Evaluating Correlation Coefficients for Clustering Gene Expression Profiles of Cancer

Pablo A. Jaskowiak[1]*
Ricardo J. G. B. Campello[1]
Ivan G. Costa[2]

[1]Institute of Mathematics and Computer Sciences
University of São Paulo - São Carlos, Brazil

[2]Center of Informatics
Federal University of Pernambuco - Recife, Brazil

August 16, 2012

1. Introduction
2. Correlation Coefficients
3. Evaluating Proximity Measures
4. Results and Discussion
5. Concluding Remarks

Outline

1. Introduction
2. Correlation Coefficients
3. Evaluating Proximity Measures
4. Results and Discussion
5. Concluding Remarks

Microarrays

- Allow expression level measurement for thousands of genes
- Huge amounts of data, the so-called gene expression data
- Cluster is usually one of the first steps employed for its analysis
    - Clustering of genes — time-course gene expression data
    - Clustering of biological samples — related to cancer

Microarrays

- Allow expression level measurement for thousands of genes
- Huge amounts of data, the so-called gene expression data
- Cluster is usually one of the first steps employed for its analysis
  - Clustering of genes — time-course gene expression data
  - **Clustering of biological samples — related to cancer**

Clustering of Cancer Samples

- Problem gained attention with the work of Golub et al.[1]
- Characterized by
  - Small number of samples (objects)
  - Large number of genes (features)
- To cope with this particular application scenario
  - Different clustering methods has been employed and developed

[1] T. R. Golub et al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring". In: *Science* (1999).

Clustering of Cancer Samples

- Problem gained attention with the work of Golub et al.[1]
- Characterized by
  - Small number of samples (objects)
  - Large number of genes (features)
- To cope with this particular application scenario
  - Different clustering methods has been employed and developed
- Studies provided guidelines for selecting clustering methods

[1] T. R. Golub et al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring". In: *Science* (1999).

Clustering of Cancer Samples

- Results are not determined solely by the clustering method
- Selecting appropriate proximity measures is an important issue
  - Selection usually depends on the application scenario
- Regarding the clustering of cancer samples
  - One usually seeks for trend (shape) similarity
  - Pearson correlation has been widely employed
  - Other correlation measures are available as alternatives

Proximity Measures for Clustering Cancer Samples

- Previous works that consider different proximity measures
    - Primarily interested in the comparison of clustering algorithms
    - Considered a small number of datasets *without* making any distinction between the clustering of genes and cancer samples

Contributions

- Comparison of proximity measures for clustering cancer samples
  - Five correlation coefficients
- Measures are compared regarding
  - Intrinsic separation ability
  - Predictive clustering ability

Outline

1 Introduction

2 Correlation Coefficients

3 Evaluating Proximity Measures

4 Results and Discussion

5 Concluding Remarks

Proximity Measures in Gene Expression

- Should capture shape or trend similarity
- Correlation coefficients capture such kind of similarity
- Different measures in the literature, we consider

Correlation	Sensibility	Time Complexity
Pearson	Magnitudes	$O(n)$
Jackknife	Magnitudes	$O(n^2)$
Spearman	Ranks	$O(n \log n)$
Kendall	Ranks	$O(n \log n)$
Rank-Magnitude	Both	$O(n \log n)$

Outline

1. Introduction
2. Correlation Coefficients
3. Evaluating Proximity Measures
4. Results and Discussion
5. Concluding Remarks

Evaluating Proximity Measures

- We evaluate proximity measures regarding their
  1. Intrinsic Separation Ability
  2. Predictive Clustering Ability

Intrinsic Separation Ability

- Concept recently introduced by[2]
- Evaluate proximity measures *without* a clustering algorithm
- The capacity that the proximity measure has to separate data points (objects) *without* the influence of a clustering algorithm

[2] R Giancarlo et al. "Distance Functions, Clustering Algorithms and Microarray Data Analysis". In: *Learning and Intelligent Optimization*. Springer, 2010.

Intrinsic Separation Ability

- Concept recently introduced by[2]
- Evaluate proximity measures *without* a clustering algorithm
- The capacity that the proximity measure has to separate data points (objects) *without* the influence of a clustering algorithm

Procedure

1. Given a *labeled* dataset with m samples (objects) $\mathbf{x_1}, \ldots, \mathbf{x_m}$
2. We start by obtaining a distance matrix D, where

$$D(i,j) = \text{distance}(\mathbf{x_i}, \mathbf{x_j}), with 1 \leq i,j \leq m, \text{ with}$$

$$0 \leq \text{distance}(\mathbf{x_i}, \mathbf{x_j}) \leq 1, \ \forall i,j$$

[2] R Giancarlo et al. "Distance Functions, Clustering Algorithms and Microarray Data Analysis". In: *Learning and Intelligent Optimization*. Springer, 2010.

