breast cancer wisconsin	0.954	0.971	+0.017	9	
cmc	0.628	0.661	+0.033	12	$R^+ = 3.5 + 9 + 12 + 5 + 6 + 14 +$
ionosphere	0.882	0.888	+0.006	5	$11 + 13 + 8 + 10 + 1.5 = 93$
iris	0.936	0.931	-0.005	3.5	
liver disorders	0.661	0.668	+0.007	6	
lung cancer	0.583	0.583	0.000	1.5	$R^- = 7 + 3.5 + 1.5 = 12$
lymphography	0.775	0.838	+0.063	14	
mushroom	1.000	1.000	0.000	1.5	
primary tumor	0.940	0.962	+0.022	11	
rheum	0.619	0.666	+0.047	13	
voting	0.972	0.981	+0.009	8	
wine	0.957	0.978	+0.021	10	

Table 2: Comparison of AUC for C4.5 with $m = 0$ and C4.5 with m tuned for the optimal AUC. The columns on the right-hand illustrate the computation and would normally not be published in an actual paper.

Two-Sample tests

Wilcoxon Signed-Ranks Test for Paired Samples

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

$$R^+ = 3.5 + 9 + 12 + 5 + 6 + 14 + 11 + 13 + 8 + 10 + 1.5 = 93$$

$$R^- = 7 + 3.5 + 1.5 = 12$$

$$T = \text{Min} \{R^+, R^-\} = 12$$

$$\alpha = 0.05, N = 14 \quad \text{dif} = 21$$

We reject the null-hypothesis

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

Two-Sample tests

Wilcoxon Signed-Ranks Test for Paired Samples

Critical value for T for N up to 25.

It $T \leq \text{dif (table-value)}$
then Reject the H_0

<i>n</i>	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

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 in SPSS

Analyze → Nonparametric Tests → 2 Related Samples Tests

- Select pair(s) of variables
- Select Wilcoxon

Two-Sample tests

Wilcoxon Signed-Ranks Test in SPSS

OUTPUT

		N	Mean Rank	Sum of Ranks
beta-endorphin conc. after (pmol/l) - beta-endorphin conc. before (pmol/l)	Negative Ranks	0 ^a	.00	.00
	Positive Ranks	11 ^b	6.00	66.00
	Ties	0 ^c		
	Total	11		

- a. beta-endorphin conc. after (pmol/l) < beta-endorphin conc. before (pmol/l)
 b. beta-endorphin conc. after (pmol/l) > beta-endorphin conc. before (pmol/l)
 c. beta-endorphin conc. before (pmol/l) = beta-endorphin conc. after (pmol/l)

Test Statistics^b

	beta-endorphin conc. after (pmol/l) - beta-endorphin conc. before (pmol/l)
Z	-2.934 ^a
Asymp. Sig. (2-tailed)	.003

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

Conclude: Reject H_0 (Wilcoxon Signed-Ranks test, $Z = -2.934$, $p = 0.003$, $n = 11$, 0).

Design of Experiments in Data Mining/Computational Intelligence

Using non-parametric tests: Data Mining/
Computational Intelligence based case studies

- Two sample tests
- Multiple comparisons
- Evolutionary Algorithms: CEC'05 Special Session on parameter optimization
- Neural network and genetic learning experiments

Cases of study: CEC'2005 Special Session on real parameter optimization

Special Session on Real-Parameter Optimization at CEC-05, Edinburgh, UK, 2-5 Sept. 2005

25 functions with real parameters, 10 variables:

f1-f5 unimodal functions f6-f25 multimodal functions

P. N. Suganthan, N. Hansen, J. J. Liang, K. Deb, Y. P. Chen, A. Auger, and S. Tiwari, "Problem definitions and evaluation criteria for the CEC 2005 special session on real parameter optimization." Nanyang Technological University, Tech. Rep., 2005, available as http://www.ntu.edu.sg/home/epnsugan/index_files/CEC-05/Tech-Report-May-30-05.pdf.

N. Hansen, "Compilation of Results on the CEC Benchmark Function Set," Institute of Computational Science, ETH Zurich, Switzerland, Tech. Rep., 2005, available as http://www.ntu.edu.sg/home/epnsugan/index_files/CEC-05/compareresults.pdf.

Source: [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](https://doi.org/10.1007/s10732-008-9080-4), *in press (2008)*.

Cases of study: CEC'2005 Special Session on real parameter optimization

- Algorithms involved in the comparison: **(11 algorithms)**
 - **BLX-GL50** (Garcia-Martinez & Lozano, 2005): Hybrid Real-Coded Genetic Algorithms with Female and Male Differentiation
 - **BLX-MA** (Molina *et al.*, 2005): Adaptive Local Search Parameters for Real-Coded Memetic Algorithms
 - **CoEVO** (Posik, 2005): Mutation Step Co-evolution
 - **DE** (Ronkkonen *et al.*,2005):Differential Evolution
 - **DMS-L-PSO**: Dynamic Multi-Swarm Particle Swarm Optimizer with Local Search
 - **EDA** (Yuan & Gallagher, 2005): Estimation of Distribution Algorithm
 - **G-CMA-ES** (Auger & Hansen, 2005): A restart Covariance Matrix Adaptation Evolution Strategy with increasing population size
 - **K-PCX** (Sinha *et al.*, 2005): A Population-based, Steady-State real-parameter optimization algorithm with parent-centric recombination operator, a polynomial mutation operator and a niched -selection operation.
 - **L-CMA-ES** (Auger & Hansen, 2005): A restart local search Covariance Matrix Adaptation Evolution Strategy
 - **L-SaDE** (Qin & Suganthan, 2005): Self-adaptive Differential Evolution algorithm with Local Search
 - **SPC-PNX** (Ballester *et al.*,2005): A steady-state real-parameter GA with PNX crossover operator

