On the evaluation of clustering results: measures, ensembles, and gene expression data analysis

Pablo Andretta Jaskowiak

Ricardo J. G. B. Campello (Advisor)
Ivan G. Costa (Co-Advisor)

November 27, 2015

Departamento de Ciências de Computação
Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade de São Paulo - São Carlos - Brasil

ICMC USP
SÃO CARLOS

Outline

1. Introduction, clustering, gene expression data
2. Relative validation of clustering results
3. Ensembles of relative validity criteria
4. Distances for clustering gene expression data
5. Biological validation of gene clustering results
6. Conclusions, contributions and future work

1 Introduction

Cluster analysis, clustering validation

Gene expression data

Motivation and lines of investigation

Introduction

- Increasing data collection and storage

- More than ever we need to make sense of data

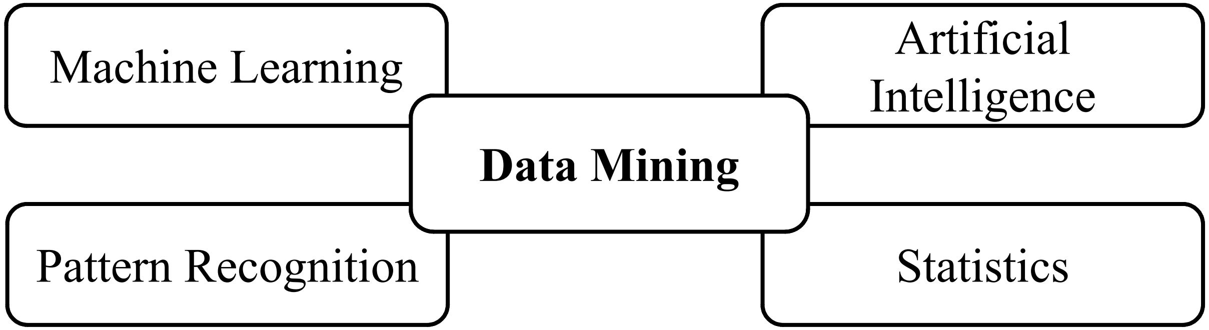


Figure adapted from Tan et al. 2006.

Cluster analysis

- Unsupervised Data Mining task
  - *Usually* there is no prior knowledge

Organize data objects into a finite set of categories (clusters), in the hope that meaningful relationships among objects will emerge from the process.

- What are clusters? How de we define them?
  - Well...

Cluster analysis

- Different clustering paradigms
  - Algorithms with different biases

- Most clustering algorithms *always* produce a result
  - Even when there are no “true” clusters...

- If we assume that there are clusters in the data
  - How many clusters?
  - Which clustering is the “best” one?

Clustering validation

- Quantitative evaluation of clustering solutions

- Three main categories (Jain and Dubes, 1988)
  - External
    - Quantify the agreement between two partitions

  - Internal
    - Quantify how well the actual partition fits the data

  - Relative
    - Internal measures that can point out the best partition from a pool

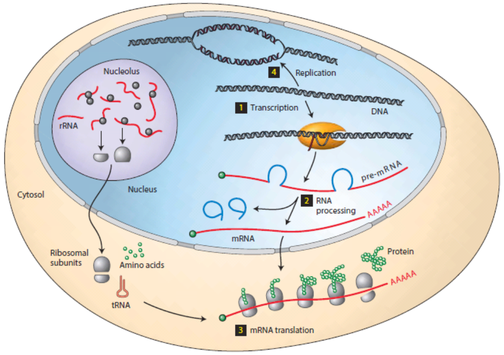
Clustering validation

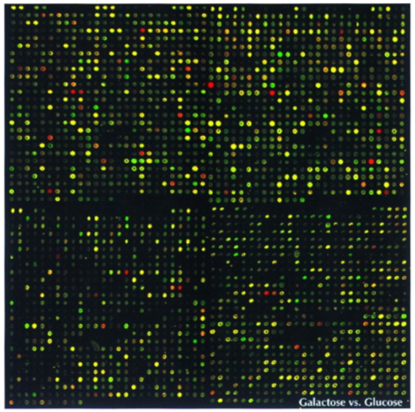
> *"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."*
>
> *Jain and Dubes, 1988*

- In a general context
  - Proposal of new relative validity measures
  - Ensembles of relative validity criteria
    - Evaluation of ad-hoc selection of members
    - Proposal of an heuristic selection of members

Gene expression data

- Understand cells and their undergoing processes





	Sample 1	Sample 2	...	Sample s
Gene 1	$e_{1,1}$	$e_{1,2}$	$\dots$	$e_{1,s}$
Gene 2	$e_{2,1}$	$e_{2,2}$	$\dots$	$e_{2,s}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Gene g	$e_{g,1}$	$e_{g,2}$	$\dots$	$e_{g,s}$

Clustering gene expression data

- Application domain with peculiarities
  - Clustering of short gene time-series
    - Large #Objects *vs* Small #Features
    - No labels for controlled experiments
    - External information
      - Gene Ontology – GO (Ashburner et al., 2000)
  - Clustering of samples
    - Small #Objects *vs* Large #Features

Clustering gene expression data

- Evaluation of distance measures
  - For different technologies
    - Microarrays and RNA-Seq
  - Using data itself and biological information
    - Proposal of new methodology
- Evaluation of gene clustering results
  - Employing data itseld and biological information

Relative validation of clustering results

Area Under the Curve (AUC)

Density-based Clustering Validation (DBCV)

Area Under the Curve (AUC)

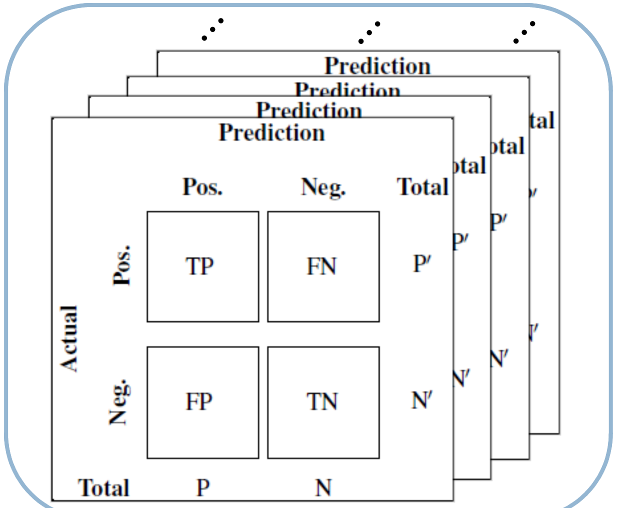
- Receiver Operating Characteristics (ROC)
  - Employed and studied in the supervised context

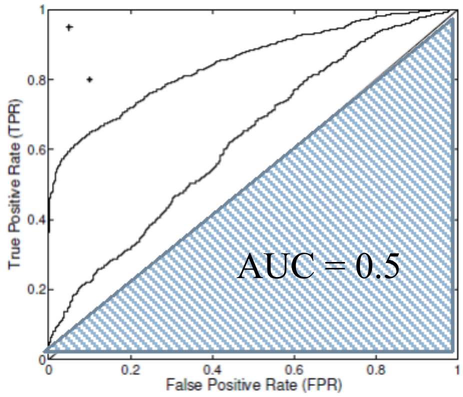
Predicted Output (Classifier)

0.9	0.6	0.8	0.7	...	...	...	...	0.8	0.2

Actual Output (True Memberships)

1	1	0	1	...	...	...	...	1	0





Area Under the Curve (AUC)

- It hasn´t been explored in the unsupervised setting

Hyphothesis 1:

The Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve can be effectivelly employed in the validation of clustering results as a relative validity criterion.

Area Under the Curve (AUC)

- How can we employ AUC in clustering validation?
  - As usual, we have a partition and pairwise distances

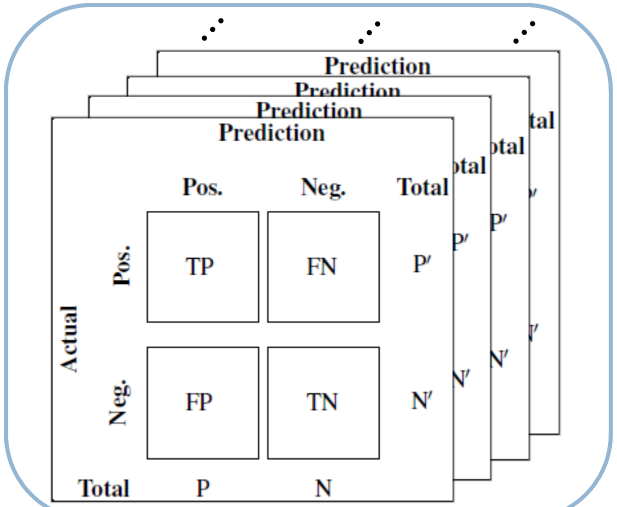
Pairwise distances (normalized)

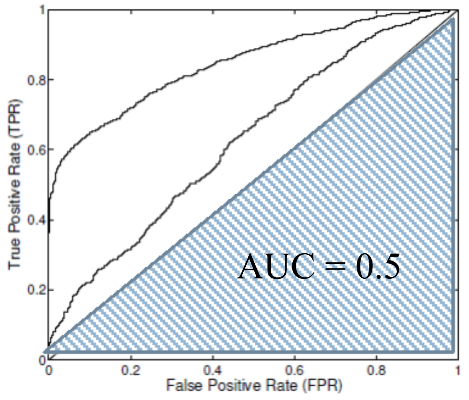
0.9	0.6	0.8	0.7	...	...	...	...	0.8	0.2

Pairwise memberships w.r.t. clusters

1	1	0	1	...	...	...	...	1	0

$\frac{n(n-1)}{2}$ pairs of objects!





