On statistical multiple comparison procedures and their application to comparing classifiers
Materials for ADA Lab ICM UW 2016-11-22 seminar

Mateusz Kobos

2016-11-22

Contents

1 Basic information about statistical tests

- In the tasks of **statistical hypothesis testing**, we have 2 hypotheses: **null hypothesis** H_0 and **alternative hypothesis** H_1. For example the null hypothesis H_0 might say that averages of random samples are equal and alternative hypothesis H_1 might say that they are not equal. We want to verify H_0 and thus possibly reject it which will mean accepting H_1.
- **Significance level** (pol. *poziom istotności*) of a test is a probability α of rejecting the null hypothesis H_0 when it is true (its value is usually chosen to be $\alpha = 5\%$).
- Types of relevant errors:

  - **Type I error** – rejecting the null hypothesis H_0 when it is true. The probability of Type I error is the significance level.
  - **Type II error** – rejecting the alternative hypothesis H_1 when it is true.

- **Power of a test** is the probability that the test correctly rejects the null hypothesis H_0 when the alternative hypothesis H_1 is true. This is equal to $1 - \text{``probability of Type II error''}$.

- Note that **statistically significant difference might not be practically significant**, i.e. the difference might be too small to have any practical consequences.

Example statistical test: t-test. t-test is one of the more popular tests; there are two basic variants:

- If we have **2 (unpaired) random samples** $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1)$ and $Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2)$, we can use the **t-test** to verify if the averages μ_1 and μ_2 are equal.
  - An example: $X_1, \ldots, X_{n_1}$ is a result of examination in a group of patients after a treatment and $Y_1, \ldots, Y_{n_2}$ is a result of examination in a control group
- If we have **2 paired random samples** $(X_1, Y_1), \ldots, (X_n, Y_n)$, where **the pairs have the same 2-dimensional normal distribution**, we can use the **paired t-test** to verify the equality of the averages. Note that the pairs are independent of each other, but the variables in the pair might be dependent.
  - An example: (X_i, Y_i) is a result of examination of patient i before (variable X_i) and after (variable Y_i) a treatment.

2 Introduction to Multiple Comparison Procedures

Motivational example.

- When you test a single statistical hypothesis at a 5% significance level, there's a 5% chance of rejecting the null hypothesis (H_0) given that it is true.
- When you test 100 statistical hypotheses and each of them is done at a 5% significance level, in 5 of them on average the null hypothesis (H_0) will be rejected even though it is true[1].

This is not what we want. We need to somehow control the significance level that encompasses the whole experiment, a.k.a. **family-wise error rate** α_{FW}.

There are **various approaches** to control this error [wikipedia16a], a.k.a. **Multiple Comparison Procedures (MCPs)**:

1. Methods where α_{FW} can be proved to never exceed given value.

2. Methods where α_{FW} can be proved not to exceed given value except under certain conditions.
3. Methods which rely on an single “omnibus test” (e.g. ANOVA) before proceeding to individual tests. These methods have “weak” control of α_{FW}.
4. Empirical methods, which control α_{FW} adaptively.

We’re going to concentrate on point 1 (Sect. 3) and point 3 (Sect. 4) here.

Note that there is an alternative technique called **False Discovery Rate** used, e.g. in genomic microarray research. It is more appropriate for exploratory research or when the results can be easily re-tested in an individual study [wikipedia16a]. We’re not going to discuss it here though.

3 Bonferroni correction and related approaches

If we want to execute k related tests (family of tests), a.k.a. **comparisons** (since they usually involve comparing averages of two or more samples with one another), we are interested in [Sheskin07, Test 21, p. 874-875]:

- α_{FW} - **familywise Type I error rate** - likelihood that at least one of the null hypotheses will be rejected given that it is true.
- α_{PC} - **per comparison Type I error rate** - likelihood that the null hypothesis in any single comparison will be rejected given that it is true.

Assuming that the comparisons are independent we have:

$$\alpha_{FW} = 1 - (1 - \alpha_{PC})^k$$

thus

$$\alpha_{PC} = 1 - \sqrt[k]{1 - \alpha_{FW}} \ .$$

This is called the **Sidak-Bonferroni** correction [Sheskin07, p. 891]. For simplicity this can be approximated as

$$\alpha_{PC} = \frac{\alpha_{FW}}{k} \ .$$

This is called the **Bonferroni correction**. We can use significance levels α_{PC} in individual hypotheses tests[2].