Cases of study: CEC'2005 Special Session on real parameter optimization

Table 1: Test of Normality of Kolmogorov-Smirnov

	f1	f2	f3	f4	f5	f6	f7	f8	f9
BLX-GL50	(.20)	* (.04)	* (.00)	(.14)	* (.00)	* (.00)	* (.04)	(.20)	* (.00)
BLX-MA	* (.01)	* (.00)	* (.01)	* (.00)	* (.00)	(.16)	(.20)	* (.00)	* (.00)
	f10	f11	f12	f13	f14	f15	f16	f17	f18
BLX-GL50	(.10)	(.20)	* (.00)	(.20)	(.20)	* (.00)	* (.00)	(.20)	* (.00)
BLX-MA	(.20)	* (.00)	* (.00)	(.20)	* (.02)	* (.00)	(.20)	(.20)	* (.00)
	f19	f20	f21	f22	f23	f24	f25		
BLX-GL50	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)		
BLX-MA	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.02)		

Table 2: Test of Normality of Shapiro-Wilk

	f1	f2	f3	f4	f5	f6	f7	f8	f9
BLX-GL50	* (.03)	(.06)	* (.00)	* (.03)	* (.00)	* (.00)	* (.01)	(.23)	* (.00)
BLX-MA	* (.00)	* (.00)	* (.01)	* (.00)	* (.00)	(.05)	(.27)	* (.03)	* (.00)
	f10	f11	f12	f13	f14	f15	f16	f17	f18
BLX-GL50	(.07)	(.25)	* (.00)	(.39)	(.41)	* (.00)	* (.00)	(.12)	* (.00)
BLX-MA	(.31)	* (.00)	* (.00)	(.56)	* (.01)	* (.00)	(.25)	(.72)	* (.00)
	f19	f20	f21	f22	f23	f24	f25		
BLX-GL50	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)		
BLX-MA	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.02)		

Cases of study: CEC'2005 Special Session on real parameter optimization

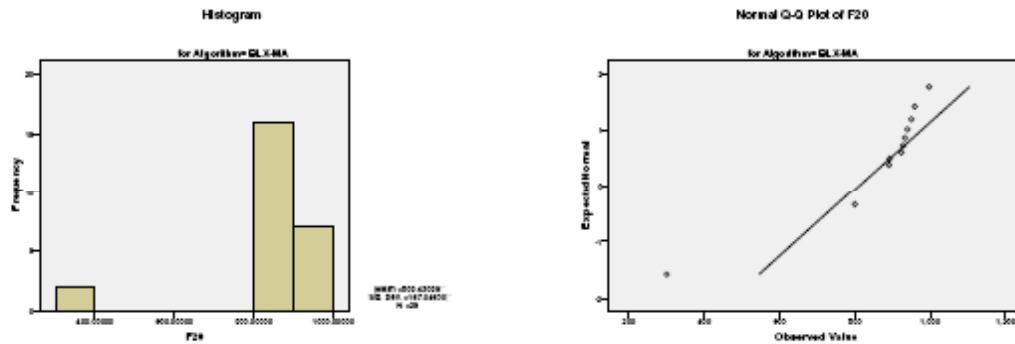


Figure 1: Example of non-normal distribution: Function f20 and BLX-GL50 algorithm: Histogram and Q-Q Graphic.

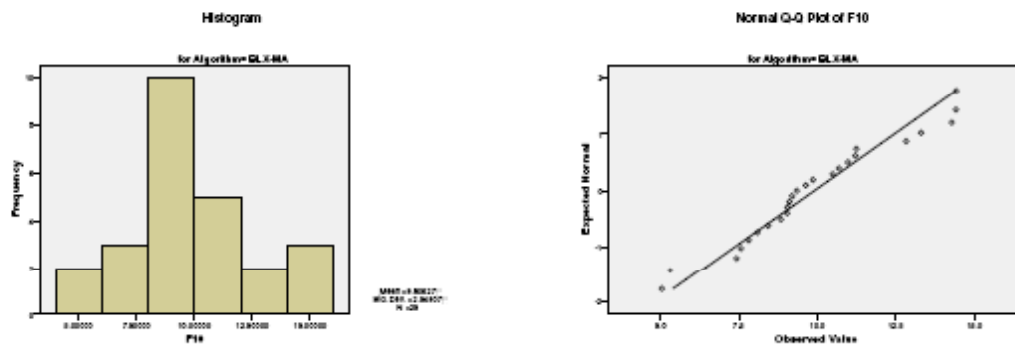


Figure 2: Example of normal distribution: Function f10 and BLX-MA algorithm: Histogram and Q-Q Graphic.