Intrinsic Separation Ability

Procedure

3. We build a binary classifier that assigns data points according to Eq. (1), where $\phi \in [0, 1]$ is a given threshold.

$$I_\phi(\mathbf{x_i}, \mathbf{x_j}) = \begin{cases} 1 & \text{if } D(i,j) \leq \phi \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

4. As we are dealing with labeled data, a desired solution for the classifier in Eq. (1). The desired solution is built upon class labels, as given by Eq. (2) for all $\mathbf{x_i}$ and $\mathbf{x_j}$.

$$J(\mathbf{x_i}, \mathbf{x_j}) = \begin{cases} 1 & \text{if } \mathbf{x_i} \text{ and } \mathbf{x_j} \text{ belong} \\ & \text{to the same cluster} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Intrinsic Separation Ability

Procedure

5. By considering values of ϕ in the interval $[0, 1]$ in Eq. (1) we obtain a set of *predicted solutions* based on the distance itself
6. This set of *predicted solutions* is then evaluated against the *desired solution*, Eq. (2), which was built upon class labels
7. In brief, one has multiple comparisons (one per each value of ϕ) to perform. These comparisons are addressed by Receiver Operating Characteristics analysis (ROC analysis)
8. With ROC analysis we then obtain a value of Area Under the Curve (AUC) for the distance in question. Such AUC value quantifies the Intrinsic Separation Ability of the distance.

Predictive Clustering Ability

- To evaluate the Predictive Clustering Ability of the proximity measures we consider 4 different clustering algorithms, namely
  - k-medoids (KM)
  - Single-Linkage (SL)
  - Average-Linkage (AL)
  - Complete-Linkage (CL)
- We generate partitions with the *same* number of clusters, as defined by the reference partition of each dataset (class labels)
- Each proximity measure is evaluated by the Adjusted Rand values obtained when it is employed with each clustering algorithm

Outline

1. Introduction
2. Correlation Coefficients
3. Evaluating Proximity Measures
4. **Results and Discussion**
5. Concluding Remarks

Experimental Setup

- We consider 35 labeled benchmark datasets proposed by[3]
  - 21 Affymetrix datasets
  - 14 cDNA datasets
- Five correlations along with Euclidean distance (EUC)
- Evaluation performed separately for
  - Intrinsic Separation Ability
  - Predictive Clustering Ability

[3] M Souto et al. "Clustering Cancer Gene Expression Data: A Comparative Study". In: *BMC Bioinformatics* (2008).

Intrinsic Separation Ability

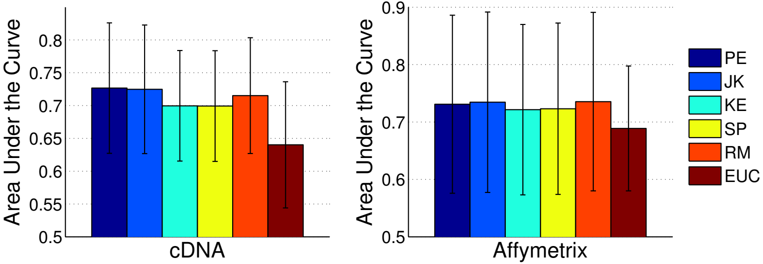

Figure : Intrinsic separation ability regarding correlation coefficients. Bars display mean results whereas error bars account for standard deviations.

Intrinsic Separation Ability

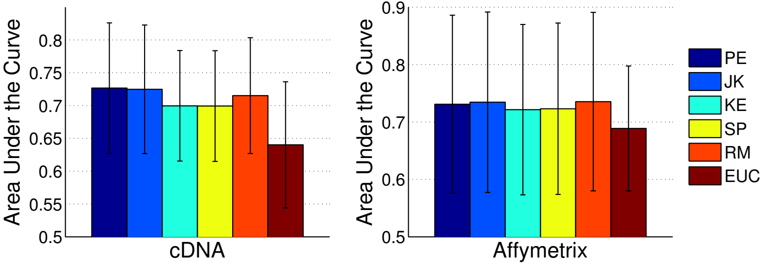

Figure : Intrinsic separation ability regarding correlation coefficients. Bars display mean results whereas error bars account for standard deviations.

