Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data

Pablo A. Jaskowiak*
Ricardo J. G. B. Campello

Computer Science Department
Institute of Mathematics and Computer Science – ICMC
University of São Paulo – São Carlos - Brazil

Outline

- Introduction
- Correlation Coefficients
- Experimental Setup
- Results
- Conclusions

Introduction

- Research focus is switching
  - From sequencing
  - To the understanding of how genomes are functioning
- Low level characterization of diseases
  - Cancer

Introduction

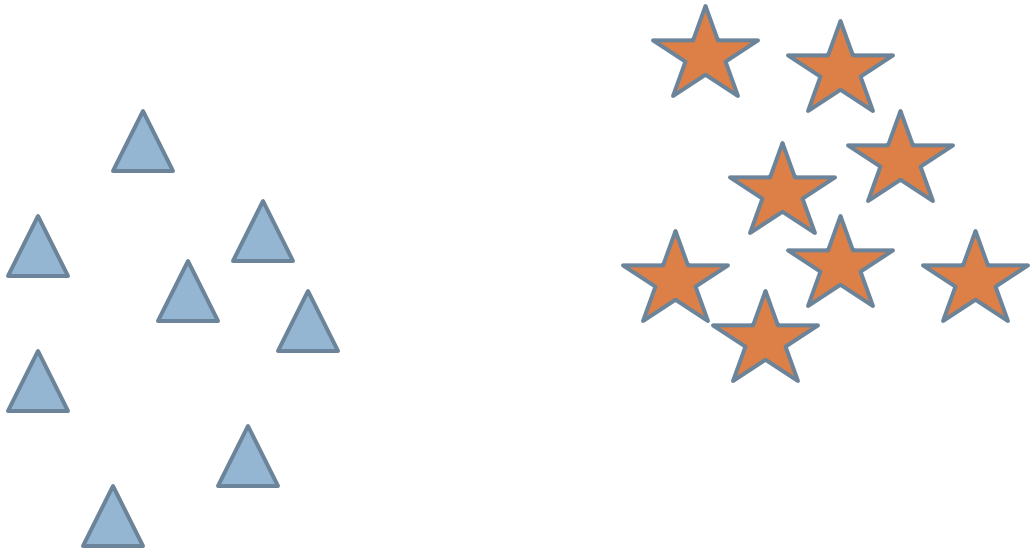
- Microarray technology
  - Expression level measurement for thousands of genes
  - Genomic picture of the system for a given state, e.g.,
    - A patient with cancer
    - A healthy patient
- Based on microarray data
  - Build classifiers or induce models
  - Predict the state of previously unseen samples
    - Cancer classification

Introduction

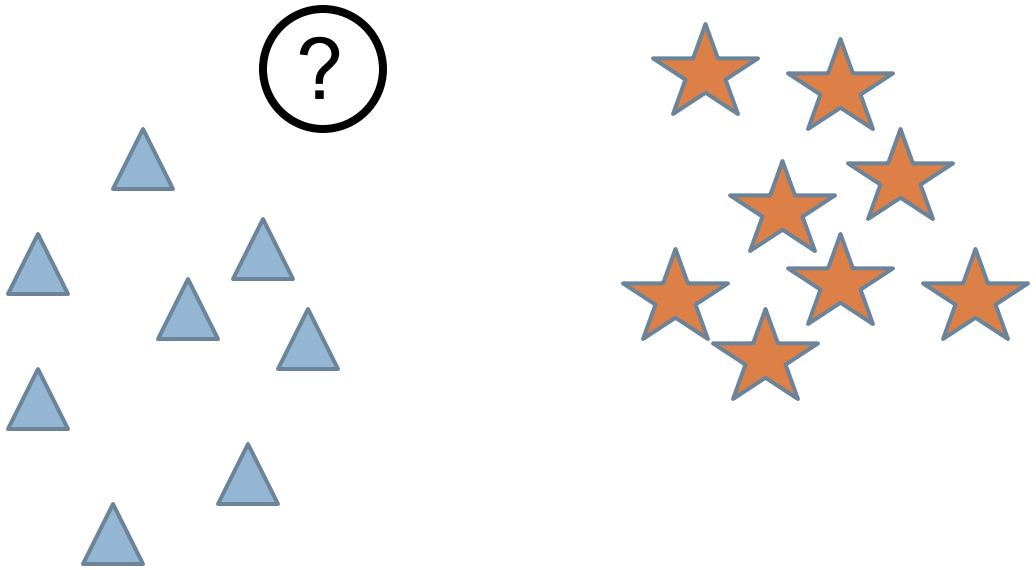
- Different classifiers have been adopted, e.g.,
  - SVMs, NN, LDA, *k*NN
- In this work we are particularly interested in
  - *k*-Nearest Neighbors classifier (*k*NN)
    - Simple
    - Has shown good results in cancer classification problems
    - Straightforward to implement
    - Even with more complex classifiers is still in use (Parry et al., 2010)

Parry et al., 2010. k-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. The Pharmacogenomics .