Cases of study: CEC'2005 Special Session on real parameter optimization

Table 4 Test of heteroscedasticity of Levene (based on means)

	f1	f2	f3	f4	f5	f6	f7	f8	f9
LEVENE	(.07)	(.07)	* (.00)	* (.04)	* (.00)	* (.00)	* (.00)	(.41)	* (.00)
	f10	f11	f12	f13	f14	f15	f16	f17	f18
LEVENE	(.99)	* (.00)	(.98)	(.18)	(.87)	* (.00)	* (.00)	(.24)	(.21)
	f19	f20	f21	f22	f23	f24	f25		
LEVENE	* (.01)	* (.00)	* (.01)	(.47)	(.28)	* (.00)	* (.00)		

Cases of study: CEC'2005 Special Session on real parameter optimization

TABLE XVI
WILCOXON TEST FOR ALL FUNCTIONS (F1-F25)

alg.	R^+	R^-	Hyp. α	Hyp. α	Hyp. α	Hyp. α
			0.01	0.02	0.05	0.1
BLX-GL50	289.5	35.5	R	R	R	R
BLX-MA	295.5	29.5	R	R	R	R
COEVO	301.0	24.0	R	R	R	R
DE	262.5	62.5	R	R	R	R
DMS-L-PSO	199.0	126.0	A	A	A	A
EDA	284.5	40.5	R	R	R	R
K-PCX	269.0	56.0	R	R	R	R
L-CMA-ES	273.0	52.0	R	R	R	R
L-SADE	209.0	116.0	A	A	A	A
SPC-PNX	305.5	19.5	R	R	R	R

G-CMAES versus the remaining algorithms.

The critical values are: 68, 76, 89 and 100 (0.01, 0.02, 0.05, 0.1)

Cases of study: CEC'2005 Special Session on real parameter optimization

Using Wilcoxon test for 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) increase. The true statistical significance for the pairwise comparison test is given by:

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

Cases of study: CEC'2005 Special Session on real parameter optimization

Cases of study I: CEC'2005 Special Session on real parameter optimization

Example on the use of Wilcoxon's test combined for multiple comparisons

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

$$\begin{aligned} p &= 1 - ((1 - 0.009) \cdot (1 - 0.016) \cdot (1 - 0.016) \cdot (1 - 0.028) \cdot (1 - 0.013) \cdot \\ &\quad \cdot (1 - 0.016) \cdot (1 - 0.026) \cdot (1 - 0.007)) = 0.123906 \end{aligned}$$

Design of Experiments in Data Mining/Computational Intelligence

Using non-parametric tests: Data Mining/
Computational Intelligence based case studies

- Two sample tests
- Multiple comparisons
- Evolutionary Algorithms: CEC'05 Special Session on parameter optimization
- Neural network and genetic learning experiments

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, assigning to the best of them the ranking 1, and to the worst the ranking k .

Under the null hypothesis, formed from supposing that the results of the algorithms are equivalent and, therefore, their rankings are also similar, the Friedman statistic

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

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 according 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).

Cases of study: Machine learning example, C4.5

	C4.5	C4.5+m	C4.5+cf	C4.5+m+cf
adult (sample)	0.763 (4)	0.768 (3)	0.771 (2)	0.798 (1)
breast cancer	0.599 (1)	0.591 (2)	0.590 (3)	0.569 (4)
breast cancer wisconsin	0.954 (4)	0.971 (1)	0.968 (2)	0.967 (3)
cmc	0.628 (4)	0.661 (1)	0.654 (3)	0.657 (2)
ionosphere	0.882 (4)	0.888 (2)	0.886 (3)	0.898 (1)
iris	0.936 (1)	0.931 (2.5)	0.916 (4)	0.931 (2.5)
liver disorders	0.661 (3)	0.668 (2)	0.609 (4)	0.685 (1)
lung cancer	0.583 (2.5)	0.583 (2.5)	0.563 (4)	0.625 (1)
lymphography	0.775 (4)	0.838 (3)	0.866 (2)	0.875 (1)
mushroom	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)
primary tumor	0.940 (4)	0.962 (2.5)	0.965 (1)	0.962 (2.5)
rheum	0.619 (3)	0.666 (2)	0.614 (4)	0.669 (1)
voting	0.972 (4)	0.981 (1)	0.975 (2)	0.975 (3)
wine	0.957 (3)	0.978 (1)	0.946 (4)	0.970 (2)
average rank	3.143	2.000	2.893	1.964

Table 6: Comparison of AUC between C4.5 with $m = 0$ and C4.5 with parameters m and/or cf tuned for the optimal AUC. The ranks in the parentheses are used in computation of the Friedman test and would usually not be published in an actual paper.

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

Cases of study: Machine learning example, C4.5

Friedman's measure: 9.28

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

Iman-Davenport's test:

$F_F = 3.69$, $F(3,3 \times 13) = 2.85$,
Therefore the null hypothesis
is rejected.