Area Under the Curve (AUC)

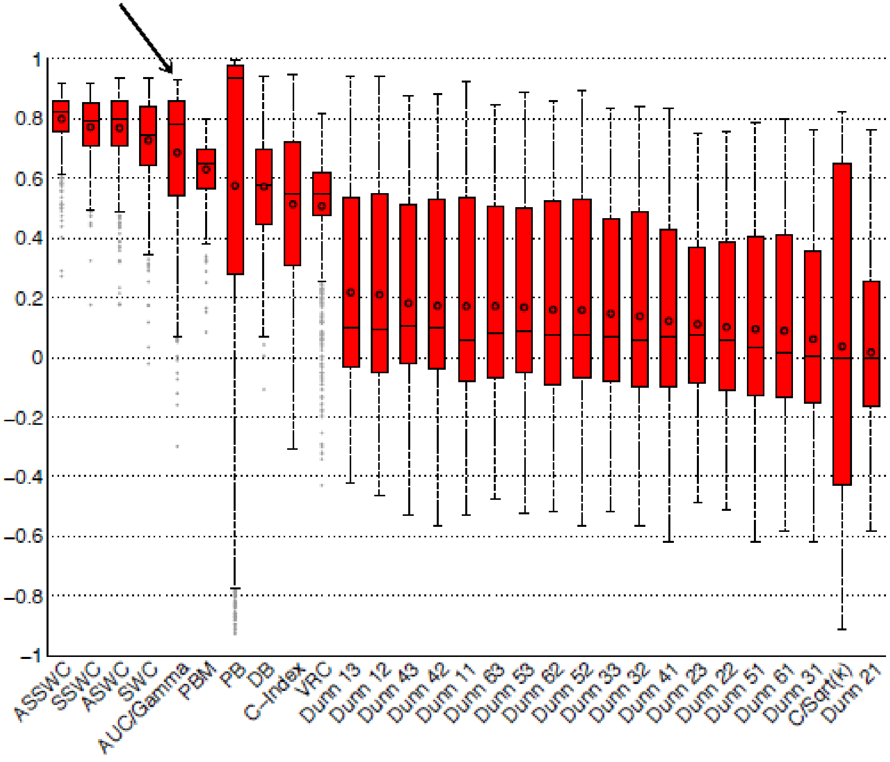
- Some of AUC properties in the context of clustering
  - Still has an expected value of 0.5 (independent of *k*)
  - Equivalent to Gamma (Baker and Hubert, 1975)
    - $AUC = \left(\frac{Gamma+1}{2}\right)$
  - Lower computational complexity than Gamma
    - AUC: $O(n^2 \log n)$
    - Gamma: $O\left(n^2 m + \frac{n^4}{k}\right)$

Area Under the Curve (AUC)

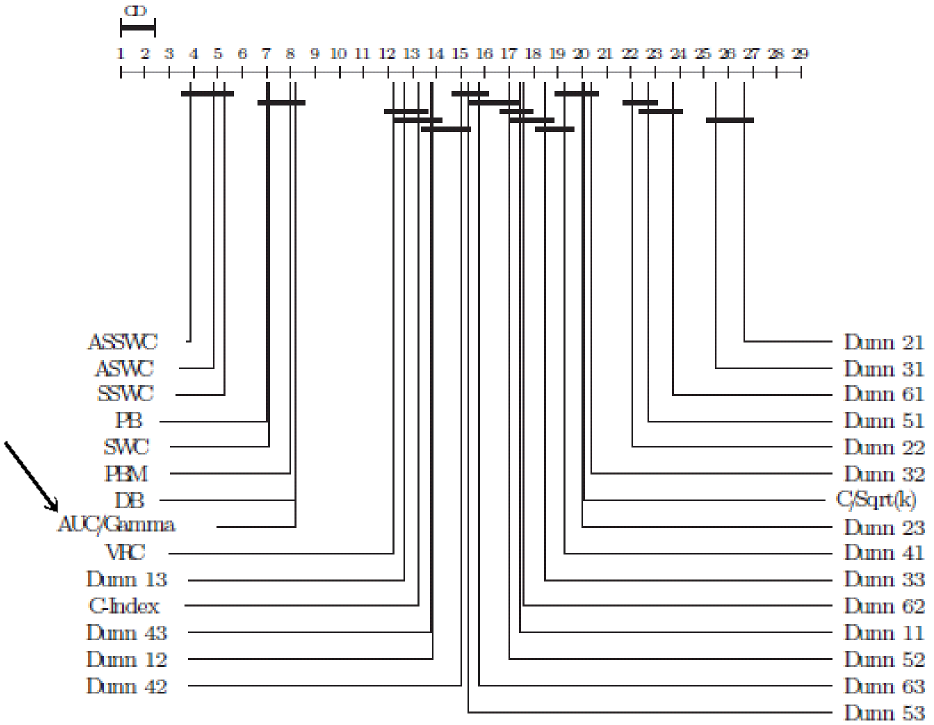
- How well does it work?
  - Replicated the expriments from *Vendramin et al., 2010*
- Datasets
  - 972 Synthetic datasets from Vendramin et al. 2010
- Partitions
  - HCA´s and k-means with $k \in [2, \lceil\sqrt{n}\rceil]$
- Criteria evaluated w.r.t. their correlation with external measure

Area Under the Curve (AUC)

- How well does it work?



(a) Results for Pearson



(b) Statistical Test Summary for Pearson

Area Under the Curve (AUC)

- Good results in comparison to other measures
- Similar to Gamma, but with lower cost
  - Appealing to relational clustering
- We believe that the initial hypothesis is valid

Relative Validation of Clustering Results

Area Under the Curve (AUC)

Density-based Clustering Validation (DBCV)

Density-based Clustering Validation (DBCV)

- Developed during author´s internship at U of A
  - Jointly supervised by Prof. Dr. Jörg Sander
  - Work done in collaboration (D. Moulavi - main author)

- Validation of arbitrary shaped clusters and noise
  - Few works on the topic to the date
    - Do not take denstities into account
    - Measures rely on parameters

Density-based Clustering Validation (DBCV)

- Based on the definition of a new core distance
  - Quantifies the density of each object w.r.t. its cluster
  - Mutual Reachability Distances (MRD)

- Each cluster is represented by a MST
  - Built on the basis of Mutual Reachability distances
  - Capture the shape and densities of each cluster

Density-based Clustering Validation (DBCV)

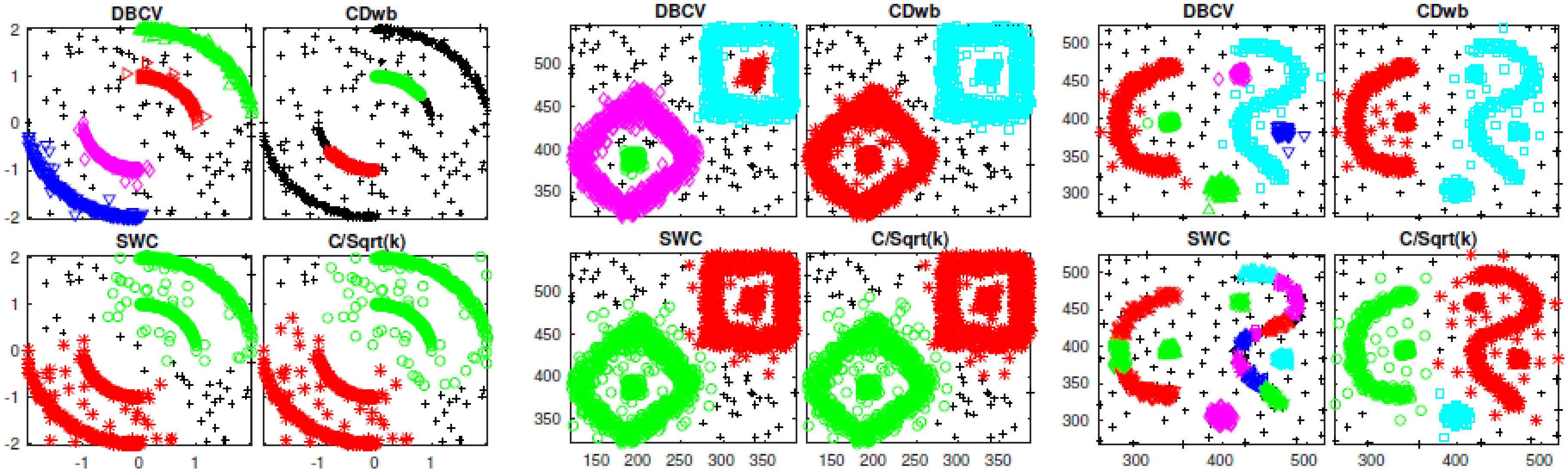
- Validation of one cluster is based on
  - Density sparsness: maximum edge of its MST
  - Density separation: minimum MRD between clusters

$$V_C(C_i) = \frac{\min_{j \neq i} \left(D_{Sep}(C_i, C_j)\right) - D_{Sp}(C_i)}{\max\left(\min_{j \neq i}(D_{Sep}(C_i, C_j)), D_{Sp}(C_i)\right)}$$

$$\text{DBCV}(\mathcal{C}) = \sum_{C_i \in \mathcal{C}} \frac{|C_i|}{|\mathbf{X}|} V_C(C_i)$$

Density-based Clustering Validation (DBCV)

- Adapted competitors to handle noise
  - Noise is discarded with proportional penalty
- Criteria evaluated on synthetic and real datasets
  - Promising results on both types of data