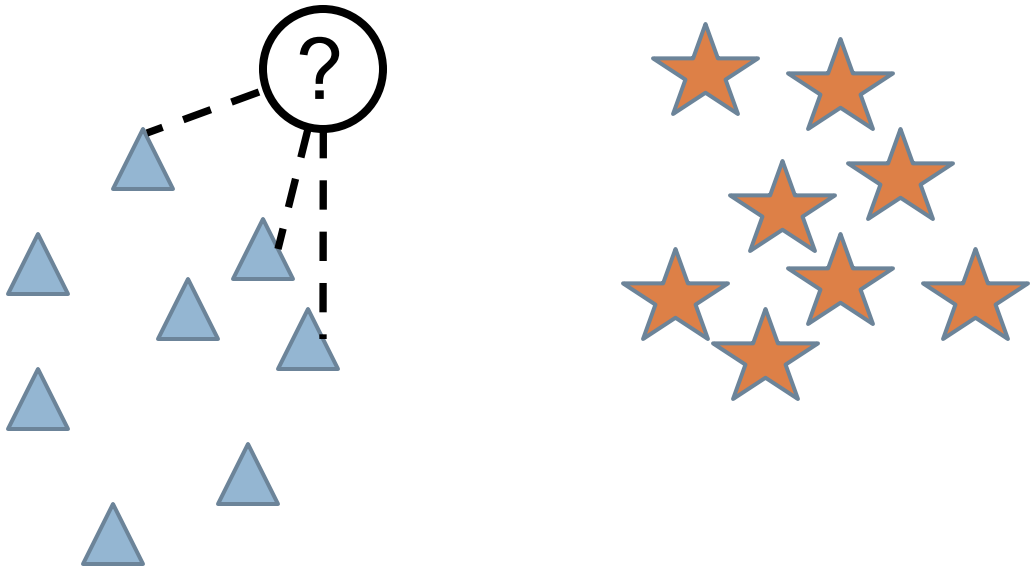
Introduction



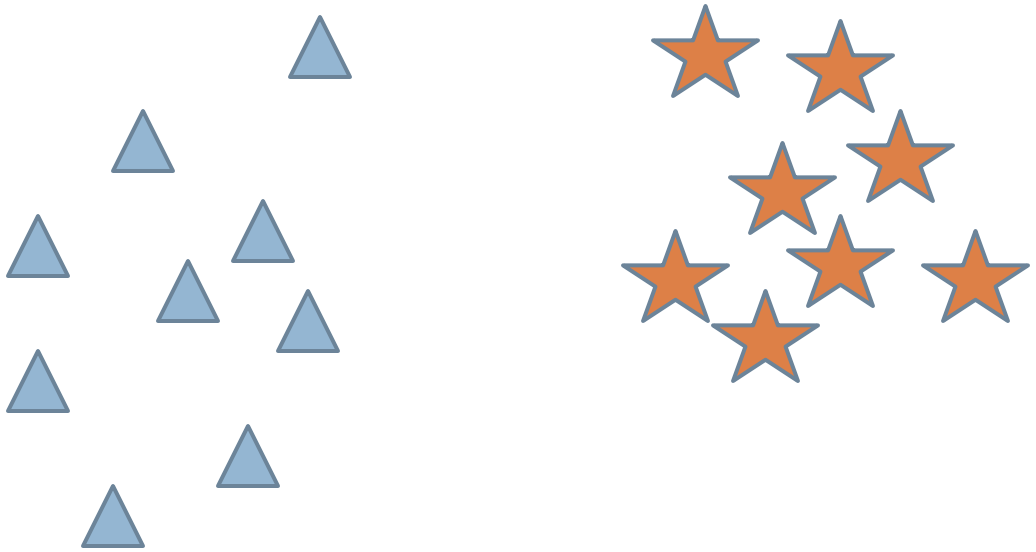
Introduction



Introduction



Introduction



Introduction

- What is close?
  - Different proximity measures (similarity or dissimilarity)
    - Different results
- Proximity is a key concept for *k*NN classifier
- Classifier depends on the definition of a proximity measure

Introduction

- Considering gene expression data
  - Trend similarity concept
    - Similarity in shape, rather than absolute differences
  - Proximity measures typically employed
    - Pearson correlation coefficient
    - Spearman correlation coefficient
    - Euclidean distance

Motivation

- *k*NN sensitive to proximity choice
- Other correlation coefficients available
- Parry et al., 2010 evaluated some proximity measures for *k*NN
  - No correlation coefficient though
- Comparison of different correlation coefficients
  - Measures sensitive to magnitudes of the values
  - Rank-based measures
  - Measures that are sensitive to both

Parry et al., 2010. k-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. The Pharmacogenomics .

Correlation Coefficients

Correlation Coefficient	Symbol	Sensibility	Time Complexity
Pearson	ρ	Magnitudes	O(n)
Jackknife	ϱ	Magnitudes	$O(n^2)$
Goodman-Kruskal	γ	Ranks	O(n log n)
Kendall	τ	Ranks	O(n log n)
Spearman	$\hat{\rho}$	Ranks	O(n log n)
Rank-Magnitude	r	Ranks and Magnitudes	O(n log n)
Weighted Goodman -Kruskal	$\hat{\gamma}$	Ranks and Magnitudes	$O(n^2)$

Experimental Setup

- 35 publicly available gene expression cancer datasets (Souto et al., 2008)
  - 14 double channel datasets (cDNA)
  - 21 single channel datasets (Affymetrix)