Properties of the Bonferroni correction:

- One of the most popular methods of dealing with the problem of multiple comparisons.
- It is very conservative (it reduces likelihood of committing Type I error but at the expense of increasing the likelihood of committing a Type II error) [Sheskin07, p. 875] [Koronacki01, p. 336].

[2] The α_{PC} approximation is slightly more conservative than the original formula [Sheskin07, endnote 16, p. 969]. For example for $c = 100$ and $\alpha_{FW} = 0.05$: for the original formula we have $1 - \sqrt[100]{1 - 0.05} \approx 0.00051$ while for the approximation we have $0.05/100 = 0.0005$.

- **Holm-Bonferroni procedure** is both more powerful than Bonferroni procedure and doesn't make any additional assumptions [wikipedia16b]. It consists of an algorithm that analyzes considered hypotheses in sequence.
- **Hochberg and Hommel procedures** are examples of more powerful methods but they require additional assumptions [wikipedia16b].

4 ANOVA

In the **analysis of variance (ANOVA)** test[3] we have:

- the null hypothesis H_0: the averages of different random samples are equal,
- the alternative hypothesis H_1: there are at least two samples with different averages.

Example [Koronacki01, Sect. 5.1]. We have couple **different types of margarine** and we want to know if all of them have the same average amount of saturated fat (we generally expect that it is true due to market competition). Here we're interested in:

- **response variable** (pol. *zmienna odpowiedzi*) - the amount of fat (in the analysis of regression we call it the dependent/output variable). It depends on
- **factor** (pol. *poziom*) - type of margarine (in the analysis of variance we call it the explanatory/independent/input variable).

We gather couple measurements of the amount of fat for each factor level and apply ANOVA to answer the question.

Ways of categorizing ANOVA tests (see Fig. 1):

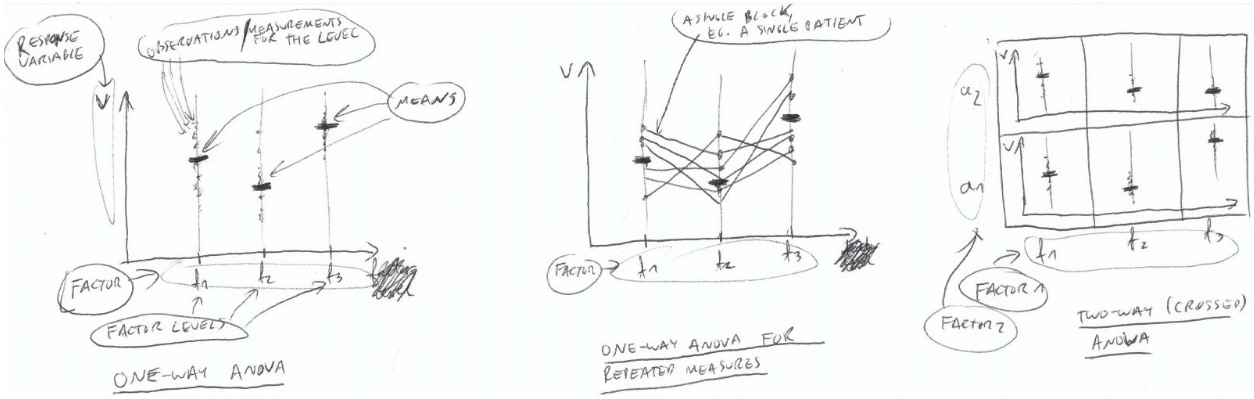

Figure 1: Symbolic representation of selected types of ANOVA.

[3] It was developed by Ronald Fisher [Sheskin07, p. 865].

- With respect to **number of factors**: one-factor, two-factor, etc. ANOVA (pl. analiza jednoczynnikowa, analiza, dwuczynnikowa etc.) – the response variable might depend on one, two or more factors.

- With respect to **existence of blocks**: independent measures (without blocks), repeated measures (with blocks).

- Being **parametric or non-parametric**. In case of parametric version, it is assumed that the samples come from the normal distribution; in case of non-parametric version, the form of the distribution is not assumed.
  - When we say "ANOVA", we usually have the parametric version (with normal distribution) in mind. The **F-test** is used in this case [Koronacki01, Sect. 5.2.1].
    - Note that before using it, you need to check if the basic assumptions of this approach hold: each factor level is normal (Shapiro-Wilk test or quantile-quantile plot) and has the same variance (Lavene or Bartlett test).
  - Non-parametric ANOVAs are usually just referred to with the names of the particular tests used, e.g. Wilcoxon test, Friedman test.