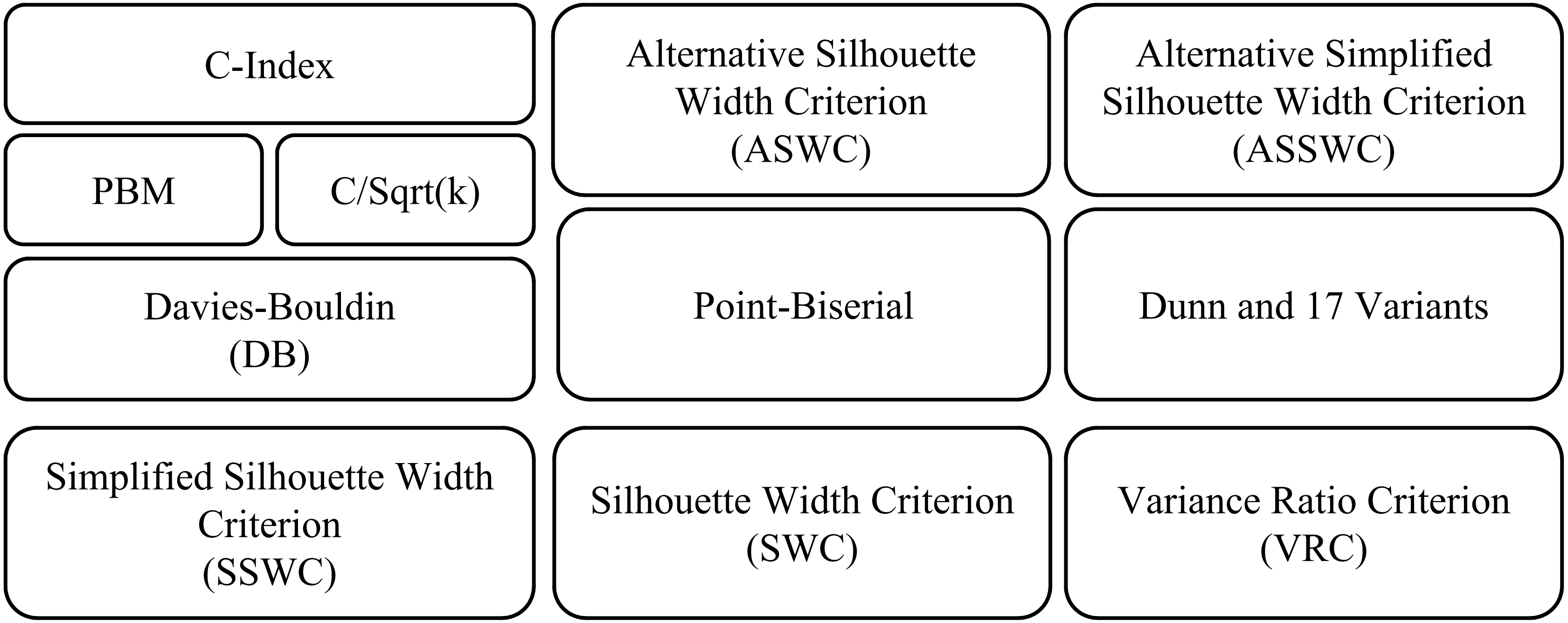
Ensembles of relative validity criteria

Ad-hoc ensembles

Ensembles based on heuristic selection

Ensembles of relative validity criteria

- Relative validity criteria



C-Index

PBM

C/Sqrt(k)

Davies-Bouldin (DB)

Alternative Silhouette Width Criterion (ASWC)

Point-Biserial

Alternative Simplified Silhouette Width Criterion (ASSWC)

Dunn and 17 Variants

Simplified Silhouette Width Criterion (SSWC)

Silhouette Width Criterion (SWC)

Variance Ratio Criterion (VRC)

These are the measures we used, but the list goes on...

Ensembles of relative validity criteria

- Different formulations, similar concepts
  - Separation and compactness

- Ensembles of validity measures
  - So far only ad-hoc approaches

- How well do these ad-hoc approaches behave?

- Can we do better?

Ensembles of relative validity criteria

Hyphothesis 2:

Ensembles of relative validity criteria built on the basis of an ad-hoc selection of their constituent members provide very limited practical benefits.

Hyphothesis 3:

Ensembles built on the basis of a simple, yet principled selection of their constituent members, perform better than those built in an ad-hoc fashion and provide more reliable evaluations than the ones obtained with individual criteria.

Ensembles of relative validity criteria

Ad-hoc ensembles

Ensembles based on heuristic selection

Ad-hoc ensembles

- Datasets
  - 972 Synthetic datasets from Vendramin et al. 2010
  - 400 datasets from ALOI (Geusebroek et al., 2005)

- Partitions
  - HCA´s and k-means with $k \in [2, \lceil \sqrt{n} \rceil]$

- 28 different relative validity criteria
  - All combinations of 3 and 5 measures

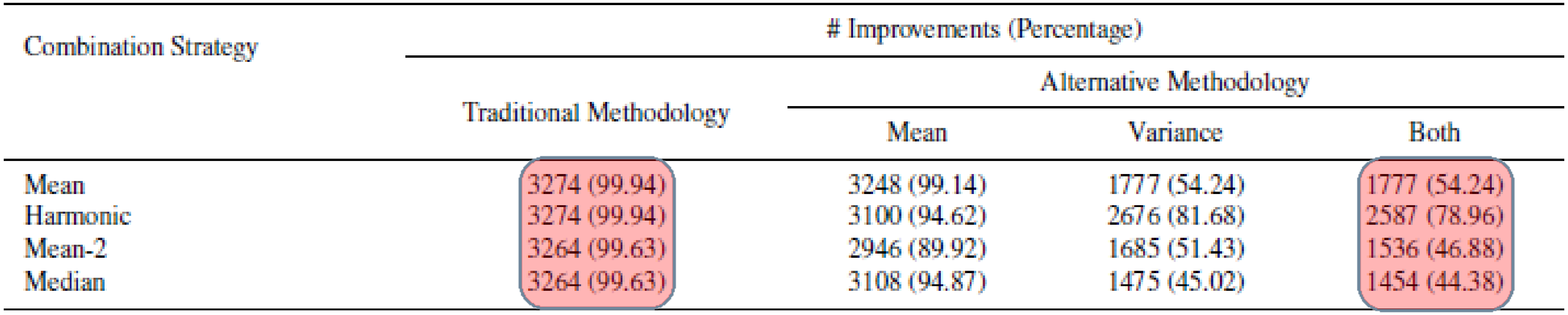
Ad-hoc ensembles

- How do we evaluate measures/ensembles?
  - Number of hits w.r.t. actual number of clusters
  - Correlation with external measure
- Different score-based combination strategies
  - Mean, Mean-2, Median, and Harmonic

Ad-hoc ensembles

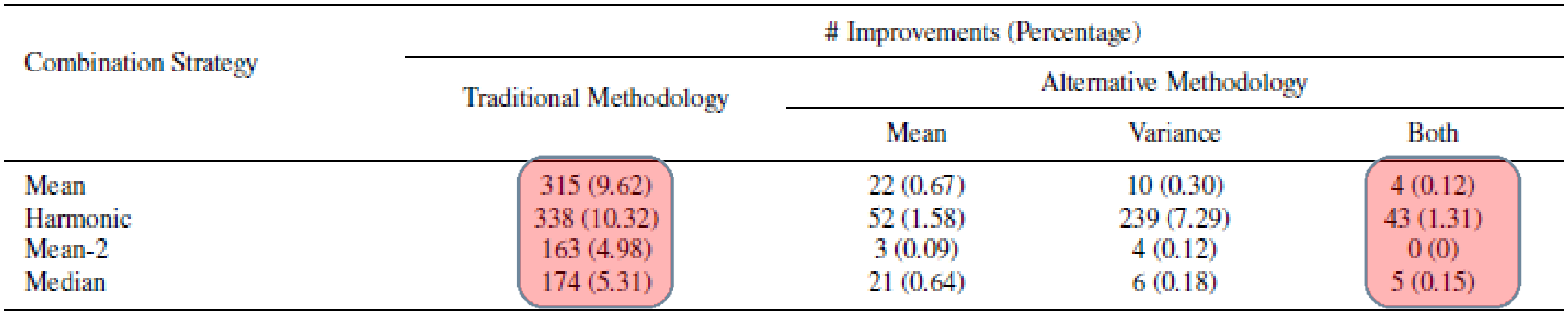
- Results for synthetic data, three criteria combinations

Improvements over the worst criterion



Combination Strategy	# Improvements (Percentage)			
	Traditional Methodology	Alternative Methodology		
		Mean	Variance	Both
Mean	3274 (99.94)	3248 (99.14)	1777 (54.24)	1777 (54.24)
Harmonic	3274 (99.94)	3100 (94.62)	2676 (81.68)	2587 (78.96)
Mean-2	3264 (99.63)	2946 (89.92)	1685 (51.43)	1536 (46.88)
Median	3264 (99.63)	3108 (94.87)	1475 (45.02)	1454 (44.38)

Improvements over all criteria



Combination Strategy	# Improvements (Percentage)			
	Traditional Methodology	Alternative Methodology		
		Mean	Variance	Both
Mean	315 (9.62)	22 (0.67)	10 (0.30)	4 (0.12)
Harmonic	338 (10.32)	52 (1.58)	239 (7.29)	43 (1.31)
Mean-2	163 (4.98)	3 (0.09)	4 (0.12)	0 (0)
Median	174 (5.31)	21 (0.64)	6 (0.18)	5 (0.15)

Ensembles of relative validity criteria

Ad-hoc ensembles

Ensembles based on heuristic selection

Ensembles based on heuristic selection

- Select ensemble members based on two principles
  - Effectiveness
  - Complementarity
- Also considered rank-based combination strategies
  - No need of score normalization
- Same configuration as in previous experiments
  - Clustering algorithms and ranges for *k*