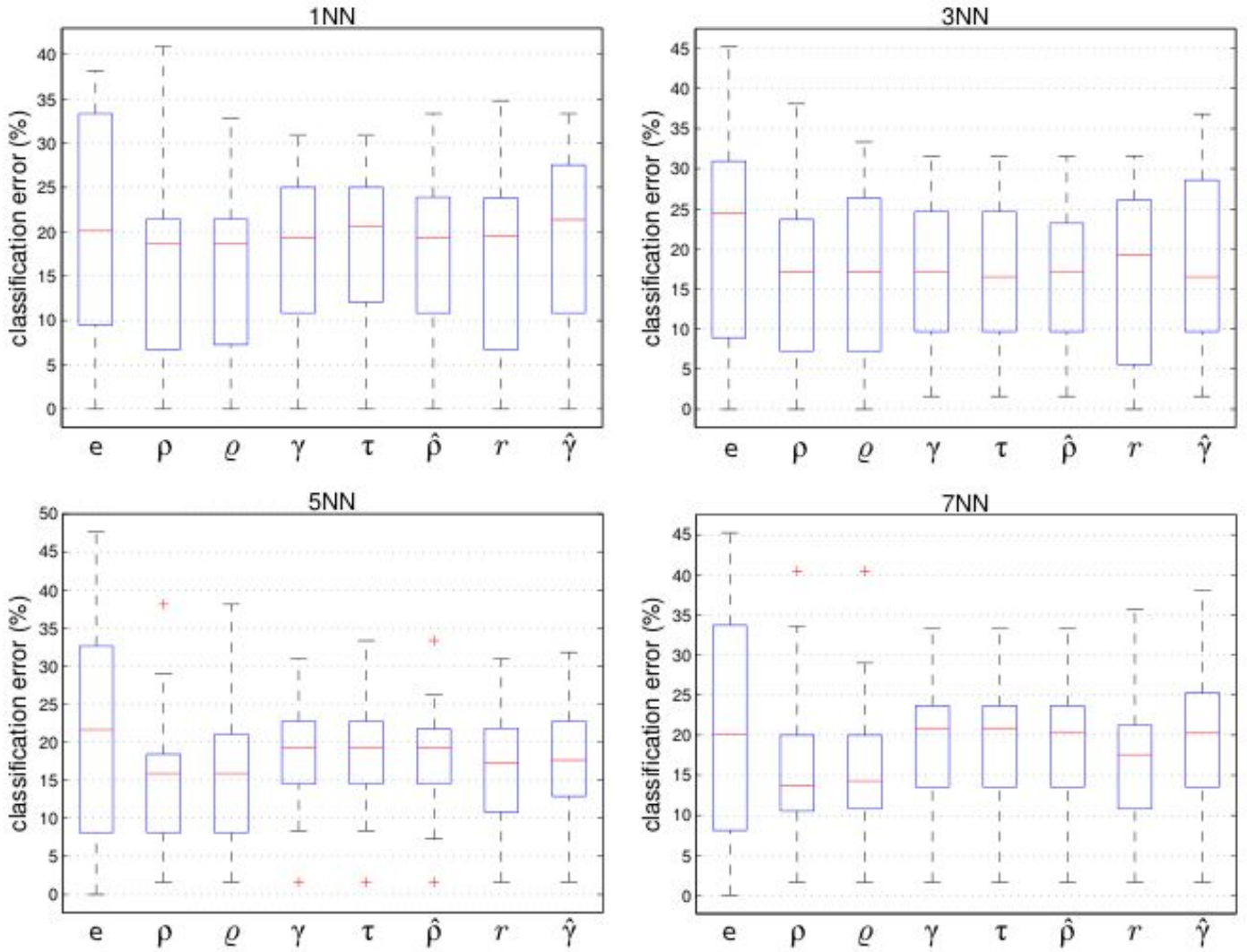
- Regarding the number of neighbors *k*, we used four values (Dudoit et al., 2000)
  - 1NN, 3NN, 5NN and 7NN

