

Clustering Mixed Numerical and Low Quality Categorical Data: Significance Metrics on a Yeast Example

Bill Andreopoulos, Aijun An, Xiaogang Wang
York University

IQIS 2005
June 17, 2005

What is clustering?

- *Clustering* aims to partition a set of objects into clusters, so that objects with similar characteristics are clustered together and different clusters contain objects with dissimilar characteristics.

Contributions, motivation, inspiration

- We designed the **M-BILCOM** clustering tool for numerical data sets that incorporates in the clustering process categorical attribute values (CAs) and confidence values (CVs) indicating the confidence that the CAs are correct.
- M-BILCOM was mainly inspired by numerical gene expression data sets from DNA microarray studies, where CAs and CVs can be derived from Gene Ontology annotations and Evidence Codes.
- One of the main advantages of this algorithm is that it offers the opportunity to apply novel significance metrics for spotting the most significant CAs in a cluster when analyzing the results.

<u>Gene x</u>	0.15, 1.0,....., 0.5, 0.75
	cytokinesis 0.5, budding 1.0, mitosis 0.8

K-Modes clustering algorithm

- k-Modes is a clustering algorithm that deals with categorical data only.
- The k-Modes clustering algorithm requires the user to specify from the beginning the number of clusters to be produced and the algorithm builds and refines the specified number of clusters.
- Each cluster has a mode associated with it. Assuming that the objects in the data set are described by m categorical attributes, the mode of a cluster is a vector $Q=\{q_1, q_2, \dots, q_m\}$ where q_i is the most frequent value for the i th attribute in the cluster of objects.
- Given a data set and the number of clusters k , the k-Modes algorithm clusters the set as follows:
 - **Select initial k modes for k clusters.**
 - **For each object X**
 - Calculate the similarity between object X and the modes of all clusters.
 - Insert object X into the cluster c whose mode is the most similar to object X .
 - Update the mode of cluster c
 - **Retest the similarity of objects against the current modes. If an object is found to be closer to the mode of another cluster rather than its own cluster, reallocate the object to that cluster and update the modes of both clusters.**
 - **Repeat 3 until no or few objects change clusters after a full cycle test of all the objects.**
- A similarity metric is needed to choose the closest cluster to an object by computing the similarity between the cluster's mode and the object. Let $X=\{x_1, x_2, \dots, x_m\}$ be an object, where x_i is the value for the i th attribute, and $Q=\{q_1, q_2, \dots, q_m\}$ be the mode of a cluster. The similarity between X and Q can be defined as:
 - $\text{similarity}(X, Q) = \sum_{i=1}^m \delta(x_i, q_i)$
 - where $\delta(x_i, q_i) = \begin{cases} 1 & (x_i = q_i); \\ 0 & (x_i \neq q_i). \end{cases}$

Overview of M-BILCOM

- M-BILCOM is a combination of MULICsoft [2] and BILCOM [1].
- The basic idea of our algorithm is to do clustering at two levels, where the first level clustering imposes an underlying framework for the second level clustering, thus simulating a Bayesian prior as described in [1].
- The categorical similarity is emphasized at the first level and the numerical similarity at the second level.
- *The level one clusters are given as input to level two and the level two clusters are the output of the clustering process.*

Overview of M-BILCOM

- The process looks as in Figure 1. As shown, both level one and level two involve the same number of clusters, four in this example.