Ensembles based on heuristic selection

- Estimating complementarity and effectiveness
  - 972 synthetic datasets

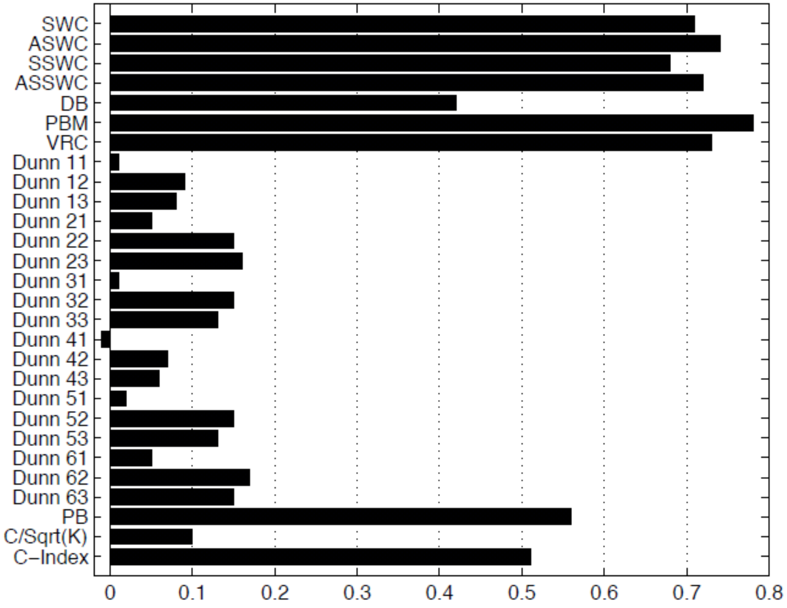
- We later evaluate the ensembles on unseen data

- Proeminent ensembles
  - Selected based on average results w.r.t. all aggregators

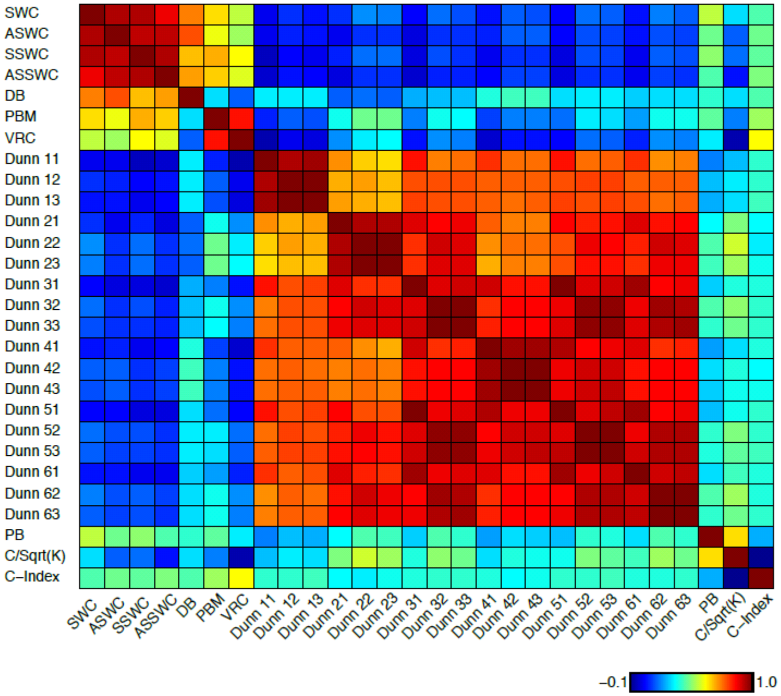
Ensembles based on heuristic selection

- Evaluations based on 972 synthetic datasets

Effectiveness



Complementarity



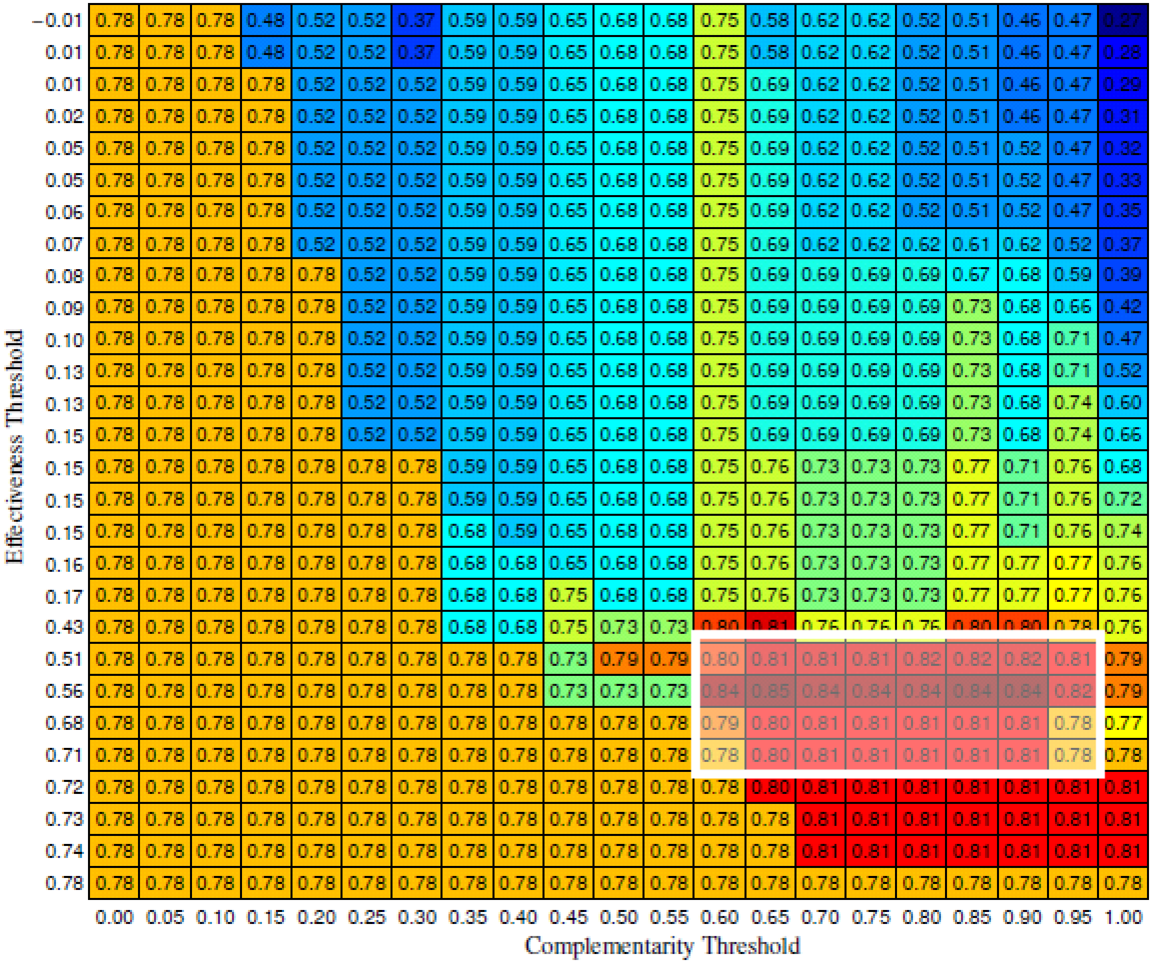
Ensembles based on heuristic selection

- How do we select the ensemble members?
  1. Add the criterion with highest effectiveness
  2. Add criteria that do not violate effectiveness and complementarity restrictions (ordered by effectiveness)

- Different thresholds are used for each restriction
  - Effectiveness: 28 thresholds (number of rel. criteria)
  - Complementarity: 0.05 increments (21 threshold in [0,1])