- Seven correlation coefficients + Euclidean distance

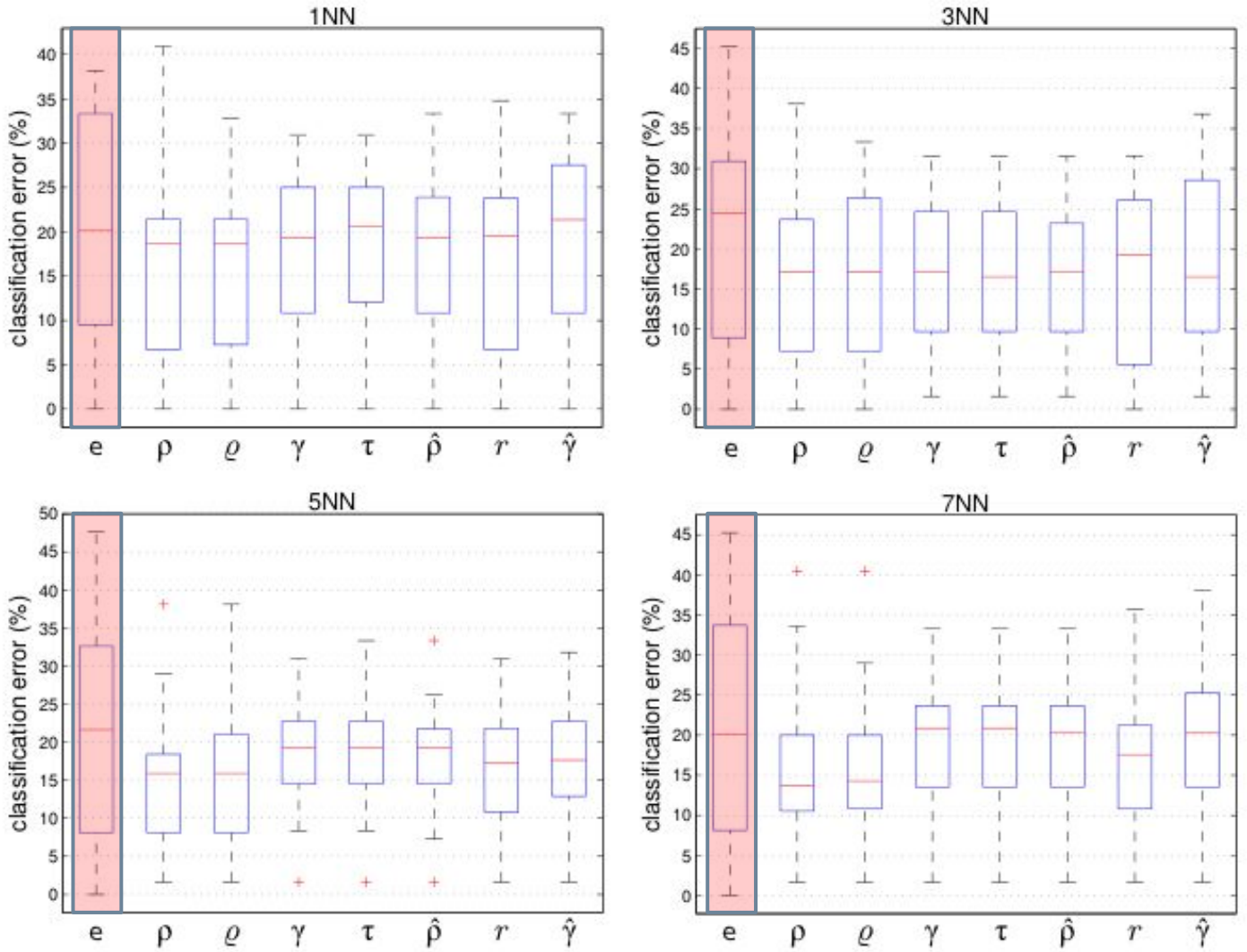
- Proximity measures evaluated by their LOOCV error

Souto et al., 2008. Clustering cancer gene expression data: a comparative study. BMC Bioinformatics.
Dudoit et al., 2000. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Stat. Assoc.