	C4.5	C4.5+m	C4.5+cf	C4.5+m+cf
adult (sample)	0.763 (4)	0.768 (3)	0.771 (2)	0.798 (1)
breast cancer	0.599 (1)	0.591 (2)	0.590 (3)	0.569 (4)
breast cancer wisconsin	0.954 (4)	0.971 (1)	0.968 (2)	0.967 (3)
cmc	0.628 (4)	0.661 (1)	0.654 (3)	0.657 (2)
ionosphere	0.882 (4)	0.888 (2)	0.886 (3)	0.898 (1)
iris	0.936 (1)	0.931 (2.5)	0.916 (4)	0.931 (2.5)
liver disorders	0.661 (3)	0.668 (2)	0.609 (4)	0.685 (1)
lung cancer	0.583 (2.5)	0.583 (2.5)	0.563 (4)	0.625 (1)
lymphography	0.775 (4)	0.838 (3)	0.866 (2)	0.875 (1)
mushroom	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)
primary tumor	0.940 (4)	0.962 (2.5)	0.965 (1)	0.962 (2.5)
rheum	0.619 (3)	0.666 (2)	0.614 (4)	0.669 (1)
voting	0.972 (4)	0.981 (1)	0.975 (2)	0.975 (3)
wine	0.957 (3)	0.978 (1)	0.946 (4)	0.970 (2)
average rank	3.143	2.000	2.893	1.964

Table 6: Comparison of AUC between C4.5 with $m = 0$ and C4.5 with parameters m and/or cf tuned for the optimal AUC. The ranks in the parentheses are used in computation of the Friedman test and would usually not be published in an actual paper.

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

Cases of study: CEC'2005 Special Session on real parameter optimization

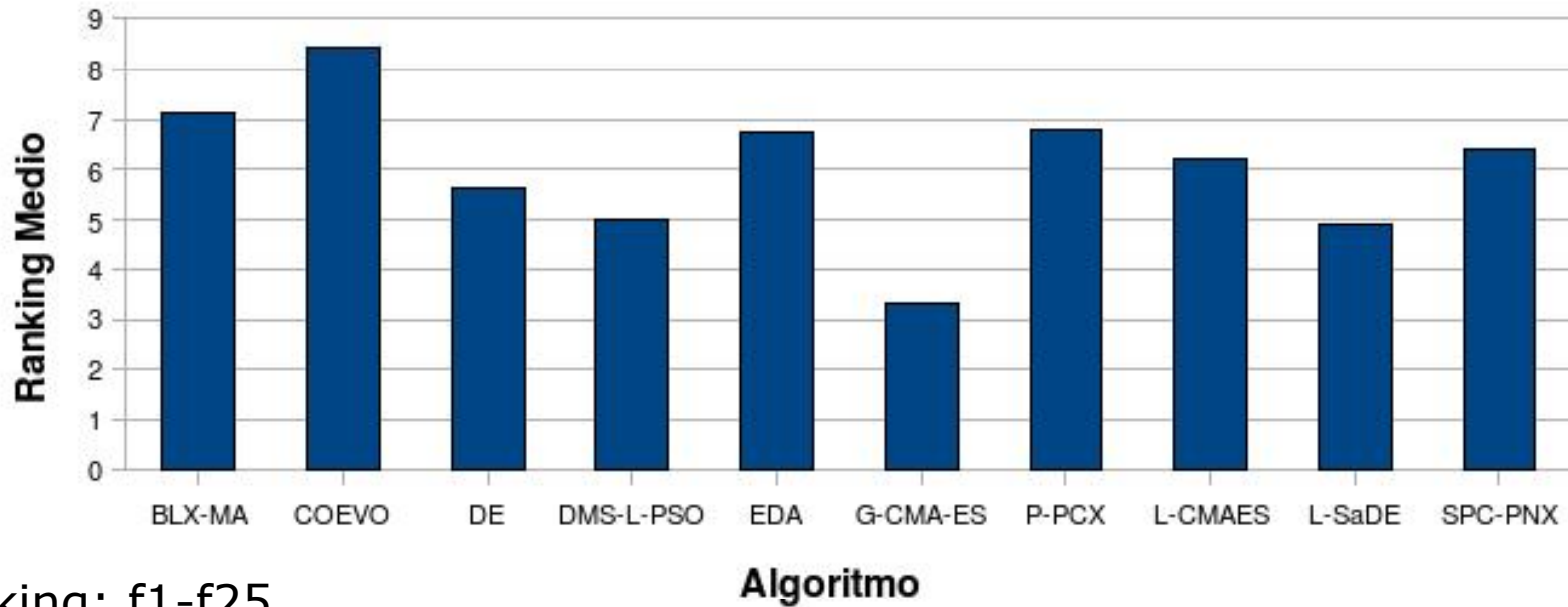
Table 7: Results of the Friedman and Iman-Davenport Tests ($\alpha = 0.05$)

	Friedman Value	Value in χ^2	p-value	Iman-Davenport Value	Value in F_F	p-value
f15-f25	26.942	18.307	0.0027	3.244	1.930	0.0011
All	41.985	18.307	< 0.0001	4.844	1.875	< 0.0001