- Statistical Tests: Friedman and Nemenyi (95% confidence level)
  - Affymetrix data: RM > EUC
  - cDNA data: PE, JK, and RM > EUC

Predictive Clustering Ability

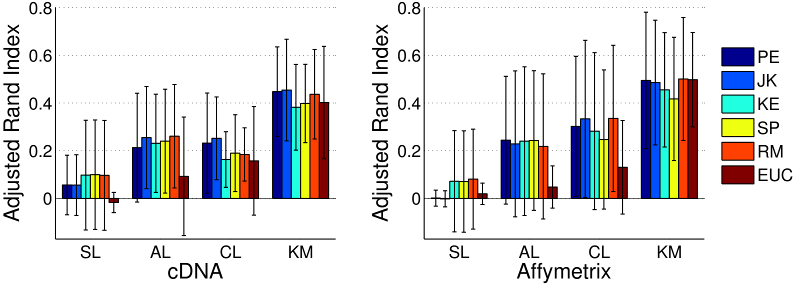

Figure : Class recovery regarding different correlation coefficients . Bars display mean results whereas error bars account for standard deviations.

Predictive Clustering Ability

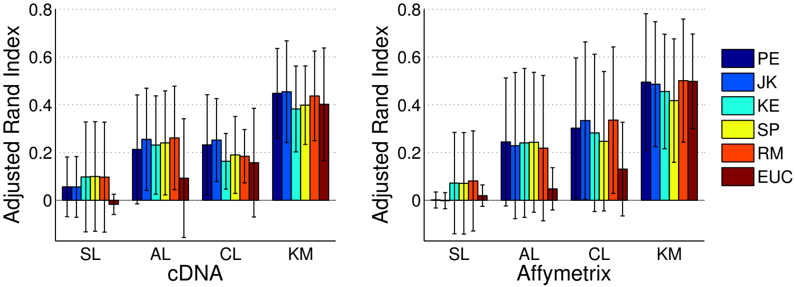

Figure : Class recovery regarding different correlation coefficients . Bars display mean results whereas error bars account for standard deviations.

- Friedman test suggests differences (95% confidence level)
- Nemenyi test was unable to identify among which pairs

Relating Intrinsic Separation and Predictive Ability

- In order to show relations between Intrinsic Separation Ability and Predictive Clustering Ability we correlate their results
- Mean Spearman correlation of
  - .82 when correlating with KM results
  - .74 when correlating with CL results
  - Poor correlation with SL results

Discussion

- All correlation coefficients provided better results than EUC, except for KM. Note, however, that regarding KM, PE, JK and RM also provided competitive or better results than EUC.
- PE and JK provided better results than other correlations.
- RM provides competitive results to PE and JK. Whereas PE and JK are based solely on the magnitude values of the sequences, RM considers also their ranks. RM may also be more robust to noise than PE and JK and arises as a good alternative to both.
- Rank-based correlations showed worst results than other correlation coefficients. This behavior can be explained by the loss of information inherent in the definition of such measures.

Outline

1 Introduction

2 Correlation Coefficients

3 Evaluating Proximity Measures

4 Results and Discussion

5 Concluding Remarks

Concluding Remarks

- We compared five correlation coefficients (along with EUC) for clustering cancer samples from microarray data, regarding
  - Intrinsic Separation Ability
  - Predictive Clustering Ability
- Among the measures, PE, JK, and RM provided good results
- JK has quadratic time complexity, for clustering cancer samples it may be an issue (great number of features per object)
- RM has moderate time-complexity and *may* be more robust to noise than PE. We do not have much information about it

Concluding Remarks

- As future work we intend to
  - Consider the scenario of gene clustering (time-course data)
  - Compare proximity measures that were specifically developed to the gene clustering scenario, specifically short time-course data
    - L J Heyer et al. "Exploring Expression Data: Identification and Analysis of Coexpressed Genes". In: *Genome Research* (1999)
    - R Balasubramaniyan et al. "Clustering of gene expression data using a local shape-based similarity measure". In: *Bioinf.* (2005)
    - C S Möller-Levet et al. "Clustering of unevenly sampled gene expression time-series data". In: *Fuzzy Sets and Systems* (2005)
    - Young Sook Son and Jangsun Baek. "A modified correlation coefficient based similarity measure for clustering time-course gene expression data". In: *Pattern Recognition Letters* (2008)
- Finally, we pretend to evaluate the behavior of the proximity measures in the presence of different levels of noise, assessing their effect in the performance of the measures.

Acknowledgements

Thank You!

Brazilian research agencies:
CAPES, CNPq, FACEPE and FAPESP.

Questions?

`pablo@icmc.usp.br`