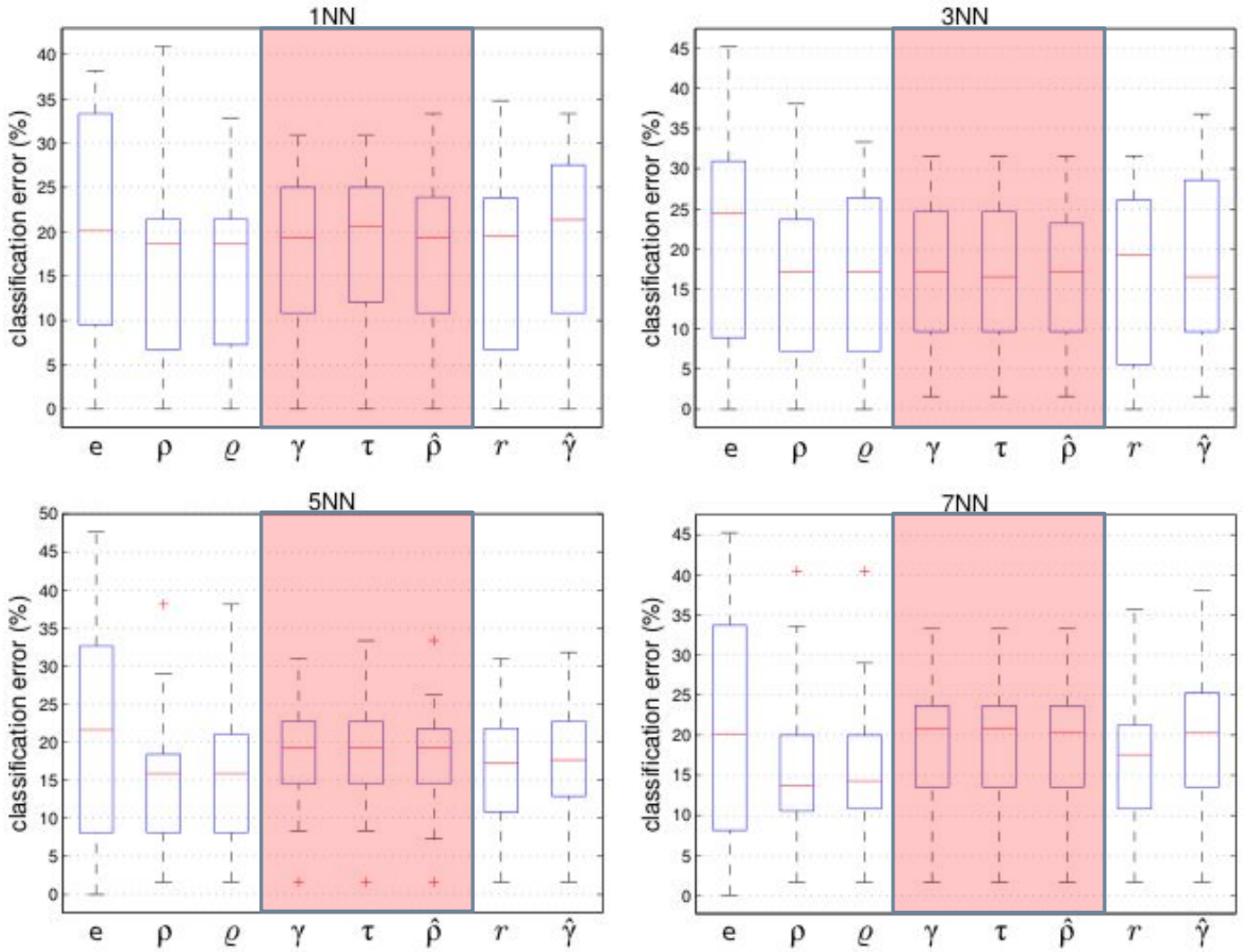
Results - cDNA



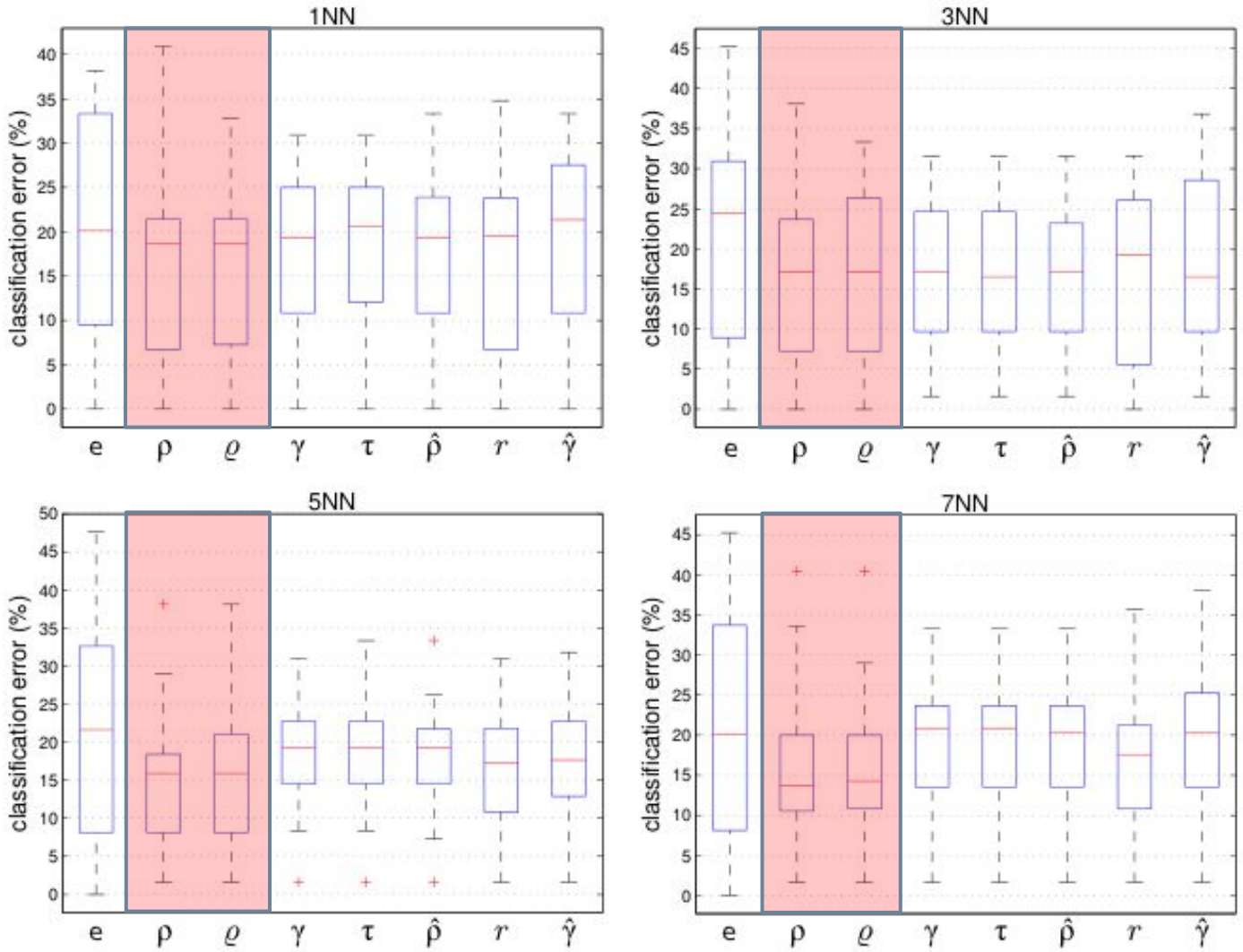
Results - cDNA



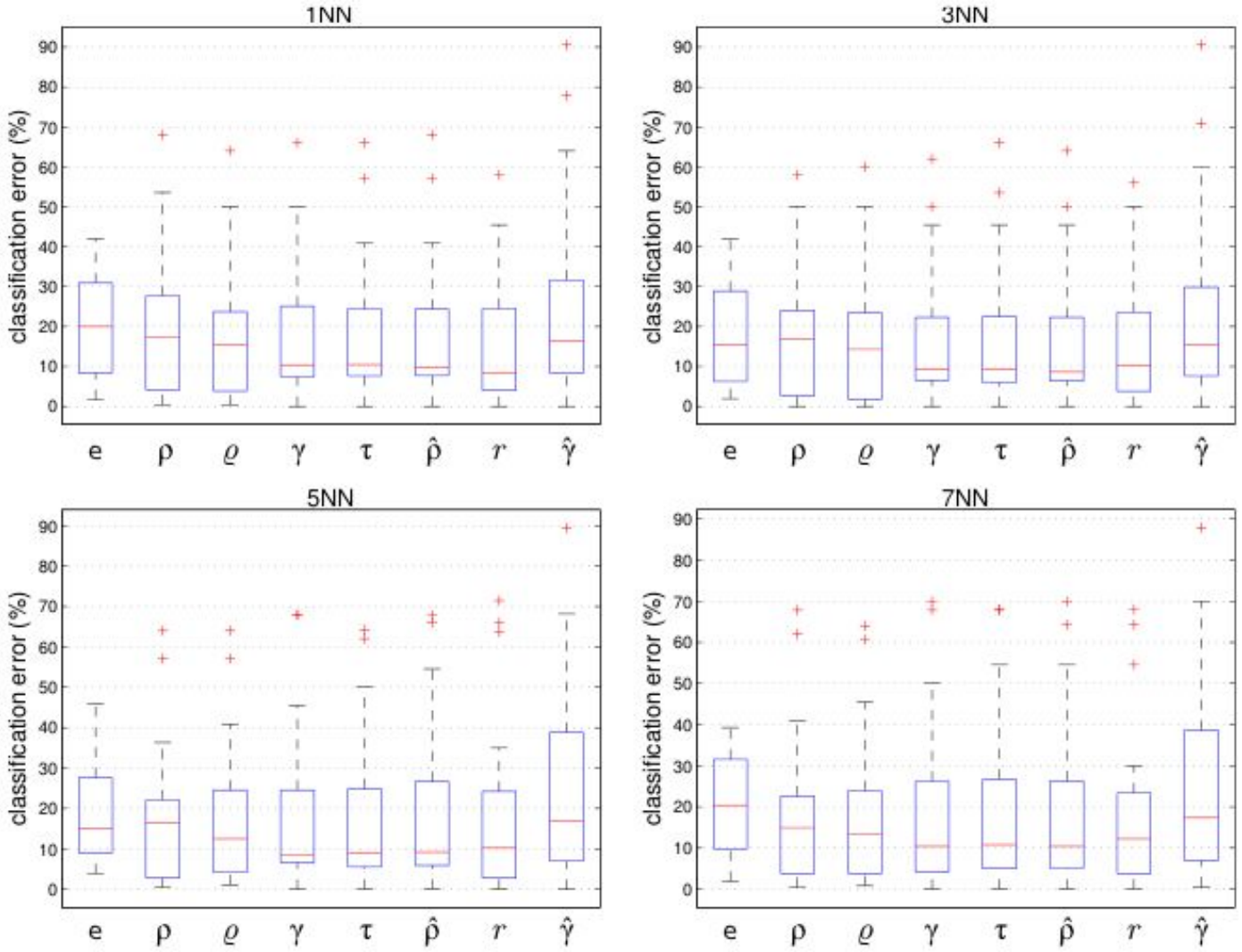
Results - cDNA



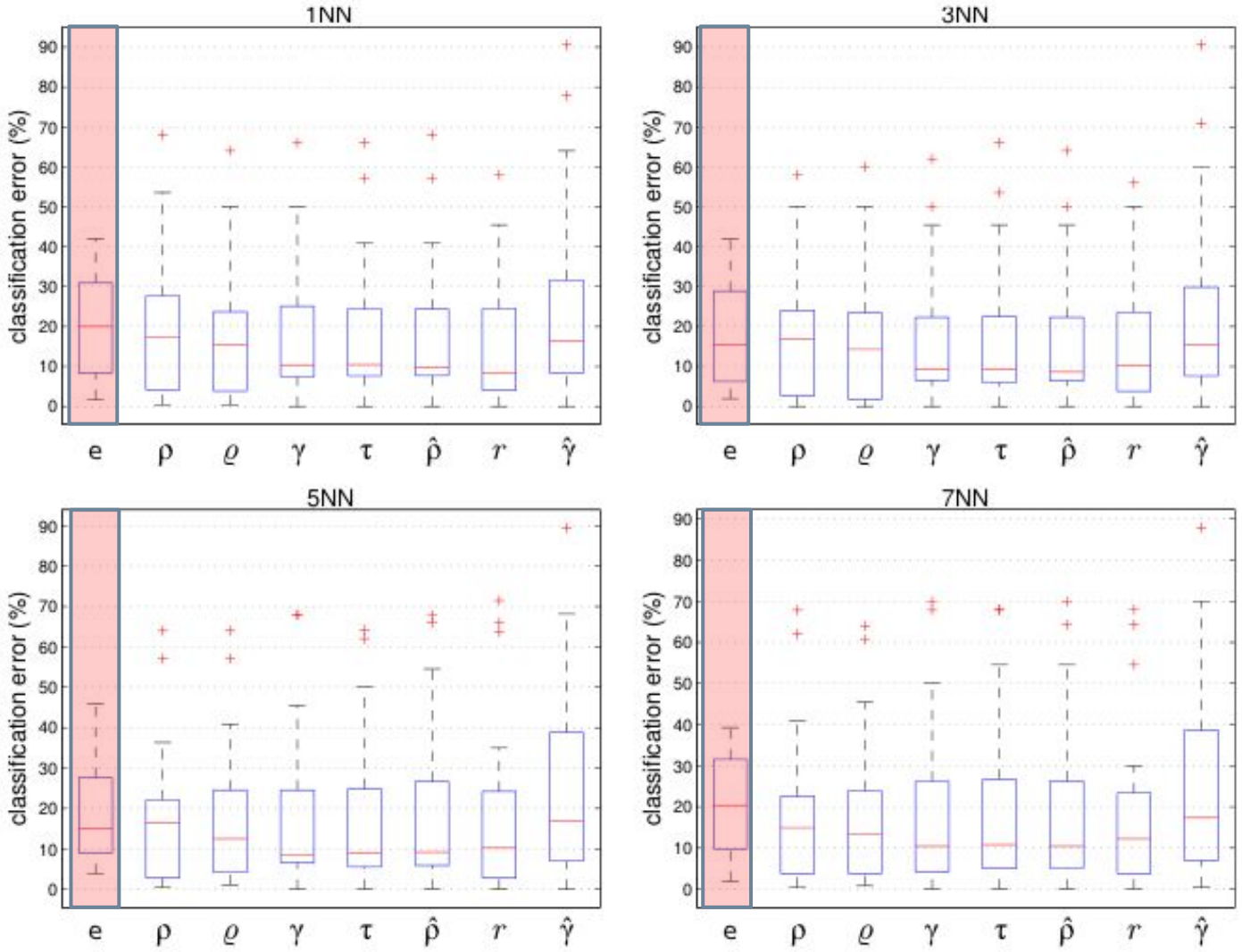
Results - cDNA



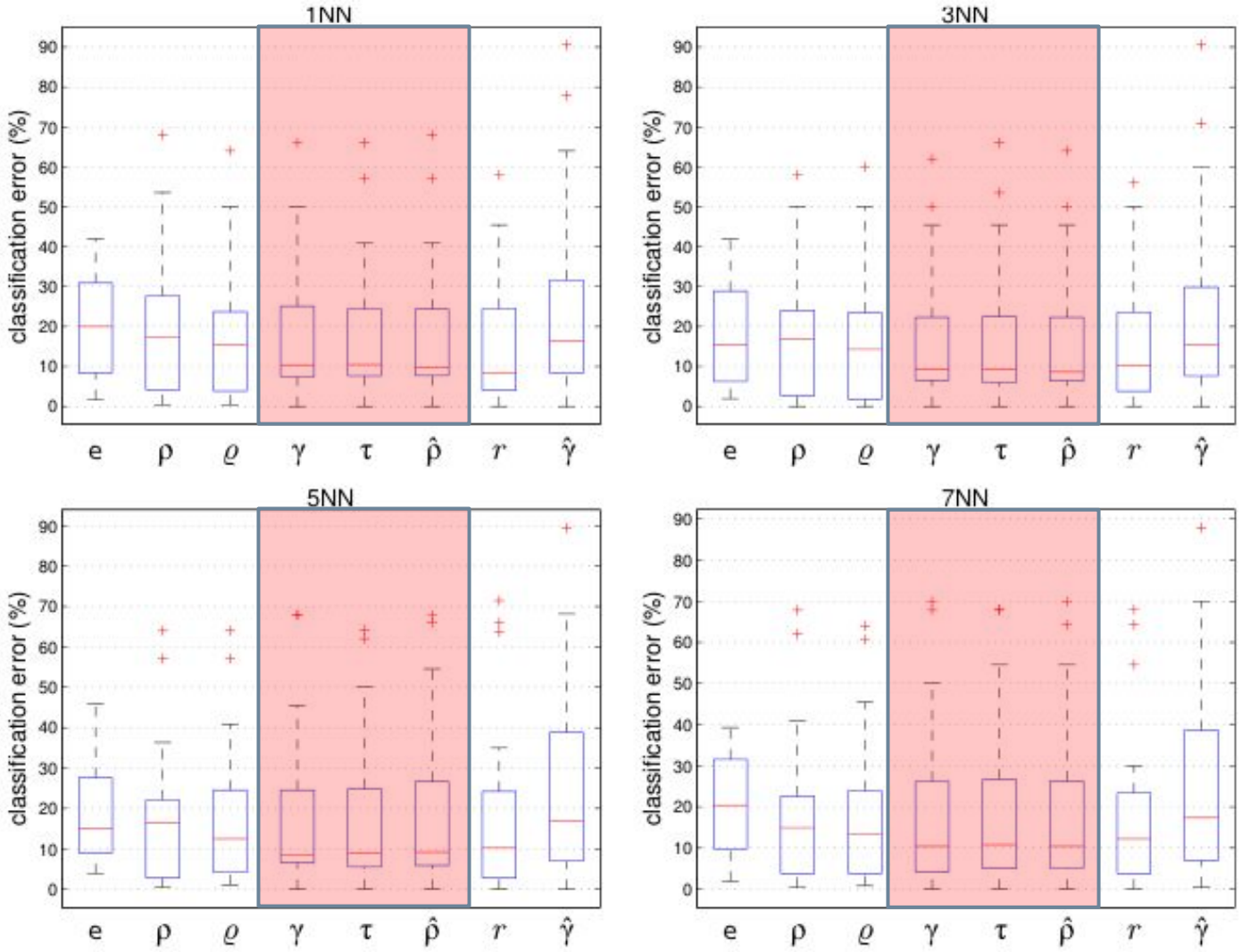
Results - Affymetrix



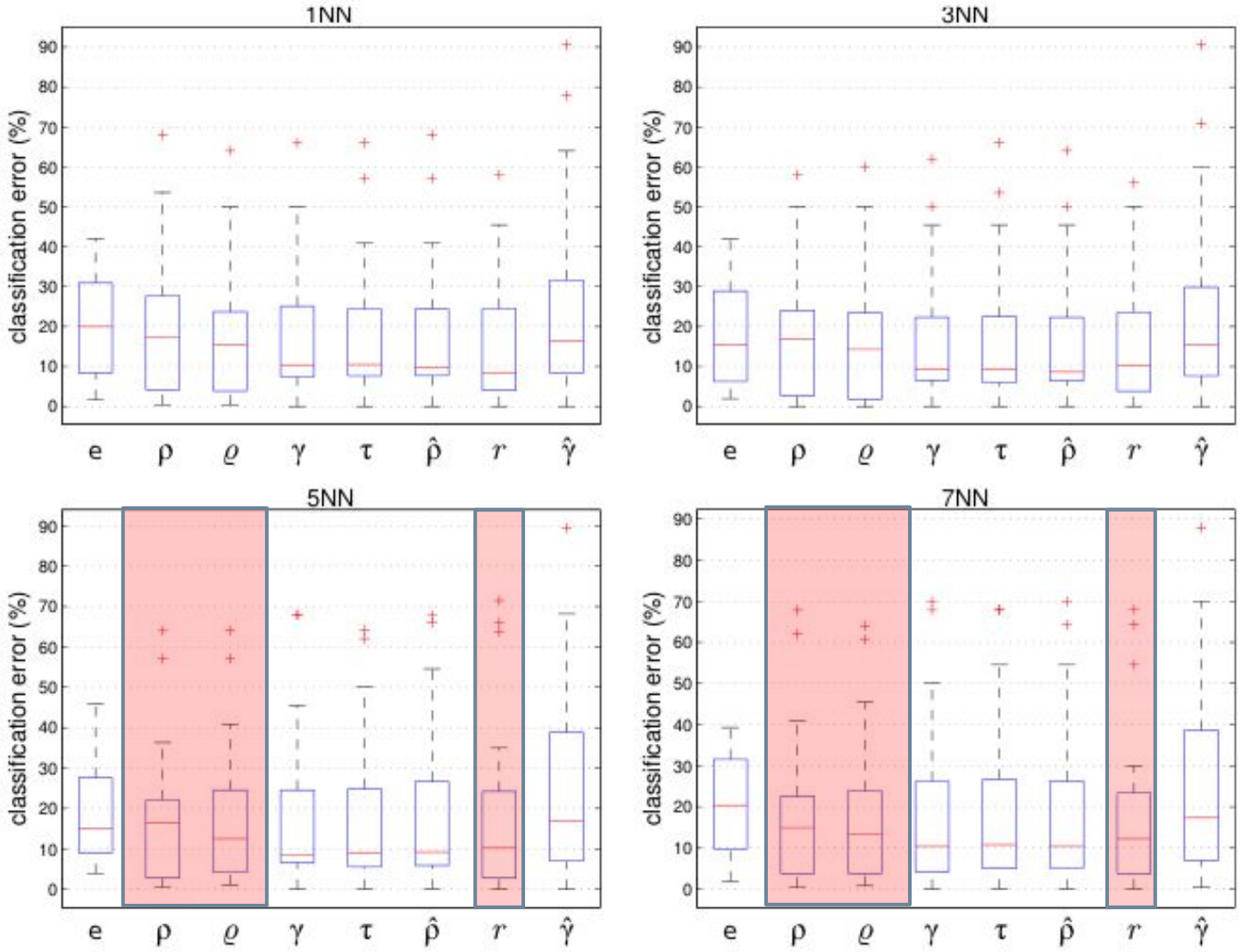
Results - Affymetrix



Results - Affymetrix



Results - Affymetrix



Results

- Great variability
  - Different datasets
- In particular datasets
  - Great differences among correlation coefficients

Results