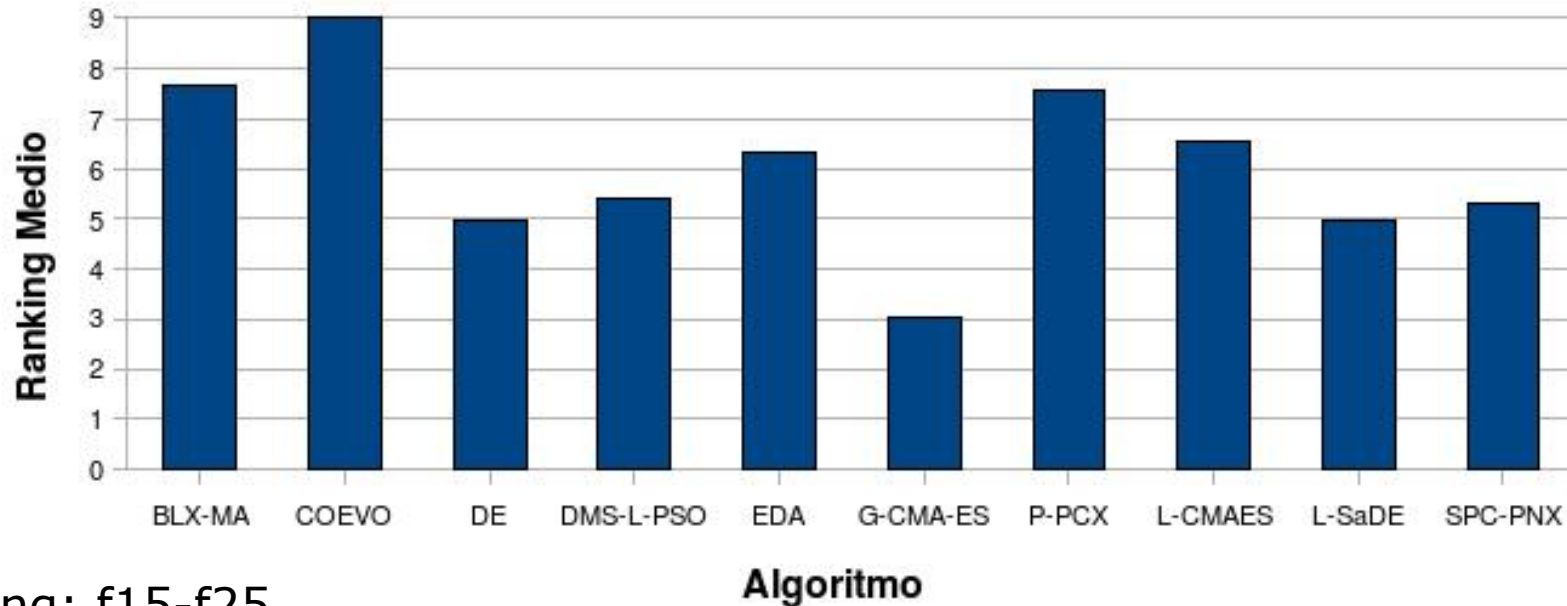
Table 8: Rankings obtained through Friedman's test and critical difference of Bonferroni-Dunn's procedure

Algorithm	Ranking (f15-f25)	Ranking (f1-f25)
BLX-GL50	5.227	5.3
BLX-MA	7.681	7.14
CoEVO	9.000	6.44
DE	4.955	5.66
DMS-L-PSO	5.409	5.02
EDA	6.318	6.74
G-CMA-ES	3.045	3.34
K-PCX	7.545	6.8
L-CMA-ES	6.545	6.22
L-SaDE	4.956	4.92
SPC-PNX	5.318	6.42

Cases of study: CEC'2005 Special Session on real parameter optimization



Ranking: f1-f25



Ranking: f15-f25

Multiple Comparisons

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

Holm's method compares each p_i with $\alpha/(k-i)$ starting from the most significant p value. If p_1 is below $\alpha/(k-1)$, 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 can not be rejected, all the remaining hypothesis are maintained as accepted. The statistic for comparing the i algorithm with the j algorithm is:

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

The value of z is used for finding the corresponding probability from the table of the normal distribution, which is compared with the corresponding value of α .

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

Cases of study: Machine learning example, C4.5

Holm's method: $SE = \sqrt{(4.5/6.14)} = 0.488$.

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

p-values are

0.607 (C4.5+cf)

0.019 (C4.5+m)

0.016 (C4.5+m+cf).

i	classifier	$z = (R_0 - R_i)/SE$	p	α/i
1	C4.5+m+cf	$(3.143 - 1.964)/0.488 = 2.416$	0.016	0.017
2	C4.5+m	$(3.143 - 2.000)/0.488 = 2.342$	0.019	0.025
3	C4.5+cf	$(3.143 - 2.893)/0.488 = 0.512$	0.607	0.050

The first one is rejected ($0.016 < 0.017$)

The second one is rejected ($0.019 < 0.025$),

The third one can not be rejected ($0.607 > 0.05$)

Cases of study: CEC'2005 Special Session on real parameter optimization

HOLM/HOCHBERG | TABLE FOR FUNCTIONS F1-F25 (G-CMA-ES IS THE CONTROL ALGORITHM)

i	algorithm	z	p	α/i 0.05	α/i 0.10
10	COEVO	5.43662	$5.43013 \cdot 10^{-8}$	0.00500	0.01000
9	BLX-MA	4.05081	$5.10399 \cdot 10^{-5}$	0.00556	0.01111
8	K-PCX	3.68837	$2.25693 \cdot 10^{-4}$	0.00625	0.01250
7	EDA	3.62441	$2.89619 \cdot 10^{-4}$	0.00714	0.01429
6	SPC-PNX	3.28329	0.00103	0.00833	0.01667
5	L-CMA-ES	3.07009	0.00214	0.01000	0.02000
4	DE	2.47313	0.01339	0.01250	0.02500
3	BLX-GL50	2.08947	0.03667	0.01667	0.03333
2	DMS-L-PSO	1.79089	0.07331	0.02500	0.05000
1	L-SADE	1.68429	0.09213	0.05000	0.10000

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 hypothesis 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.

