Introduction to classifier comparison

Javier Sáez

Universidad de Granada

January 18, 2023

Index

Notation

- D is a classifier (may have a sub-index)
- $E_{i,j}$ refers to the error of the classifier i in the partition/dataset j.

Book: Kuncheva [2014]

General Guidelines

- Choose and fix the procedures in advance.

General Guidelines

- Choose and fix the procedures in advance.
- Compare modified versions of classifiers with the original one. Try not to compare very different classifiers.

General Guidelines

- Choose and fix the procedures in advance.
- Compare modified versions of classifiers with the original one. Try not to compare very different classifiers.
- Make sure that all the information is used by all the classifiers (avoid clever initialisations).

General Guidelines

- Choose and fix the procedures in advance.
- Compare modified versions of classifiers with the original one. Try not to compare very different classifiers.
- Make sure that all the information is used by all the classifiers (avoid clever initialisations).
- Do NOT look at test data.

General Guidelines

- Choose and fix the procedures in advance.
- Compare modified versions of classifiers with the original one. Try not to compare very different classifiers.
- Make sure that all the information is used by all the classifiers (avoid clever initialisations).
- Do NOT look at test data.
- Give also the complexity of the classifier: training and running times, memory requirements, computational requirements.

Two classifiers in one fixed set - McNemar test (Continuity corrected version) [Dietterich, 1998]

	D_2 correct	D_2 wrong
D_1 correct	N_{11}	N_{10}
D_1 wrong	N_{01}	N_{00}

Two classifiers in one fixed set - McNemar test (Continuity corrected version) [Dietterich, 1998]

	D_2 correct	D_2 wrong
D_1 correct	N_{11}	N_{10}
D_1 wrong	N_{01}	N_{00}

$H_0 \equiv$ there is no difference between the accuracies.

$$s = \frac{(|N_{01} - N_{10}| - 1)^2}{N_{01} + N_{10}} \approx \chi^2(1)$$

Given α, if $s > F^{-1}_{\chi^2(1)}(1 - \alpha)$, we reject $H_0 \implies$ the classifiers have significantly different accuracies.

Sources of variation in classifier metrics

- Choice of testing set. Single experiment might lead to not very accurate results

Sources of variation in classifier metrics

- Choice of testing set. Single experiment might lead to not very accurate results
- Choice of training sets (unstable classifiers)

Sources of variation in classifier metrics

- Choice of testing set. Single experiment might lead to not very accurate results
- Choice of training sets (unstable classifiers)
- Randomness of the training algorithm

Sources of variation in classifier metrics

- Choice of testing set. Single experiment might lead to not very accurate results
- Choice of training sets (unstable classifiers)
- Randomness of the training algorithm
- Randomly mislabeled objects

Sources of variation in classifier metrics

- Choice of testing set. Single experiment might lead to not very accurate results
- Choice of training sets (unstable classifiers)
- Randomness of the training algorithm
- Randomly mislabeled objects

Simple suggestion: use multiple training and test sets!

Two models - single data set [Nadeau and Bengio, 1999]

Using a single dataset it is common to partition it and run experiments multiple times.

T-test: test whether the means of two populations are different

Two models - single data set [Nadeau and Bengio, 1999]

Using a single dataset it is common to partition it and run experiments multiple times.

T-test: test whether the means of two populations are different

Problem: errors in the T testing partitions are not completely independent (K-fold)

Two models - single data set [Nadeau and Bengio, 1999]

Using a single dataset it is common to partition it and run experiments multiple times.

T-test: test whether the means of two populations are different

Problem: errors in the T testing partitions are not completely independent (K-fold)

$$d_j = E_{1,j} - E_{2,j}, \quad \forall j = 1, \ldots, T$$

$H_0 \equiv$ mean of these differences is 0.

Two models - single data set [Nadeau and Bengio, 1999]

Using a single dataset it is common to partition it and run experiments multiple times.
T-test: test whether the means of two populations are different
Problem: errors in the T testing partitions are not completely independent (K-fold)

$$d_j = E_{1,j} - E_{2,j}, \quad \forall j = 1, \dots, T$$

$H_0 \equiv$ mean of these differences is 0.
Std of mean difference:

"Independent"
$\sigma'_d = \frac{\sigma_d}{\sqrt{T}}$

One split
$\sigma'_d = \sigma_d \sqrt{\frac{1}{T} + \frac{N_{\text{testing}}}{N_{\text{Training}}}}$

K-fold
$\sigma'_d = \sigma_d \sqrt{\frac{1}{K} + \frac{1}{K-1}}$

Two models - single data set [Nadeau and Bengio, 1999]

Algorithm:

1. Calculate d_j, and then the mean m_d and standard deviation s_d (empirical)
2. Calculate the amended standard error s'_d as one of the previous cases
3. Calculate the test statistic $t_d = \frac{m_d}{s'_d}$ and the degrees of freedom $df = T - 1$.
4. Calculate the p-value:
   - Two tailed t-test: $p = 2F_t(-|t_d|, df)$
   - Set $H_1 \equiv$ "D_1 has lower error than D_2", one tailed test, $p = F_t(t_d, df)$
5. Reject H_0 if $p < \alpha$

Two models - multiple datasets: Wilcoxon signed rank test

T-test not appropiate: errors in different dataset are hardly commensurable.

Let $d_j = E_{1,j} - E_{2,j},\ \forall j = 1, \ldots, N$ be the difference of the errors in the N datasets.