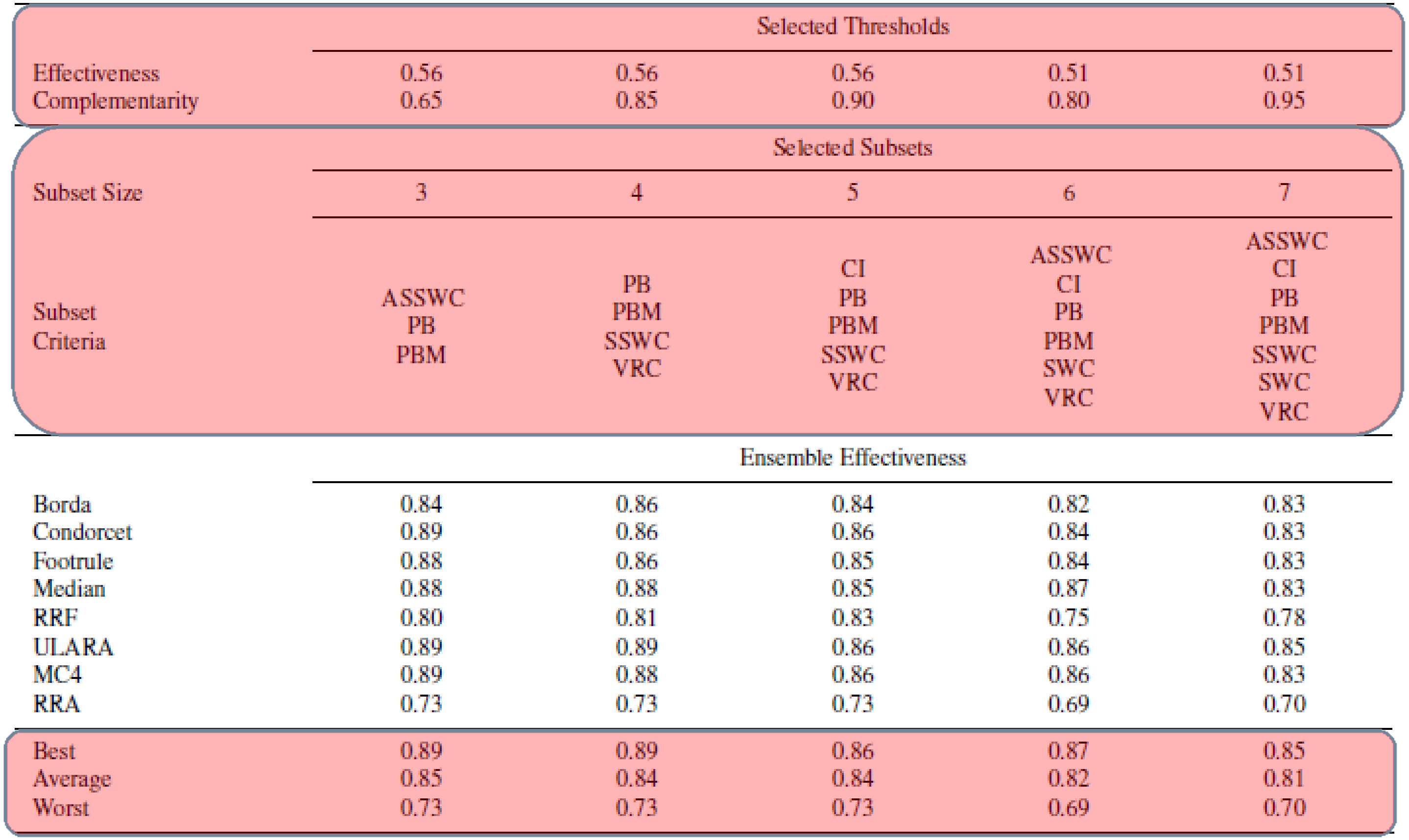
Ensembles based on heuristic selection

- □ Results w.r.t. average for all combination methods



	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
−0.01	0.78	0.78	0.78	0.48	0.52	0.52	0.37	0.59	0.59	0.65	0.68	0.68	0.75	0.58	0.62	0.62	0.52	0.51	0.46	0.47	0.27
0.01	0.78	0.78	0.78	0.48	0.52	0.52	0.37	0.59	0.59	0.65	0.68	0.68	0.75	0.58	0.62	0.62	0.52	0.51	0.46	0.47	0.28
0.01	0.78	0.78	0.78	0.78	0.52	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.62	0.62	0.52	0.51	0.46	0.47	0.29
0.02	0.78	0.78	0.78	0.78	0.52	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.62	0.62	0.52	0.51	0.46	0.47	0.31
0.05	0.78	0.78	0.78	0.78	0.52	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.62	0.62	0.52	0.51	0.52	0.47	0.32
0.05	0.78	0.78	0.78	0.78	0.52	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.62	0.62	0.52	0.51	0.52	0.47	0.33
0.06	0.78	0.78	0.78	0.78	0.52	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.62	0.62	0.52	0.51	0.52	0.47	0.35
0.07	0.78	0.78	0.78	0.78	0.52	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.62	0.62	0.62	0.61	0.62	0.52	0.37
0.08	0.78	0.78	0.78	0.78	0.78	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.69	0.69	0.69	0.67	0.68	0.59	0.39
0.09	0.78	0.78	0.78	0.78	0.78	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.69	0.69	0.69	0.73	0.68	0.66	0.42
0.10	0.78	0.78	0.78	0.78	0.78	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.69	0.69	0.69	0.73	0.68	0.71	0.47
0.13	0.78	0.78	0.78	0.78	0.78	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.69	0.69	0.69	0.73	0.68	0.71	0.52
0.13	0.78	0.78	0.78	0.78	0.78	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.69	0.69	0.69	0.73	0.68	0.74	0.60
0.15	0.78	0.78	0.78	0.78	0.78	0.52	0.52	0.59	0.59	0.65	0.68	0.68	0.75	0.69	0.69	0.69	0.69	0.73	0.68	0.74	0.66
0.15	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.59	0.59	0.65	0.68	0.68	0.75	0.76	0.73	0.73	0.73	0.77	0.71	0.76	0.68
0.15	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.59	0.59	0.65	0.68	0.68	0.75	0.76	0.73	0.73	0.73	0.77	0.71	0.76	0.72
0.15	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.68	0.59	0.65	0.68	0.68	0.75	0.76	0.73	0.73	0.73	0.77	0.71	0.76	0.74
0.16	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.68	0.68	0.65	0.68	0.68	0.75	0.76	0.73	0.73	0.73	0.77	0.77	0.77	0.76
0.17	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.68	0.68	0.75	0.68	0.68	0.75	0.76	0.73	0.73	0.73	0.77	0.77	0.77	0.76
0.43	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.68	0.68	0.75	0.73	0.73	0.80	0.81	0.76	0.76	0.76	0.80	0.80	0.78	0.76
0.51	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.73	0.79	0.79	0.80	0.81	0.81	0.81	0.82	0.82	0.82	0.81	0.79
0.56	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.73	0.73	0.73	0.84	0.85	0.84	0.84	0.84	0.84	0.84	0.82	0.79
0.68	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.79	0.80	0.81	0.81	0.81	0.81	0.81	0.78	0.77
0.71	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.80	0.81	0.81	0.81	0.81	0.81	0.78	0.78
0.72	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.80	0.81	0.81	0.81	0.81	0.81	0.81	0.81
0.73	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.81	0.81	0.81	0.81	0.81	0.81	0.81
0.74	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.81	0.81	0.81	0.81	0.81	0.81	0.81
0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78

Ensembles based on heuristic selection



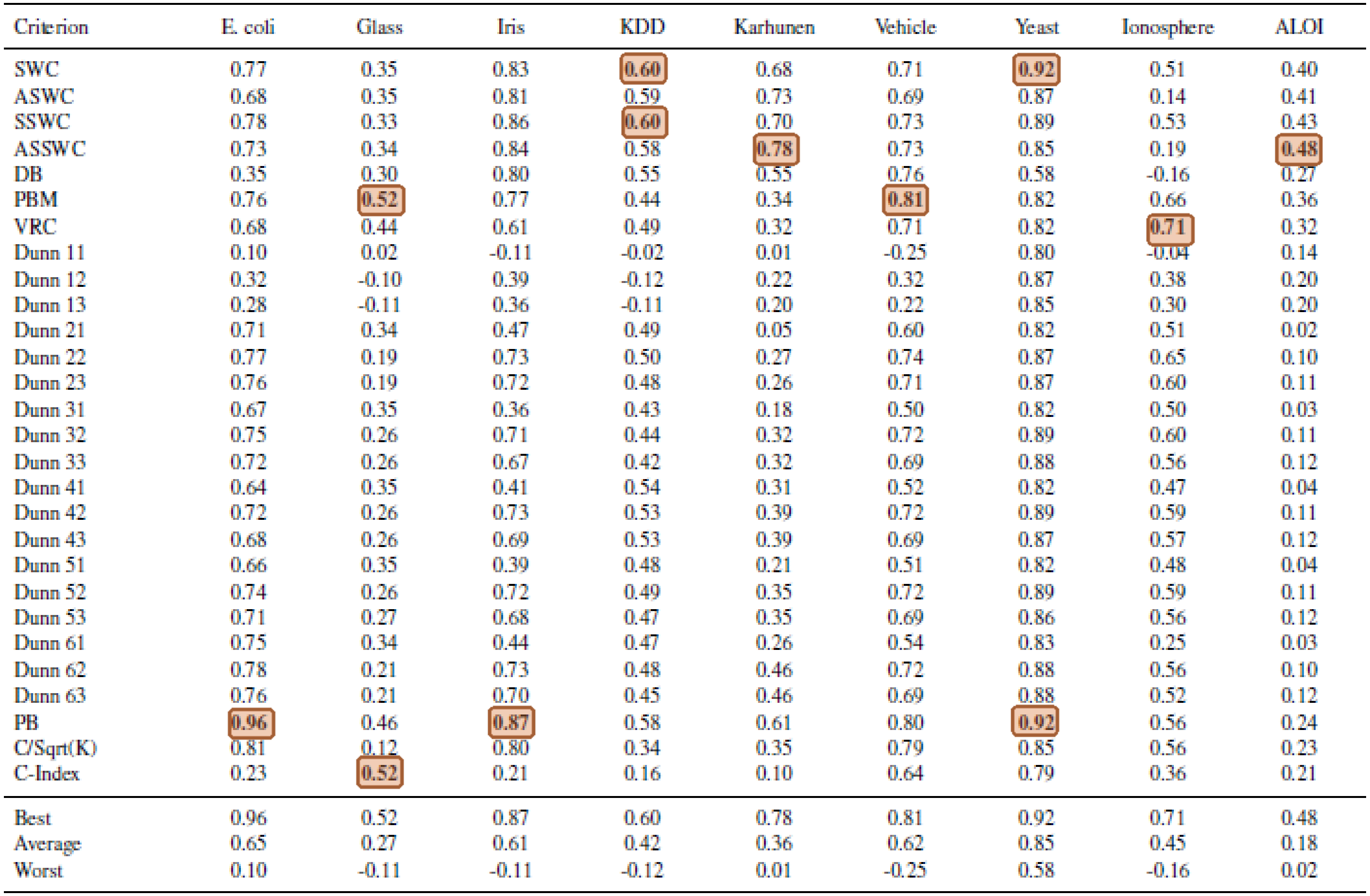
	Selected Thresholds				
Effectiveness Complementarity	0.56 0.65	0.56 0.85	0.56 0.90	0.51 0.80	0.51 0.95
	Selected Subsets				
Subset Size	3	4	5	6	7
Subset Criteria	ASSWC PB PBM	PB PBM SSWC VRC	CI PB PBM SSWC VRC	ASSWC CI PB PBM SWC VRC	ASSWC CI PB PBM SSWC SWC VRC
	Ensemble Effectiveness				
Borda	0.84	0.86	0.84	0.82	0.83
Condorcet	0.89	0.86	0.86	0.84	0.83
Footrule	0.88	0.86	0.85	0.84	0.83
Median	0.88	0.88	0.85	0.87	0.83
RRF	0.80	0.81	0.83	0.75	0.78
ULARA	0.89	0.89	0.86	0.86	0.85
MC4	0.89	0.88	0.86	0.86	0.83
RRA	0.73	0.73	0.73	0.69	0.70
Best	0.89	0.89	0.86	0.87	0.85
Average	0.85	0.84	0.84	0.82	0.81
Worst	0.73	0.73	0.73	0.69	0.70

Ensembles based on heuristic selection

- Evaluation of selected ensemble members
  - Different collection of datasets
  - ALOI datasets
    - 400 datasest (results depicted as a single value)
  - Seven UCI datasets
    - E. Coli, Glass, Iris, Kdd, Karhunen, Vehicle, and Ionosphere
  - Datasets from Yeung et al. 2001
    - Yeast Galactose