Cases of study: CEC'2005 Special Session on real parameter optimization

HOLM/HOCHBERG | TABLE FOR FUNCTIONS F1-F25 (G-CMA-ES IS THE CONTROL ALGORITHM)

i	algorithm	z	p	α/i 0.05	α/i 0.10
10	COEVO	5.43662	$5.43013 \cdot 10^{-8}$	0.00500	0.01000
9	BLX-MA	4.05081	$5.10399 \cdot 10^{-5}$	0.00556	0.01111
8	K-PCX	3.68837	$2.25693 \cdot 10^{-4}$	0.00625	0.01250
7	EDA	3.62441	$2.89619 \cdot 10^{-4}$	0.00714	0.01429
6	SPC-PNX	3.28329	0.00103	0.00833	0.01667
5	L-CMA-ES	3.07009	0.00214	0.01000	0.02000
4	DE	2.47313	0.01339	0.01250	0.02500
3	BLX-GL50	2.08947	0.03667	0.01667	0.03333
2	DMS-L-PSO	1.79089	0.07331	0.02500	0.05000
1	L-SADE	1.68429	0.09213	0.05000	0.10000

Cases of study: CEC'2005 Special Session on real parameter optimization

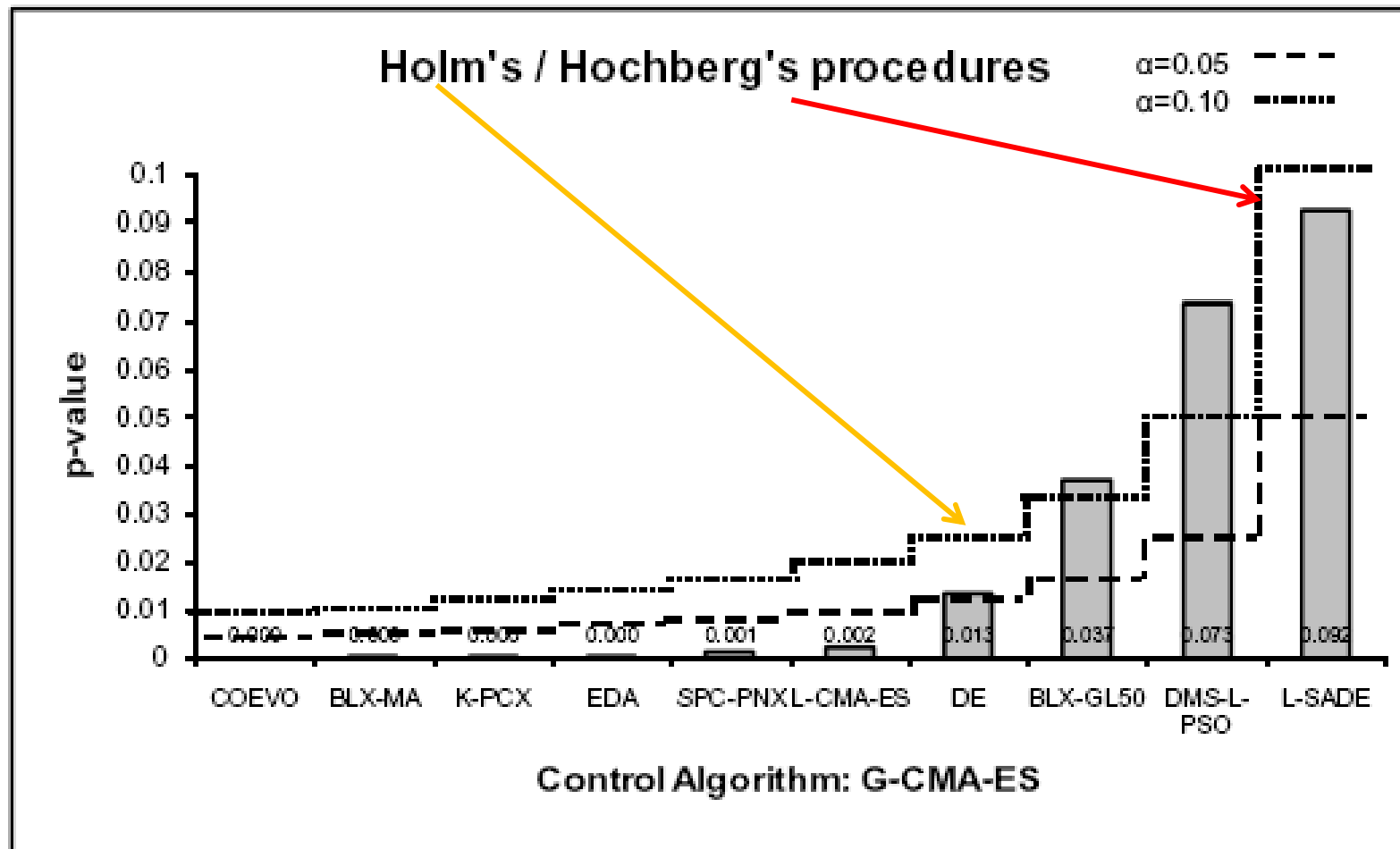


Fig. 11. Holm's/Hochberg's procedure for all functions (f1-f25).

