# Statistical Comparisons of Classifiers over Multiple Data Sets

Peiqian Li

23.01.2008

# Outline

- Motivation
- Statistics and Tests for Comparison of Classifiers
  - Comparisons of Two Classifiers
    - Averaging over data sets
    - Paired t-test
    - Wilcoxon signed-ranks test
  - Comparisons of Multiple Classifiers
    - ANOVA
    - Friedman test
- Conclusion

# Motivation

- comparing two learning algorithms on a single data set
- comparisons of more algorithms on multiple data sets
  - more essential to typical machine learning studies
- no established procedure over multiple data sets

# Statistics and Tests for Comparison of Classifiers

- $k$ learning algorithms on $N$ data sets
- $c_i^j$: performance score of the $j$-th algorithm on the $i$-th data set
- statistically significantly different ?
- which are the particular algorithms that differ in performance
- fundamental difference
- sample size = number of data sets

# Averaging over data sets

- *"it is debatable whether error rates in different domains are commensurable, and hence whether averaging error rates across domains is very meaningful"* -- Webb (2000)
- results not comparable → averages meaningless
- susceptible to outliers

# Paired t-test

- A common way to test whether the difference is non-random
- $d_i = c_i^1 - c_i^2$
- $t$ statistic $\bar{d} / \sigma_{\bar{d}}$
  - Student distribution with $N - 1$ degrees of freedom
- Weaknesses
  - Commensurability
  - Differences distributed normally
  - affected by outliers

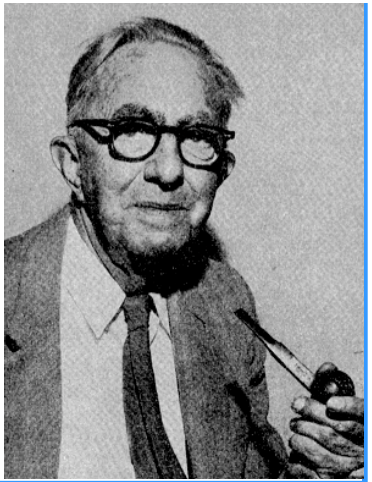
# Wilcoxon signed-ranks test

- ranks the differences for each data set
- compares the ranks for the positive and the negative differences.

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i), R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$

$$T = \min(R^+, R^-)$$

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



| | C4.5 | C4.5+m |
|---|---|---|
| adult (sample) | 0.763 | 0.768 |
| breast cancer | 0.599 | 0.591 |
| breast cancer wisconsin | 0.954 | 0.971 |
| cmc | 0.628 | 0.661 |
| ionosphere | 0.882 | 0.888 |
| iris | 0.936 | 0.931 |
| liver disorders | 0.661 | 0.668 |
| lung cancer | 0.583 | 0.583 |
| lymphography | 0.775 | 0.838 |
| mushroom | 1.000 | 1.000 |
| primary tumor | 0.940 | 0.962 |
| rheum | 0.619 | 0.666 |
| voting | 0.972 | 0.981 |
| wine | 0.957 | 0.978 |

| d | rank | |
|---|---|---|
| 0.000 | 1 | 1.5 |
| 0.000 | 2 | |
| -0.005 | 3 | 3.5 |
| +0.005 | 4 | |
| +0.006 | 5 | 5 |
| +0.007 | 6 | 6 |
| -0.008 | 7 | 7 |
| +0.009 | 8 | 8 |
| +0.017 | 9 | 9 |
| +0.021 | 10 | 10 |
| +0.022 | 11 | 11 |
| +0.033 | 12 | 12 |
| +0.047 | 13 | 13 |
| +0.063 | 14 | 14 |