The basic **non-parametric ANOVA tests** [Koronacki01, Sect. 9.5]:

- in case of indepedent measures: **Kruskall-Wallis test**
  - (you can use **Bonferroni procedure** as post hoc comparisons after rejecting the null hypothesis)

- in case of dependent measures: **Friedman test**

5 Post Hoc Procedures

Note that if we apply ANOVA and reject the null hypothesis, we don't know which averages differ. We need to use additional tests – called **post hoc procedures** – to find out. The **main difference between these procedures** is whether we want to control α_{FW} or not, and if so, to what degree.

Note that there might be cases when [Sheskin07, p. 876, 3rd paragraph]:

- ANOVA says that there are statistically significant differences between the averages but the post hoc precedure is not able to find any statistically significant differences.

- ANOVA says that there are no statistically significant differences between the averages but the post hoc procedure shows statistically significant differences between some averages.

5.1 Planned and unplanned experiments

(This section is based mostly on [Sheskin07, p.875-876]).

We usually use post hoc comparisons after rejecting the null hypothesis in ANOVA test (parametric or not) in order to find out which averages differ significantly. However, there is no agreement among the researchers on how to compare these averages [Sheskin07, p. 874, the 3rd paragraph from the bottom]. In general, this depends on whether these comparisons have been planned or not [4]:

- In **planned comparisons** (a.k.a. a priori comparisons), the researcher knows which averages he wants to compare before gathering the data.
- In **unplanned comparisons** (a.k.a. post hoc, or a posteriori comparisons), the researcher gathers data, compares the averages of the groups and based on this information decides which pairs of averages he would like to inspect with respect to the significance of the difference.

More details:

- In the **planned comparisons**:
  - The researcher can do the comparisons that he planned regardless of the result of ANOVA test (the sources agree on that), some researchers say that you don't even need to do the ANOVA test.
  - **It's not necessary to adjust the significance level of individual tests with respect to familywise significance level** (most sources agree on that). However, in certain cases (when there are lots of planned comparisons) one can argue that adjusting the significance level is necessary.
    - To be more precise, you can skip adjusting the significance level if the number of planned comparisons is not larger than $df_{BG} = \text{“number of factors”} - 1$ [Sheskin07, p. 883, paragraph 1], [Sheskin07, p. 872, eq. 21.8]. However, if it's larger, you should apply the adjustment.
- In the **unplanned comparisons**:
  - For many years, the consensus was that you are allowed to execute post hoc comparisons only if ANOVA test resulted earlier in rejecting the null hypothesis. However, recently many researchers (including the author of [Sheskin07]) claim that you can apply post hoc comparisons regardless of the result of ANOVA test.
  - **You need to adjust the significance level of individual tests with respect to familywise significance level.** The sources agree; however, there is no consensus on how exactly this should be done.
    - Some sources claim that using familywise significance level of $\alpha_{FW} = .05$ is too conservative and you can use values as large as $\alpha_{FW} = .25$. Note that when the number of factor levels (groups) is $k = 3$, it would be hard to justify assuming $\alpha_{FW} = .25$; however, this might make sense when the number of factor levels (groups) is $k = 10$ [Sheskin07, p. 876].

See [Sheskin07] for a list of test one can apply in case of both planned and unplanned comparisons.

[4] See [Sheskin07, p. 877, paragraphs 2 and 3] for a motivational example showing why it is important to distinguish between these two cases and why it is much more likely to commit Type I error in case of unplanned experiments.

5.2 Post hoc procedures after rejecting null hypothesis in one-factor ANOVA

After rejecting null hypothesis in one-factor ANOVA, we can use a **modified version of t-test** to compare pairs of averages [Koronacki01, Sect. 5.2.3]. The variance is estimated better since it takes into consideration variances on all factor levels.

There are approaches to computing the per comparison significance level α_{PC} in this case [Koronacki01, Sect. 5.2.3]:

- **Bonferroni procedure**. It's disadvantage: it's very conservative. Hochberg procedure is an improved version of this procedure [VanBelle02, Sect. 6.12].
- **Tukey's procedure**. It's recommended if the samples have the same size [Koronacki01, p. 336] [VanBelle02, Sect. 6.12]. It's a very popular method [Montgomery00, p. 102].
- **Scheffe procedure**. It's recommended if the samples do not have the same size [Koronacki01, p. 336] [VanBelle02, Sect. 6.12] but it's very conservative [VanBelle02, Sect. 6.12]. It's less powerful than Tukey's procedure if the samples have the same size [VanBelle02, Sect. 6.12].