Cases of study: CEC'2005 Special Session on real parameter optimization

Table 10: p -values on functions f1-f25 (G-CMA-ES is the control algorithm)

G-CMA-ES vs.	z	unadjusted p	Bonferroni-Dunn p	Holm p	Hochberg p
CoEVO	5.43662	$5.43013 \cdot 10^{-8}$	$5.43013 \cdot 10^{-7}$	$5.43013 \cdot 10^{-7}$	$5.43013 \cdot 10^{-7}$
BLX-MA	4.05081	$5.10399 \cdot 10^{-5}$	$5.10399 \cdot 10^{-4}$	$4.59359 \cdot 10^{-4}$	$4.59359 \cdot 10^{-4}$
K-PCX	3.68837	$2.25693 \cdot 10^{-4}$	0.002257	0.001806	0.001806
EDA	3.62441	$2.89619 \cdot 10^{-4}$	0.0028961	0.002027	0.002027
SPC-PNX	3.28329	0.00103	0.0103	0.00618	0.00618
L-CMA-ES	3.07009	0.00214	0.0214	0.0107	0.0107
DE	2.47313	0.01339	0.1339	0.05356	0.05356
BLX-GL50	2.08947	0.03667	0.3667	0.11	0.09213
DMS-L-PSO	1.79089	0.07331	0.7331	0.14662	0.09213
L-SaDE	1.68429	0.09213	0.9213	0.14662	0.09213

- In practice, Hochberg's method is more powerful than Holm's one (but this difference is rather small), in this the results are in favour of Hochberg's method.

Source: [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).

Design of Experiments in Data Mining/Computational Intelligence

Using non-parametric tests: Data Mining/
Computational Intelligence based case studies

- Two sample tests
- Multiple comparisons
- Evolutionary Algorithms: CEC'05 Special Session on parameter optimization
- Neural network and genetic learning experiments

Neural Network and Genetics-Based Machine Learning Experiments

NN and GBML does not verify parametric conditions.

Similar studies can be presented with them.

J. Luengo, S. García, F. Herrera, **A Study on the Use of Statistical Tests for Experimentation with Neural Networks: Analysis of Parametric Test Conditions and Non-Parametric Tests.** *Expert Systems with Applications, in press (2008).*

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*

Cases of study: Genetics-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?**

Cases of study: Genetics-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

Cases of study: Genetics-based machine learning

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 "*"

Cases of study: Genetics-based machine learning

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

Cases of study: Genetics-based machine learning

Analyzing parametric tests

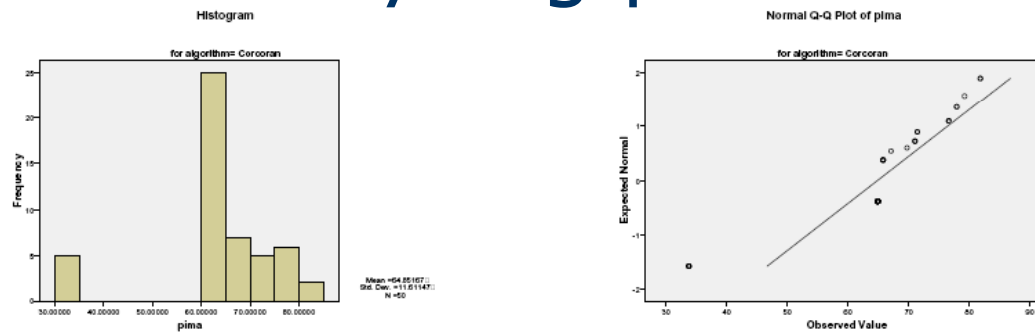


Figure 1: Results of Pitts-GIRLA over pima data set in 10fcv: Histogram and Q-Q Graphic.

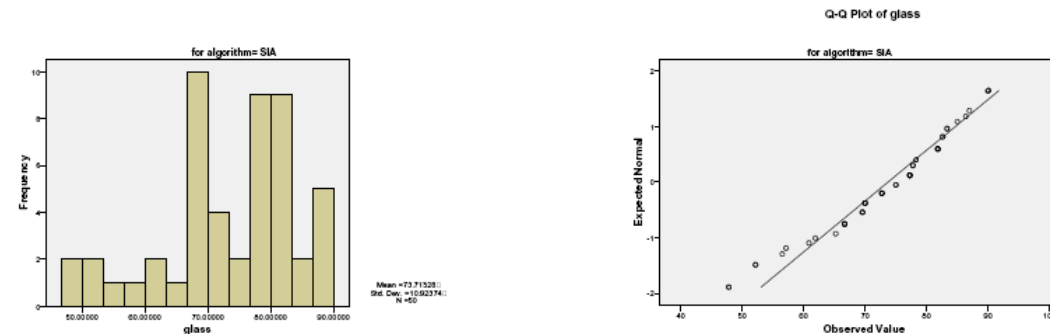


Figure 2: Results of SIA over glass data set in 10fcv: Histogram and Q-Q Graphic.

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

Cases of study: Genetics-based machine learning

**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)	(.38)	(.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 homogeneous (we reject the null hypothesis).