$$R^+ = 91.5 + 1.5 = 93 \qquad R^- = 10.5 + 1.5 = 12$$

$$T = 12 < 21$$

# Wilcoxon signed-ranks test

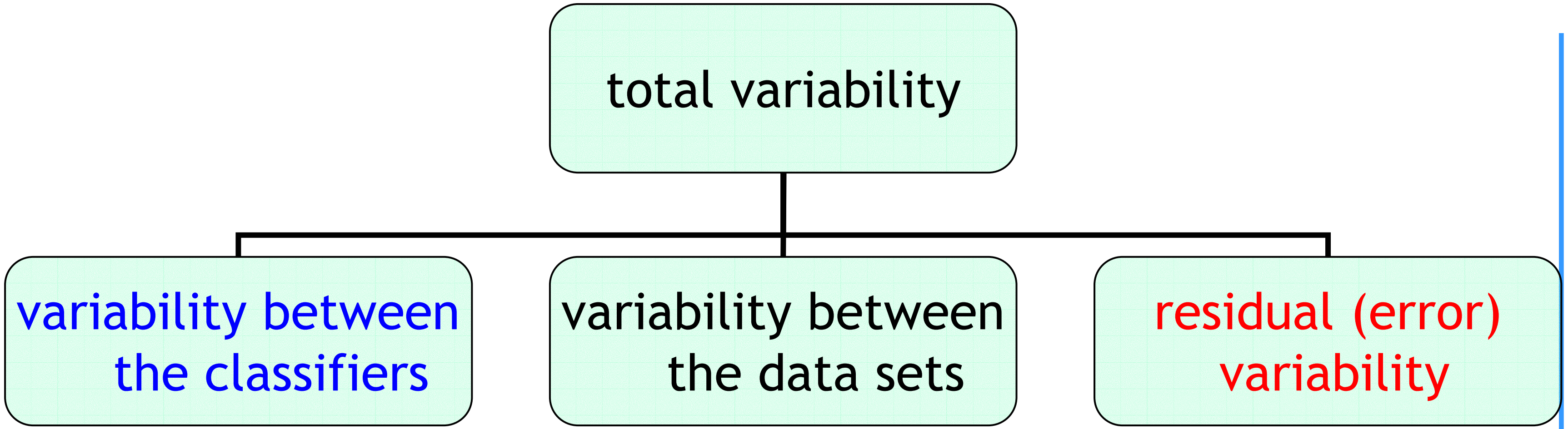
- more sensible than t-test
  - commensurability: only qualitatively
  - does not assume normal distributions: safer
  - Outliers: less effect
- less powerful or more powerful
  - assumptions of the paired t-test

# Comparisons of Multiple Classifiers

- well-known statistical problem
- control the *family-wise error*
  - probability of making at least one Type 1 error
- Statistics offers powerful specialized procedures
  - ANOVA
  - non-parametric counterpart: Friedman test

# ANOVA

- *repeated-measures ANOVA*(*within-subjects ANOVA*)
  - common statistical method
  - between more than two related sample means



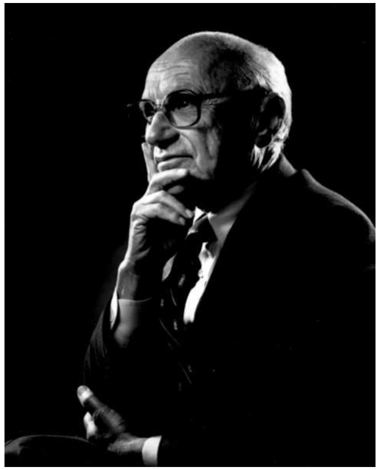
# probably violated assumptions

- normal distributions
  - minor problem
- sphericity
  - homogeneity of variance
  - requires random variables have equal variance
  - Violations of these assumptions have an even greater effect on the post-hoc tests

# Friedman test

- ranks algorithms for each data set eparately
- average ranks of algorithms $R_j = \frac{1}{N}\sum_i r_i^j$

| | 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** |



# Friedman test

- $$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$
  - according to $\chi_F^2$ with $k-1$ degrees of freedom
- $$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$
  - according to the F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom

# Friedman test

| | C4.5 | C4.5+m | C4.5+cf | C4.5+m+cf |
|---|---|---|---|---|
| … | … | … | … | … |
| average rank (Rj) | **3.143** | **2.000** | **2.893** | **1.964** |

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

$$= \frac{12 \cdot 14}{4 \cdot 5}\left[\left(3.143^2 + 2.000^2 + 2.893^2 + 1.964^2\right) - \frac{4 \cdot 5^2}{4}\right] = 9.28$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} = \frac{13 \cdot 9.28}{14.3 - 9.28} = 3.69$$

$$F\left((k-1),(k-1)\cdot(N-1)\right) = F(3,39) \xrightarrow{\alpha=0.05} 2.85$$

# post-hoc test

- Nemenyi test (Nemenyi,1963)
  - is used when all classifiers are compared to each other
  - $CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$

| #classifiers | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| $q_{0.05}$ | 1.960 | 2.343 | 2.569 | 2.728 | 2.850 | 2.949 | 3.031 | 3.102 | 3.164 |
| $q_{0.10}$ | 1.645 | 2.052 | 2.291 | 2.459 | 2.589 | 2.693 | 2.780 | 2.855 | 2.920 |

$\alpha$=0.05: CD=1.25

$\alpha$=0.10: CD=1.16

C4.5 - C4.5+m

C4.5 - C4.5+m+cf

| | |
|---|---|
| C4.5 | **3.143** |
| C4.5+m | **2.000** |
| C4.5+cf | **2.893** |
| C4.5+m+cf | **1.964** |

# post-hoc test

- Bonferroni correction
  - all classifiers are compared with a control classifier
  - more powerful than the Nemenyi test
  - $z = (R_i - R_j) \Big/ \sqrt{\dfrac{k(k+1)}{6N}}$
  - find the corresponding probability from the table of normal distribution
  - compared with an appropriate $\alpha$

# Conclusion

- Wilcoxon signed-ranks test & Friedman test
  - Appropriate
    - assume some, but limited commensurability
  - safer than parametric tests
    - do not assume normal distributions or homogeneity
  - stronger than the other tests studied

# Danke für Ihre Aufmerksamkeit