Ensembles based on heuristic selection

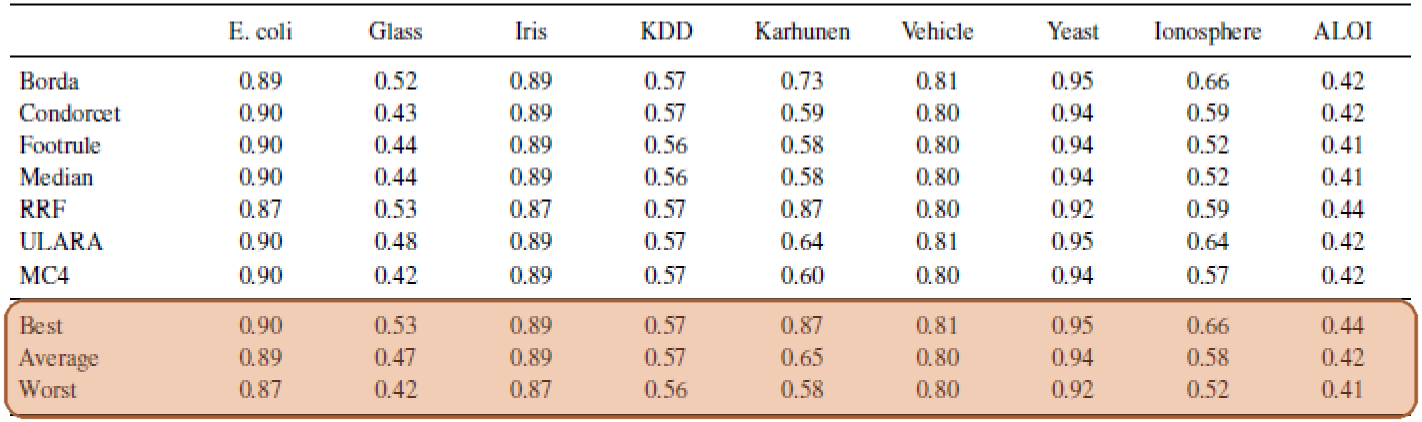
How effective are single criterion on these datasets



Criterion	E. coli	Glass	Iris	KDD	Karhunen	Vehicle	Yeast	Ionosphere	ALOI
SWC	0.77	0.35	0.83	0.60	0.68	0.71	0.92	0.51	0.40
ASWC	0.68	0.35	0.81	0.59	0.73	0.69	0.87	0.14	0.41
SSWC	0.78	0.33	0.86	0.60	0.70	0.73	0.89	0.53	0.43
ASSWC	0.73	0.34	0.84	0.58	0.78	0.73	0.85	0.19	0.48
DB	0.35	0.30	0.80	0.55	0.55	0.76	0.58	-0.16	0.27
PBM	0.76	0.52	0.77	0.44	0.34	0.81	0.82	0.66	0.36
VRC	0.68	0.44	0.61	0.49	0.32	0.71	0.82	0.71	0.32
Dunn 11	0.10	0.02	-0.11	-0.02	0.01	-0.25	0.80	-0.04	0.14
Dunn 12	0.32	-0.10	0.39	-0.12	0.22	0.32	0.87	0.38	0.20
Dunn 13	0.28	-0.11	0.36	-0.11	0.20	0.22	0.85	0.30	0.20
Dunn 21	0.71	0.34	0.47	0.49	0.05	0.60	0.82	0.51	0.02
Dunn 22	0.77	0.19	0.73	0.50	0.27	0.74	0.87	0.65	0.10
Dunn 23	0.76	0.19	0.72	0.48	0.26	0.71	0.87	0.60	0.11
Dunn 31	0.67	0.35	0.36	0.43	0.18	0.50	0.82	0.50	0.03
Dunn 32	0.75	0.26	0.71	0.44	0.32	0.72	0.89	0.60	0.11
Dunn 33	0.72	0.26	0.67	0.42	0.32	0.69	0.88	0.56	0.12
Dunn 41	0.64	0.35	0.41	0.54	0.31	0.52	0.82	0.47	0.04
Dunn 42	0.72	0.26	0.73	0.53	0.39	0.72	0.89	0.59	0.11
Dunn 43	0.68	0.26	0.69	0.53	0.39	0.69	0.87	0.57	0.12
Dunn 51	0.66	0.35	0.39	0.48	0.21	0.51	0.82	0.48	0.04
Dunn 52	0.74	0.26	0.72	0.49	0.35	0.72	0.89	0.59	0.11
Dunn 53	0.71	0.27	0.68	0.47	0.35	0.69	0.86	0.56	0.12
Dunn 61	0.75	0.34	0.44	0.47	0.26	0.54	0.83	0.25	0.03
Dunn 62	0.78	0.21	0.73	0.48	0.46	0.72	0.88	0.56	0.10
Dunn 63	0.76	0.21	0.70	0.45	0.46	0.69	0.88	0.52	0.12
PB	0.96	0.46	0.87	0.58	0.61	0.80	0.92	0.56	0.24
C/Sqrt(K)	0.81	0.12	0.80	0.34	0.35	0.79	0.85	0.56	0.23
C-Index	0.23	0.52	0.21	0.16	0.10	0.64	0.79	0.36	0.21
Best	0.96	0.52	0.87	0.60	0.78	0.81	0.92	0.71	0.48
Average	0.65	0.27	0.61	0.42	0.36	0.62	0.85	0.45	0.18
Worst	0.10	-0.11	-0.11	-0.12	0.01	-0.25	0.58	-0.16	0.02

Ensembles based on heuristic selection

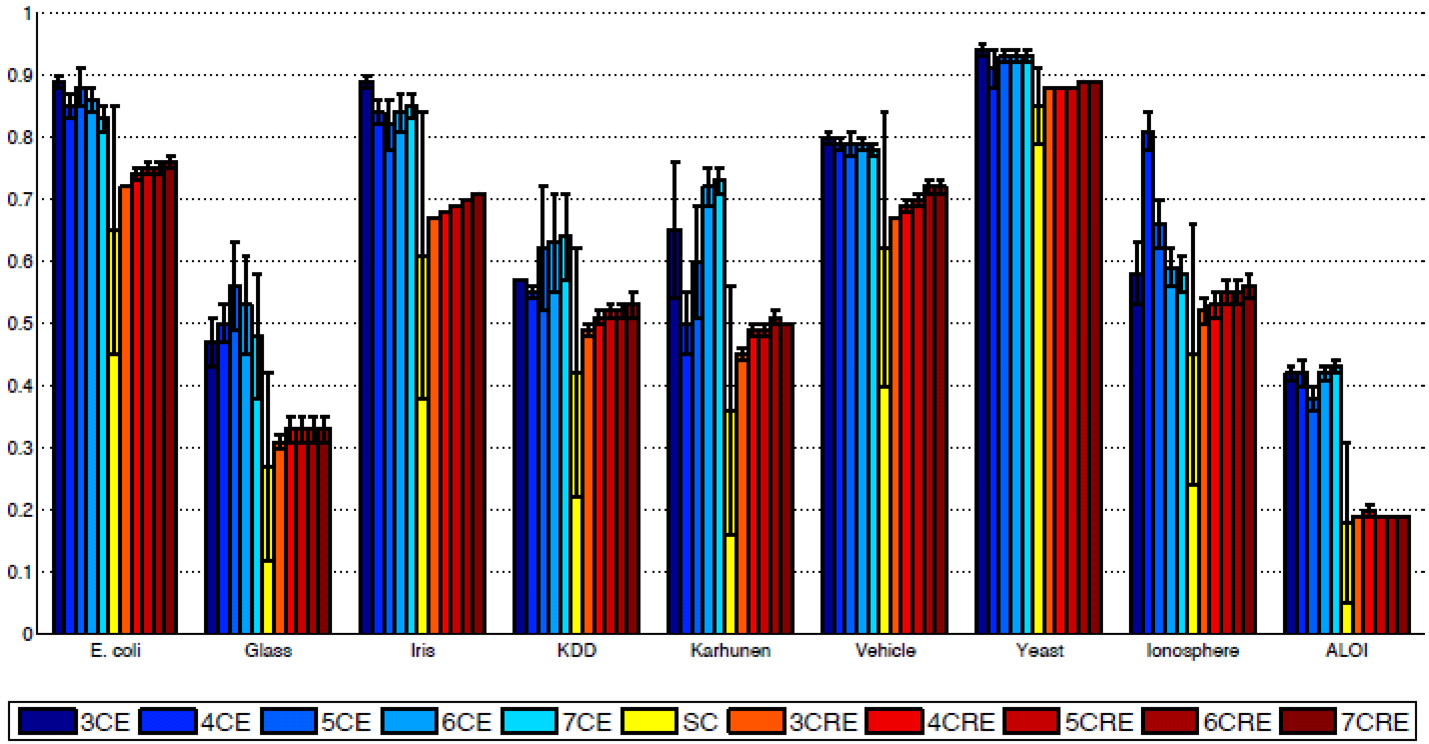
How effective are the ensembles on these datasets



	E. coli	Glass	Iris	KDD	Karhunen	Vehicle	Yeast	Ionosphere	ALOI
Borda	0.89	0.52	0.89	0.57	0.73	0.81	0.95	0.66	0.42
Condorcet	0.90	0.43	0.89	0.57	0.59	0.80	0.94	0.59	0.42
Footrule	0.90	0.44	0.89	0.56	0.58	0.80	0.94	0.52	0.41
Median	0.90	0.44	0.89	0.56	0.58	0.80	0.94	0.52	0.41
RRF	0.87	0.53	0.87	0.57	0.87	0.80	0.92	0.59	0.44
ULARA	0.90	0.48	0.89	0.57	0.64	0.81	0.95	0.64	0.42
MC4	0.90	0.42	0.89	0.57	0.60	0.80	0.94	0.57	0.42
Best	0.90	0.53	0.89	0.57	0.87	0.81	0.95	0.66	0.44
Average	0.89	0.47	0.89	0.57	0.65	0.80	0.94	0.58	0.42
Worst	0.87	0.42	0.87	0.56	0.58	0.80	0.92	0.52	0.41