6 Comparing classifiers

You can find description of procedures that are recommended and not recommended for comparing classifiers in [Demsar06], [Garcia08] (the latter is accompanied by "scientific-quality" software that implements these tests).

In this section we're going to consider cases where the algorithms are tested on one (Sect. 6.2) and more (Sect. 6.3) data sets.

6.1 Introduction

The general requirement. The general requirement for the methods presented below is that the algorithm has to be tuned on a specially selected **tuning set** – a surrogate of the real testing set. When the algorithm training has finished i.e. its parameters are fixed, it can be tested using the testing set, which yields the final algorithm error [Salzberg97].

The probability of type I error is increased when sample variables are dependent. In statistical tests, we usually assume that random variables $X_1, \ldots, X_n$ constituting random sample are independent.

- But if this assumption is not true, their realizations land close by (in an extreme situation of $X_1 = \ldots = X_n$ they would land in the same point). Thus, their variance will be smaller than the variance in a situation of independent random variables.
- However, in statistical test we estimate the variance using the sample points and an assumption of the independence. If the points are dependent, their variance is small, and as a result, the variance is underestimated.
- **If the variance is underestimated, it increases the probability of type I error in the test** (more than the nominal significance level α would suggest) [Nadeau03, p. 240, p. 276].

6.2 Single dataset

The standard **assumption of independence of random sample's variables doesn't hold if the same dataset is used** in experiments comparing the classifiers. This is because, e.g in a 10-fold cross-validation evaluation method we have 10 results but they are not independent since they were generated from very similar data sets.

That's why using standard t-test in this case results in an output that is too optimistic (i.e., it is easier to commit Type I error). There are many heuristic modifications of the standard tests that try to make them more conservative but this is done on an expense of statistical power of the test (i.e. it's more difficult than necessary to reject the null hypothesis – the Type II error is higher than nominal significance level set).

- **Comparison of 2 algorithms – recommended tests** from the best one to the worst one:
  - **10 × 10-fold CV (cross-validation) with corrected resampled (paired) t-test** (result on each fold is a single value, i.e. we have two samples, each with 100 elements) [Witten05, p.157]. The corrected resampled (paired) t-test is a partially heuristic modification of the paired t-test; this test is called "corrected repeated k-fold cv test" in [BouckaertFrank04].
  - **100 × holdout with corrected resampled (paired) t-test** – a method proposed in [Nadeau03]. According to [BouckaertFrank04] the method is characterized by slightly worse replicability than 10 × 10-fold CV method.
- **Comparison of 2 algorithms – NOT recommended tests**
  - 10-fold CV with paired t-test (result on each fold is a single value) [Dietterich98] – quite high Type I error
  - 10 × 10-fold CV with sign test (average over runs is a single value) [Bouckaert04] – worse replicability than 10 × 10 CV with (paired) t-test
  - 10 × 10-fold CV with paired t-test (result on each fold is a single value) [Bouckaert04] – very high Type I error
  - 10 × 10-fold CV with paired t-test (average from a single CV evaluation is a single value) [Bouckaert04] – high Type I error
  - 100 × holdout with paired t-test [Bouckaert04] – high Type I error
  - holdout with difference of two proportions test [Dietterich98]
  - 30 × holdout with paired t-test [Dietterich98] – very high Type I error
- **Comparison of many algorithms – recommended tests**
  - **30 × 2-fold CV with (parametric or nonparametric) (independent measures or repeated measures) ANOVA** (average from a CV is a single value) with 10 internal repeats for non-deterministic algorithms [Pizarro02]. Specific post hoc tests should be used to select the best algorithm.
    - ∗ My comment: But the samples are not independent which is against the assumptions of the test, and this method doesn't deal with this problem.

  - In [Salzberg97] the same method as in comparison of 2 algorithms, i.e. **CV with Binomial test**, is proposed.
    - My comment: Probably the **Bonferroni adjustment** should be used here to compensate for many experiments performed, but its disadvantage is that it makes the test weak.

6.3 Multiple datasets

When doing experiments on multiple datasets, the average result for a dataset of a given algorithm is treated as a single value. This is a much simpler case than the one with a single dataset, since the assumption of independence is not violated (due to different data sets used).