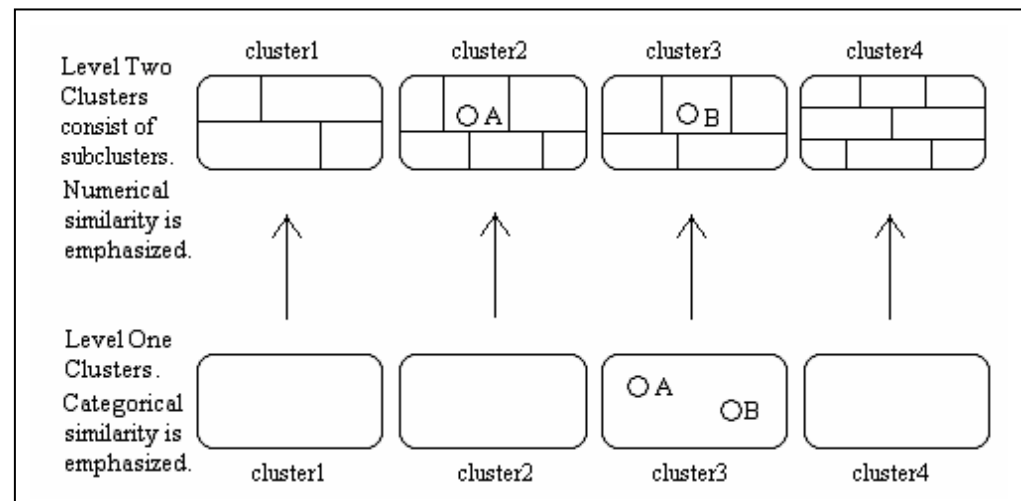
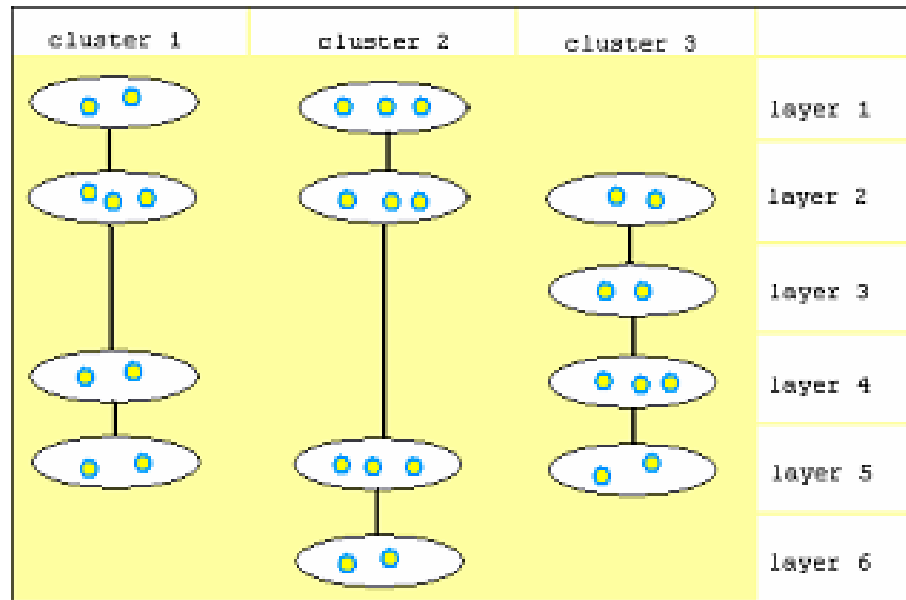


Figure 1. Overview of M-BILCOM clustering.

Overview of M-BILCOM

- Different types of data are used at levels one and two.
- At the first level the categorical data used represent something that has been observed to be true about the data set objects before the experiment takes place. For example the data at the first level might look as follows: SEX:male; STEROID:yes; FATIGUE:no; ANOREXIA:no .
- At the second level, on the other hand, the numerical data used represent the results of an experiment involving the data set objects. For example the data at the second level might look as follows: BILIRUBIN:0.39; ALBUMIN:2.1; PROTIME:10.

First Level Clustering: MULICsoft



- Figure 2. MULICsoft results. Each cluster consists of one or more different layers representing different similarities of the objects attached to the cluster.

How to distinguish first from second level objects?

- The question remains of which objects to be clustered at the first level of M-BILCOM.
- The first level objects are those whose comparison to the mode of the closest cluster yields a result that is greater than or equal to a threshold *minimum_mode_similarity*. The rest of the objects are used at the second level.
- The reason we choose to insert in the first level clusters just the objects whose similarity to the closest mode yields a value higher than a threshold *minimum_mode_similarity* is because the objects that yield a low similarity to the closest mode are more likely to be inserted in the wrong cluster, as we show in [1,2].
- Thus, the objects whose classification in clusters based on categorical similarity is not reliable enough, are clustered in the second level instead, where the numerical similarity of objects to second level clusters is more influential.