Ensembles based on heuristic selection

- Results on datasets *not* used to select members



Ensembles of relative validity criteria

- Ad-hoc ensembles
  - Should be avoided
    - Unless the behavior of relative measures is knwon
  - Can avoid only the performance of the worst measure

- Heuristic selection of ensembles
  - Selection of ensemble members
    - Effectiveness and Complementarity
  - Simple heuristic, yet good results on unseen data

Distances for clustering gene expression data

Clustering algorithm dependent/independent evaluation

Results on microarray and RNA-Seq datasets

Distances for clustering gene expression data

- Distance selection is a key issue in clustering
- A number of measures in the literature
- Some specifically designed to short gene time-series
  - No evaluation of these measures
- Expansion of the work performed during the Master´s

Distances for clustering gene expression data

- Two main types of evaluation, w.r.t clustering algorithm
  - Dependent
    - Performance of clustering algorithm and distance measure *pair*
    - Measured w.r.t. ARI, if labels are available
    - Measured regarding # of enriched terms, if not
  - Independent
    - Intrinsic Separation Ability (Giancarlo, 2011)
    - Intrinsic Biological Separation Ability

Distances for clustering gene expression data

- Intrinsic Biological Separation Ability
  - Distance matrix (from data)
  - Biological distance matrix (semantic similarities from GO)

$$I_{\phi_1}(\mathrm{x}_i, \mathrm{x}_j) = \begin{cases} 1 & \text{if } \mathrm{D}(i,j) \leq \phi_1 \\ 0 & \text{otherwise} \end{cases} \qquad J_{\phi_2}(\mathrm{x}_i, \mathrm{x}_j) = \begin{cases} 1 & \text{if } \mathrm{B}(i,j) \leq \phi_2 \\ 0 & \text{otherwise} \end{cases}$$

  - Considering two thresholds, multiple ROC analyses
    - Measures thes agreement between them

Distances for clustering gene expression data

Hyphothesis 4:

External information, in the form of semantic similarities from the GO, can be employed to evaluate the suitability of distances among pairs of gene time-series for the task of clustering, independently from the bias of a particular clustering algorithm.

Distances for clustering gene expression data

- Microarray data
  - Evaluated a total of 15 distance measures
    - Considered with 4 clustering algorithms (SL, CL, AL, KM)
  - Distance measures evaluated on two settings
    - 35 cancer benchmark data (de Souto et al, 2008)
    - 17 yeast time course data (Jaskowiak et al, 2013)
  - Also considered different noise levels during evaluation

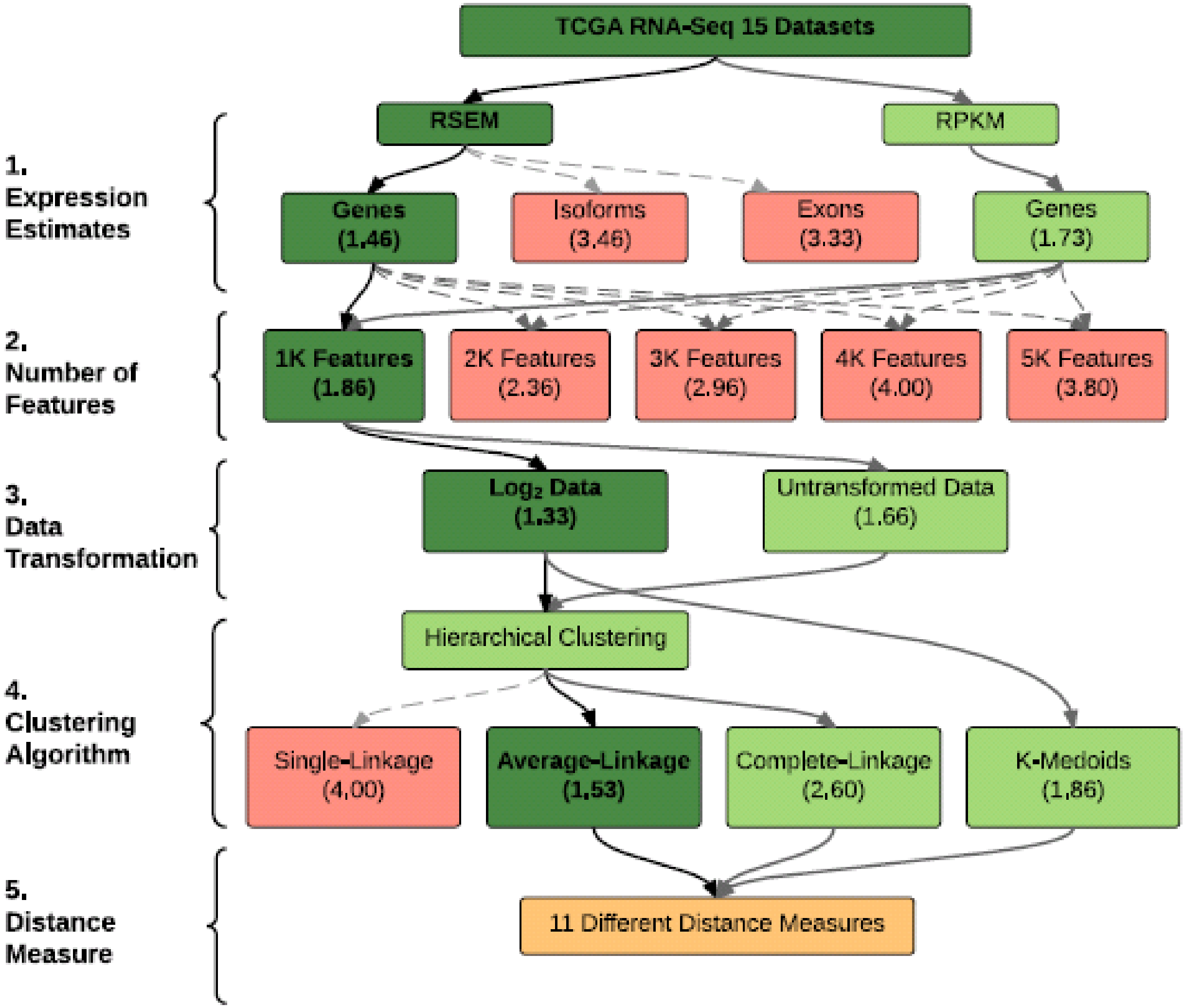
Distances for clustering gene expression data

- Microarray data
  - Different methodologies provided compatible results
  - Cancer datasets
    - Pearson and Symmetric Rank-Magnitude (robustness to noise)
  - Time-series datasets
    - YR1, YS1, and Jackknife

Distances for clustering gene expression data

- Also performed experiments on RNA-Seq data
  - Obtained raw data, compiled, pre-processed, ...
- Analysed the clustering of cancer samples
- Different experimental factors
  - Expression estimates, final number of features, whether to log-transform the data, clustering algorithm, and distance

Distances for clustering gene expression data



Distances for clustering gene expression data

- RNA-Seq data
  - Preference for gene quantifications (RPKM or RSEM)
  - About 1K features
  - Log-transformation improves value based measures
  - Average-Linkage, k-medoids
    - Rank-based measures (Spearman, Kendall, Goodman-Kruskal)

5

Biological validation of gene clustering results

Semantic similarities employed with relative measures

Problems with external index, BHI

Biological validation of gene clustering results

- Previous work evaluated semantic similarities from the GO in limited context (Bolshakova et al., 2006)
  - Small number of genes (total of 63)

- Evaluate the potential of semantic similarities

- Combine their evaluations with data based ones

Biological validation of gene clustering results

Hyphothesis 5:

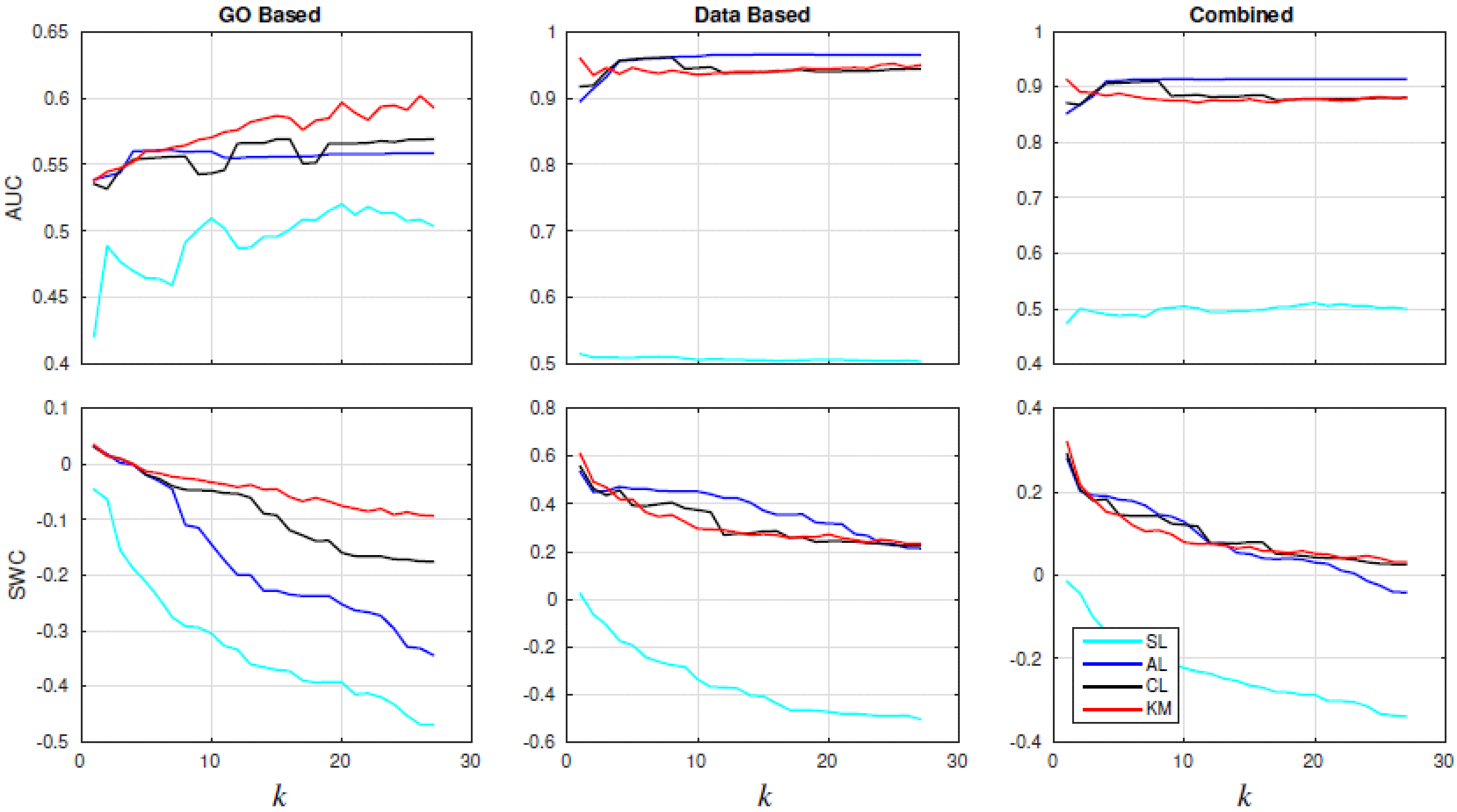
External information, in the form of semantic similarities from the GO, can be employed in the relative evaluation of clustering results, whether alone or combined with statistical similarities from the data.