- **comparison of 2 algorithms – recommended tests**
  - **Wilcoxson signed-rank test** – a standard non-parametric test [Demsar06]
- **comparison of 2 algorithms – NOT recommended tests**
  - **Comparing averages computed over different datasets** [Demsar06] – the results on different datasets are not comparable, thus their averages are meaningless. The average is also prone to outliers.
  - **Paired t-test** [Demsar06] – disadvantages: 1) the results on different datasets are not comparable which makes t-test meaningless, 2) samples are (generally) not normally distributed which is against the assumptions of the test, 3) the test is prone to outliers which decrease test's power.
- **comparison of many algorithms – recommended test**
  - **Friedman test**, or, even better, its modification by **Iman and Davenport** which yields a more powerful test [Demsar06] [Garcia08]. After rejection of the null hypothesis using this test, post hoc tests can be executed:
    - comparison of all classifiers with the base classifier
      - **Holm's test** [Demsar06]
      - **Bonferroni-Dunn test** [Demsar06] – less powerful than Holm's test but easier to visualize
    - comparison of all classifiers with each other
      - **Nemenyi test** [Demsar06] – a very conservative test [Garcia08]
      - **Shaffer static procedure** [Garcia08] – a more powerful procedure than the Nemenyi test
      - **Bergmann-Hommel procedure** [Garcia08] – the best perfoming procedure, but also the most difficult to understand and computationally expensive. This procedure examines logical relations between individual tests and excludes from considerations those combinations of the results that are impossible.
- **comparison of many algorithms – NOT recommended tests**

- **Repeated-measures parametric ANOVA** [Demsar06] – disadvantages: 1) assumption that samples come from a normal distribution is (generally) not met. However, many statisticians would not object using parametric ANOVA unless the distributions were, for instance, clearly bi-modal [Demsar06, p.10]. 2) sphericity assumption (a property analogous to property that all the variances are the same in independent measures parametric ANOVA) is (generally) not met.
  - ∗ comparison of all classifiers with the base classifier
    - · **Dunnett test** [Demsar06]
  - ∗ comparison of all classifiers with each other
    - · **Tukey test** [Demsar06]

See Fig.2 for an illustration of comparing many algorithms using many datasets.