- Considering 1NN results

Dataset	e	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
West-2001	30.6	16.3	14.3	6.1	8.2	8.2	8.2	16.3
Bitner-2000	34.2	13.2	13.2	18.4	21.1	18.4	15.8	29.0

Results

- Considering 1NN results

Dataset	e	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
West-2001	30.6	16.3	14.3	6.1	8.2	8.2	8.2	16.3
Bitner-2000	34.2	13.2	13.2	18.4	21.1	18.4	15.8	29.0

Results

- Considering 1NN results

Dataset	e	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
West-2001	30.6	16.3	14.3	6.1	8.2	8.2	8.2	16.3
Bitner-2000	34.2	13.2	13.2	18.4	21.1	18.4	15.8	29.0

Conclusions

- We compared eight different proximity measures for cancer classification regarding gene expression data
  - 7 correlation coefficients + Euclidean distance
- Considering different datasets
  - Large differences were found among correlation coefficients

Conclusions

- cDNA data
  - Euclidean distance is not a good alternative
  - Goodman-Kruskal and Kendall
    - Good alternatives to the commonly employed Spearman
- Affymetrix data
  - Rank-Magnitude appears as a promising alternative to
    - Pearson
    - Euclidean distance
  - Rank-based measures displayed similar results among themselves

Conclusions

- In real application scenarios
  - Exploratory analysis is seemingly the best choice
  - When there is no difference among measures
    - Employ the least computationally expensive one

- As future work
  - For particular datasets great differences were observed
  - Investigate possible relations between characteristics of the datasets and the results produced by the correlations

Acknowledgements

FAPESP

CAPES

CNPq

Questions?

Pablo Andretta Jaskowiak

pablo@icmc.usp.br

ICMC USP SÃO CARLOS