Biological validation of gene clustering results

- Considered two relative measures
  - SWC and AUC
- Evaluations on realistic gene clustering datasets
  - 17 benchmark datasets (Jaskowiak et al., 2013)
- Four clustering algorithms
  - SL, AL, CL, KM

Biological validation of gene clustering results

- Results regarding one of the datasets (*elutriation*)



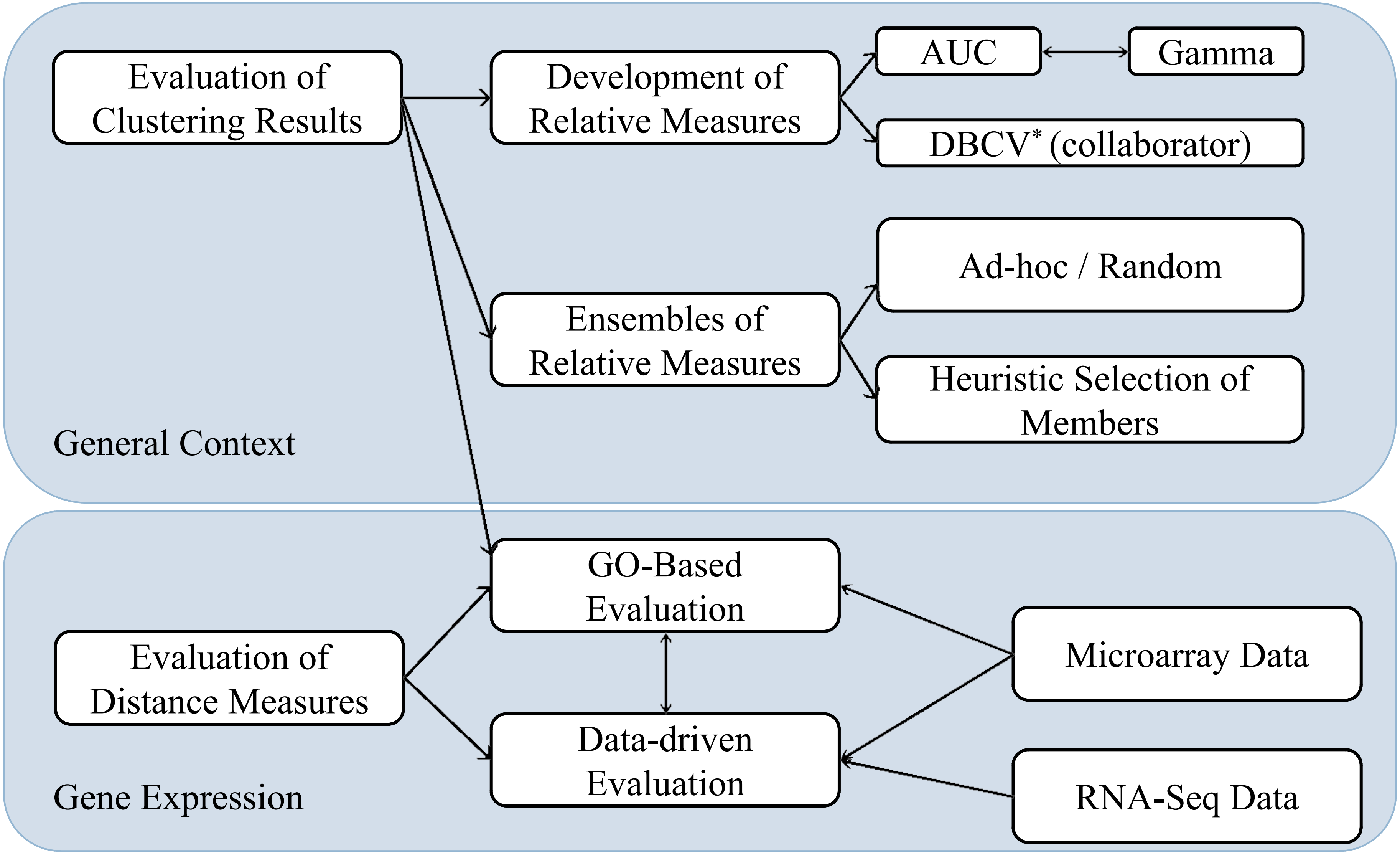
Biological validation of gene clustering results

- External measures in gene time-series evaluation
  - Biological Homogeneity Index (BHI)
    - One of the most commonly employed measures
    - Depends on term selection (external labels)
  - Undesired properties
    - Violates cluster completeness
  - If term selection is done
    - Other external measures should be preferred

6 Conclusions and future work

Contributions, publications, and future work

Conclusions



Evaluation of Clustering Results

Development of Relative Measures

AUC

Gamma

DBCV* (collaborator)

Ensembles of Relative Measures

Ad-hoc / Random

Heuristic Selection of Members

General Context

GO-Based Evaluation

Evaluation of Distance Measures

Data-driven Evaluation

Microarray Data

RNA-Seq Data

Gene Expression

Conclusions

- Publications directly related to the author´s thesis
  - Journals
    - JASKOWIAK, P.A.; MOULAVI D.; FURTADO, A.C.S.; CAMPELLO, R.J.G.B.; ZIMEK, A.; SANDER, J. On Strategies for Building Efective Ensembles of Relative Clustering Validity Criteria. Knowledge and Information Systems (KAIS) --- In Print.
    - JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; COSTA, I. G.. On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics, v. 15, p. S2, 2014.
    - JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; COSTA, I. G.. Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (Print), v. 10, p. 845-857, 2013.

Conclusions

- Publications directly related to the author´s thesis
  - Conferences
    - MOULAVI, D.; JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; ZIMEK, A.; SANDER, J.. Density-Based Clustering Validation. In: SIAM International Conference on Data Mining, 2014, Philadelphia, US. Proc. of the 14th SIAM International Conference on Data Mining, 2014. p. 1-9.

    - JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; COSTA, I. G.. Evaluating Correlation Coefficients for Clustering Gene Expression Profiles of Cancer. In: VII Brazilian Symposium on Bioinformatics, 2012, Campo Grande, v. 7409. p. 120-131.

    - VENDRAMIN, L.; JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.. On the Combination of Relative Clustering Validity Criteria. In: 25th International Conference on Scientific and Statistical Database Management, 2013, Baltimore, US, New York: ACM Press, 2013. p. 1-12.

Conclusions

- Publications done in collaboration
  - Journals
    - de SOUTO, M.C.P.; JASKOWIAK, P.A.; COSTA, I. G. Impact of missing data imputation methods on gene expression clustering and classification. BMC Bioinformatics, p.09, 2015.
    - BARROS, R. C.; JASKOWIAK, P. A.; CERRI, R.; CARVALHO, A. C. P. L. F.. A framework for bottom-up induction of oblique decision trees. Neurocomputing, v. 135, p. 3-12, 2014.

Conclusions

- Publications done in collaboration
  - Conferences
    - JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.. A Cluster Based Hybrid Feature Selection Approach. 2015 Brazilian Conference on Intelligent Systems (BRACIS 2015).
    - JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.. Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data. In: VI Brazilian Symposium on Bioinformatics, 2011, Brasília. Proc. of the 6th Brazilian Symposium on Bioinformatics. p. 1-8.
    - BARROS, R. C.; CERRI, R.; JASKOWIAK, P. A.; CARVALHO, A. C. P. L. F.. A Bottom-Up Oblique Decision Tree Induction Algorithm. In: International Conference on Intelligent Systems Design and Applications, 2011, Córdoba. Proc. of the 11th International Conference on Intelligent Systems Design and Applications, 2011. p. 450-456.

Conclusions

- Future works
  - Further developments regarding AUC
    - Consider other related measures, *e.g.*, AUPR
    - Publish the results we obtained so far
  - Density-based clustering validation
    - Different graph models and density estimates
  - Meta validation of clustering results
    - Automatic selection of measures / construction of ensembles

Conclusions

- Future works
  - Analysis of RNA-Seq data
    - Publish the results we obtained so far
    - Evaluation of feature selection methods
  - Evaluation of gene clustering results
    - Investigate different external measures
    - How selection of terms impact their performance

Acknowledgments

FAPESP

CAPES

ICMC USP
SÃO CARLOS

UNIVERSITY OF
ALBERTA

Centro
de Informática
U · F · P · E