- $H_0 \equiv$ the components of the vector $\mathbf{d} = (d_1, \ldots, d_N)$ come from a continuous, symmetric distribution with zero median.
- $H_1 \equiv$ the distribution does not have zero median.

Scipy implementation

Wilcoxon signed rank test

1. Rank the absolute values of the distances $|d_i|$ in **increasing order**
2. If positions $j, \ldots, j+k$ are tied, the rank of **all** of them becomes the mean of the ranks. Each dataset will have a rank r_i.
3. Split ranks into positive and negative depending on the sign of d_i, and calculate the sums:

$$R^+ = \sum_{d_i>0} r_i + \frac{1}{2}\sum_{d_i=0} r_i, \quad R^- = \sum_{d_i<0} r_i + \frac{1}{2}\sum_{d_i=0} r_i$$

4. Take as the test statistic $T = \min(R^+, R^-)$

Wilcoxon signed rank test

1. Rank the absolute values of the distances $|d_i|$ in **increasing order**
2. If positions $j, \ldots, j+k$ are tied, the rank of **all** of them becomes the mean of the ranks. Each dataset will have a rank r_i.
3. Split ranks into positive and negative depending on the sign of d_i, and calculate the sums:

$$R^+ = \sum_{d_i > 0} r_i + \frac{1}{2} \sum_{d_i = 0} r_i, \quad R^- = \sum_{d_i < 0} r_i + \frac{1}{2} \sum_{d_i = 0} r_i$$

4. Take as the test statistic $T = \min(R^+, R^-)$

Check the value of the statistic in a *Wilcoxon* table. It is special due to the discrete nature of the Binomial distribution.

Multiple models - multiple datasets: Friedman test

Consider that we have N datasets and M classifiers. Algorithmically, the test can be summarized as:

1. Rank the classifiers in each of the N datasets. Ties are shared equally as in the previous test. Let r_i^j be the rank of classifier j on the dataset i.

Multiple models - multiple datasets: Friedman test

Consider that we have N datasets and M classifiers. Algorithmically, the test can be summarized as:

1. Rank the classifiers in each of the N datasets. Ties are shared equally as in the previous test. Let r_i^j be the rank of classifier j on the dataset i.
2. Fixing j (a classifier), we calculate $R_j = \frac{1}{N} \sum_{i=1}^{N} R_i^j$, the average rank of model j, for each j.

Multiple models - multiple datasets: Friedman test

Consider that we have N datasets and M classifiers. Algorithmically, the test can be summarized as:

1. Rank the classifiers in each of the N datasets. Ties are shared equally as in the previous test. Let r_i^j be the rank of classifier j on the dataset i.
2. Fixing j (a classifier), we calculate $R_j = \frac{1}{N}\sum_{i=1}^{N} R_i^j$, the average rank of model j, for each j.
3. Calculate the test statistic:

$$T = \frac{12N}{M(M+1)}\left(\sum_{j=1}^{M} R_j^2 - \frac{M(M+1)^2}{4}\right) \sim \chi^2(M-1).$$

Multiple models - multiple datasets: Friedman test

Consider that we have N datasets and M classifiers. Algorithmically, the test can be summarized as:

1. Rank the classifiers in each of the N datasets. Ties are shared equally as in the previous test. Let r_i^j be the rank of classifier j on the dataset i.
2. Fixing j (a classifier), we calculate $R_j = \frac{1}{N}\sum_{i=1}^{N} R_i^j$, the average rank of model j, for each j.
3. Calculate the test statistic:

$$T = \frac{12N}{M(M+1)}\left(\sum_{j=1}^{M} R_j^2 - \frac{M(M+1)^2}{4}\right) \sim \chi^2(M-1).$$

$H_0 \equiv$ all classifier models are equivalent.
Scipy Implementation

Iman and Davenport amendment

Iman showed [Iman and Davenport, 1980] that the previous test has shown to be very conservative in many cases and proposed the following statistic:

$$F_F = \frac{(N-1)x_F^2}{N(M-1) - x_F^2} \sim F\left((M-1),(M-1)(N-1)\right)$$

Post-hoc test

H_0 rejected. Where are the differences?

Two classifiers are declared different if their average ranks differ by more than a critical value.

Post-hoc test

H_0 rejected. Where are the differences?
Two classifiers are declared different if their average ranks differ by more than a critical value.

$$z = \frac{R_i - R_j}{\sqrt{\frac{M(M+1)}{6N}}}, \quad \forall i, j = 1, \ldots, M$$

Post-hoc test

H_0 rejected. Where are the differences?
Two classifiers are declared different if their average ranks differ by more than a critical value.

$$z = \frac{R_i - R_j}{\sqrt{\frac{M(M+1)}{6N}}}, \quad \forall i, j = 1, \ldots, M$$

This statistic follows a standard Gaussian distribution.

- If we compare with all other classifiers,

$$p\text{-value} < \frac{2\alpha}{M(M-1)}$$

- If we compare one classifier with all other:

$$p\text{-value} < \frac{\alpha}{M-1}$$

Thank you for your attention

Bibliography

Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley Publishing, 2nd edition, 2014. ISBN 1118315235.

Thomas G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 1998.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. In *Advances in Neural Information Processing Systems*. MIT Press, 1999.

Ronald Iman and James Davenport. Approximations of the critical region of the friedman statistic. *Communications in Statistics-Theory and Methods*, 9:571–595, 01 1980.