Cases of study: Genetics-based machine learning

Wilcoxon Signed-Ranks Test for Paired Samples

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

Table 11 Wilcoxon's test applied over the all possible comparisons between the five algorithms in classification rate

Comparison	Classification rate		
	R^+	R^-	p -value
Pitts-GIRLA - XCS	0.5	104.5	0.001
Pitts-GIRLA - GASSIST-ADI	0	105	0.001
Pitts-GIRLA - HIDER	1	104	0.001
Pitts-GIRLA - CN2	6	99	0.004
XCS - GASSIST-ADI	89	16	0.022
XCS - HIDER	53	52	0.975
XCS - CN2	78	27	0.109
GASSIST-ADI - HIDER	20	85	0.041
GASSIST-ADI - CN2	52	53	0.975
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

Cases of study: Genetics-based machine learning

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	p-value	Iman-Davenport Value	Value in F_F	p-value
Classification rate	28.987	9.487	< 0.0001	13.920	2.55	< 0.0001
Cohen's kappa	26.729	9.487	< 0.0001	11.871	2.55	< 0.0001

- 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

Cases of study: Genetics-based machine learning

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

Classification rate (XCS is the control)					
i	algorithm	unadjusted p	PH_{Bonf}	PH_{Holm}	PH_{Hoch}
1	Pitts-GIRLA	$1.745 \cdot 10^{-6}$	$6.980 \cdot 10^{-6}$	$6.980 \cdot 10^{-6}$	$6.980 \cdot 10^{-6}$
2	CN2	0.01428	0.08711	0.04283	0.04283
3	GASSIST-ADI	0.02702	0.10810	0.05405	0.05405
4	HIDER	0.67571	1.00000	0.67571	0.67571
Cohen's kappa (XCS is the control)					
i	algorithm	unadjusted p	PH_{Bonf}	PH_{Holm}	PH_{Hoch}
1	Pitts-GIRLA	$5.576 \cdot 10^{-5}$	$2.230 \cdot 10^{-5}$	$2.230 \cdot 10^{-5}$	$2.230 \cdot 10^{-5}$
2	CN2	0.01977	0.07908	0.05931	0.05931
3	GASSIST-ADI	0.13517	0.54067	0.27033	0.27033
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), in this our study the results are the same.

Design of Experiments in Data Mining/Computational Intelligence

Outline

- Introduction
- Conditions for the safe use of parametric tests
- Using non-parametric tests: Data Mining/
Computational Intelligence based case studies
 - Two sample tests/Multiple comparisons
 - Evolutionary Algorithms: CEC'05 Special Session on parameter optimization
 - Neural network and genetic learning experiments
- **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.

Lessons learned

Wilcoxon's test

- ❑ 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.

Lessons learned

Multiple comparison

- ❑ A multiple comparison must be carried out first by using a statistical method for testing the differences among the related samples means. Then to 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.

Lessons learned

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.**

Lessons learned

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

Lessons learned

Advantages of Nonparametric Tests

(only one run for all cases of test)

If we have a set of data sets/benchmark functions, we must apply a parametric test for each data set/benchmark function.

We only need to use a non-parametric test for comparing the algorithms on the whole set of benchmarks.

Concluding Remarks

- ❑ Nonparametric tests are a very useful tool for comparing algorithms in a design of experiments in Computational Intelligence and HAIS.
- ❑ This talk presents the use of nonparametric tests for comparing a control algorithm against a set of algorithms.
- ❑ There are also nonparametric tests for comparing a sets of algorithms based on ranking, procedures for performing all pairwise comparisons, among them: Nemenyi, Shaffer, Bergmann-Hommel procedures.
- ❑ There are other kind of nonparametric algorithms as: permutation based procedures, etc.

Concluding Remarks

More on Nonparametric Tests

All pairwise comparisons

S. García, F. Herrera, An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *Journal of Machine Learning Research, in press (2008)*.

Proposals of statistical procedures for comparing $n \times n$ classifiers. An easy way of obtaining adjusted and comparable p-values in multiple comparison procedures

Concluding Remarks

More on Nonparametric Tests

All pairwise comparisons

Nemenyi, 1963

Shaffer, 1986

Bergmann and Hommel, 1988

<http://sci2s.ugr.es/publications/ficheros/garcia08a-JMLR.pdf>

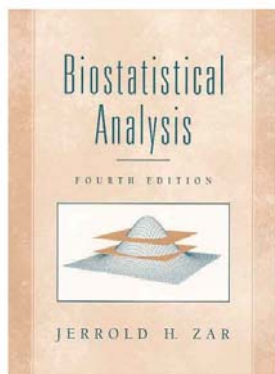
Concluding Remarks

Design of Experiments in Data Mining/Computational Intelligence

They are not the objective of our talk, but they are two additional important questions:

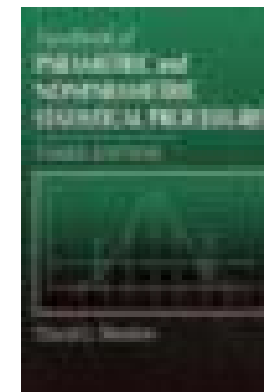
- ❑ Benchmark functions/data sets ... are very important.
- ❑ To compare with the state of the art is a necessity.

Bibliography



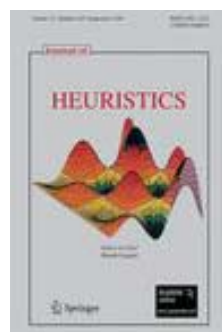
J.H. Zar, Biostatistical Analysis, Prentice Hall, 1999.

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



JMLR

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



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*, *in press* (2008).

Design of Experiments in Data Mining/Computational Intelligence

Thanks!!!