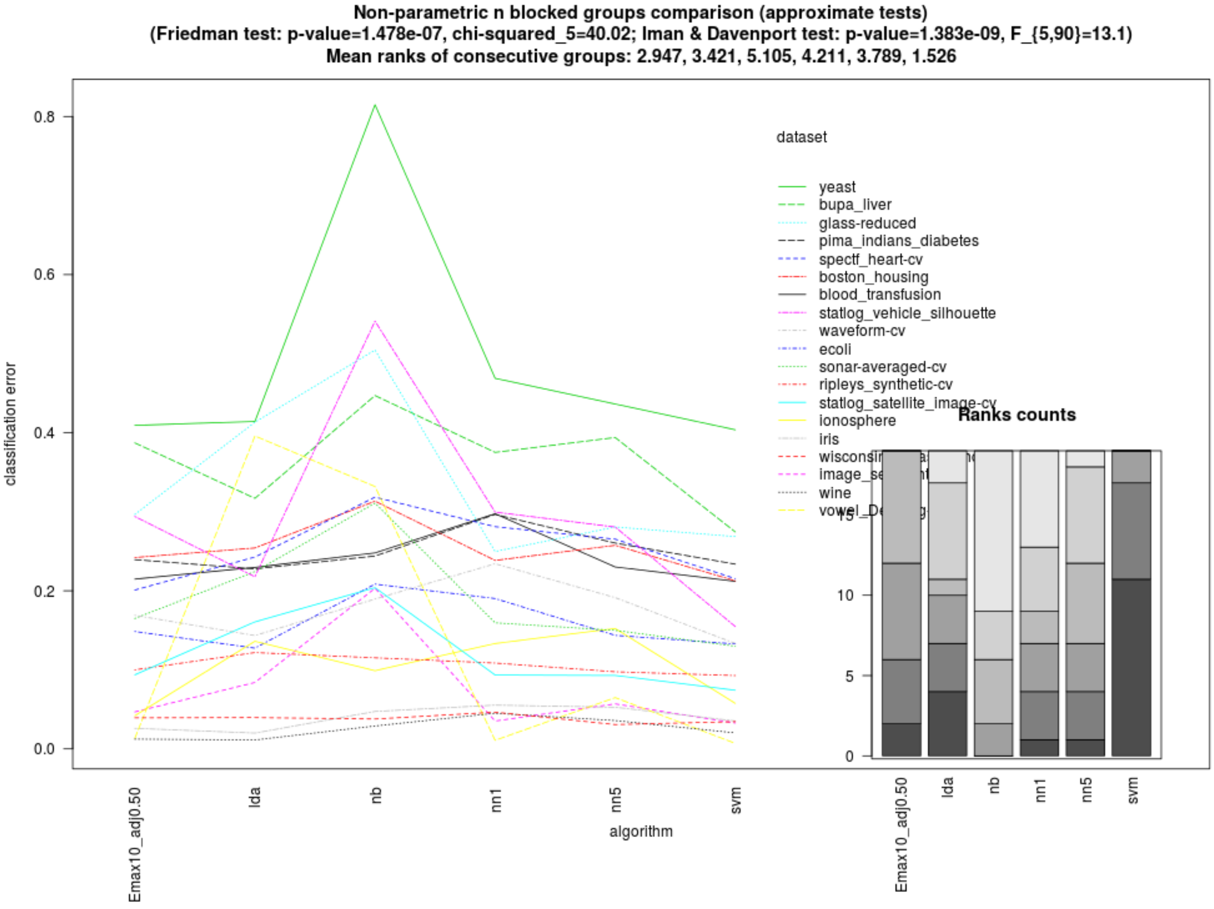

Figure 2: An example visualization of comparison of couple classifiers on different data sets from my PhD-related experiments.

7 Final comments and conclusions

My comments:

- From my experience, MCPs are rarely used in the literature (in the field of experimental computer science, machine learning).
- In my PhD-related work I used: 1) Iman and Davenport tests followed by Holm procedure, 2) Bonferroni correction, and 3) corrected resampled paired t-test. It was much easier to obtain statistically significant results without using MCPs!

Conclusions:

- **Limit the number of statistical tests** [Sheskin07, p. 882, last paragraph] – test only what is necessary.
- **Consider using ANOVA followed by post hoc procedure** when comparing the results of many experiments at the same time.
- **Consider using Bonferroni or Holm-Bonferroni corrections** when doing many statistical test at the same time.
- It's best to **find the description of statistical procedure for your use-case** in the literature since otherwise it's pretty easy to use one in a situation where the basic assumptions aren't met and thus obtain erroneous results.

References

[BouckaertFrank04] Bouckaert, Frank: "Evaluating the replicability of significance tests for comparing learning algorithms", PAKDD, 2004

[Bouckaert04] Bouckaert: "Estimating replicability of classifier learning experiments", ICML conference, 2004

[Demsar06] Demsar: "Statistical comparisons of classifiers over multiple data sets", Journal of Machine Learning Research, 2006

[Dietterich98] Dietterich: "Approximate statistical tests for camparing supervised classification learning algorithms", Neural Computation, 1998

[Garcia08] Garcia, Herrer: "An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons", Journal of Machine Learning Research, 2008

[Koronacki01] Koronacki, Mielniczuk: "Statystyka dla kierunków technicznych i przyrodniczych", WNT, 2001

[Montgomery00] Montgomery: "Design and Analysis of Experiments", 5th edition, Wiley, 2000

[Nadeau03] Nadeau, Bengio: "Inference for the Generalization Error", Machine Learning, 52, 239–281, 2003

[Pizarro02] Pizarro, Guerrero, Galindo: "Multiple comparison procedures applied to model selection", Neurocomputing, 2002

[Salzberg97] Salzberg: "On comparing classifiers: pitfalls to avoid and a recommended approach", Data Mining and Knowledge Discovery, 1997

[Sheskin07] Sheskin: "Handbook of parametric and nonparametric statistical procedures", 4th edition, Chapman & Hall /CRC, 2007

[VanBelle02] van Belle: "Statisical rules of thumb", 1st edition, Wiley, 2002

[wikipedia16a] Wikipedia "Multiple Comparisons Problem", https://en.wikipedia.org/wiki/Multiple_comparisons_problem, access time: 2016-11-20

[wikipedia16b] Wikipedia "Holm-Bonferroni method", https://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method, access time: 2016-11-20

[Witten05] Witten, Frank: "Data mining, practical machine learning tools and techniques", 2nd ed., 2005