The MULICsoft similarity metric

- All CAs in an object have CVs ("weights") in the range 0.0 to 1.0 associated with them.
- The similarity metric used in MULICsoft for computing the similarity between a mode μ and an object o considers both the CAs and their CVs. Our similarity metric amplifies the object positions having high CVs, at pairs of CAs between an object o and a mode μ that have identical values.

$$\text{similarity}(o, \mu) = \sum_{i=1}^m \frac{6 - (4 \times w_i)}{5 - (4 \times w_i)} \times \sigma(o_i, \mu_i)$$

$$\sigma(o_i, \mu_i) = \begin{cases} 1 & (o_i = \mu_i); \\ 0 & (o_i \neq \mu_i). \end{cases}$$

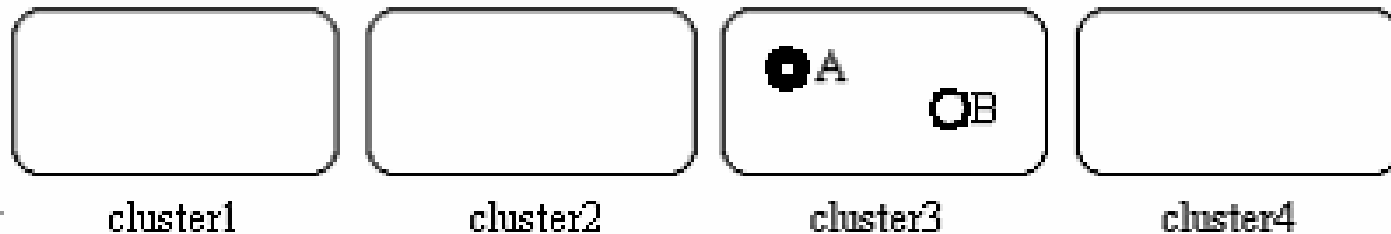
Second level clustering: BILCOM

- The first level result is the input to the second level.
- The second level clusters all of the data set objects, including the objects clustered at the first level.
- The second level uses numerical data type similarity and the first level result as a prior.
- The second level clustering consists of 5 steps, whose rationale is to simulate maximizing the numerator of a Bayesian equation.
- The second level result is the output of the BILCOM process.

Second level clustering: BILCOM

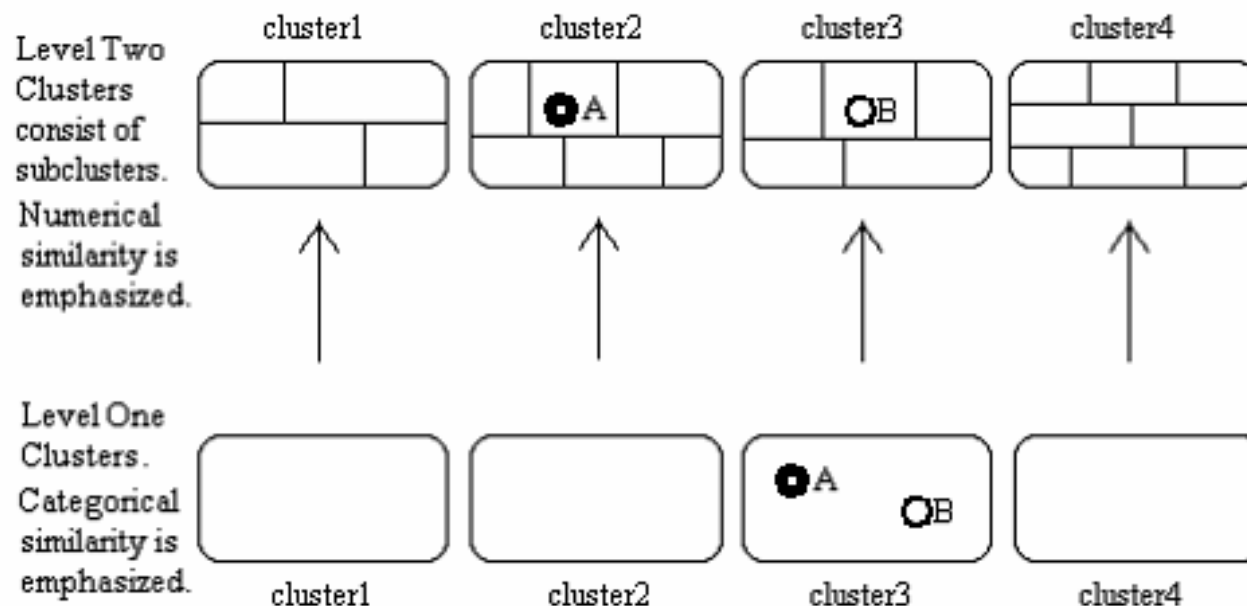
- *Step 1.* One object in each first level cluster is set as a *seed*, while all the rest of the objects in the cluster are set as *centers*.

Level One
Clusters.
Categorical
similarity is
emphasized.



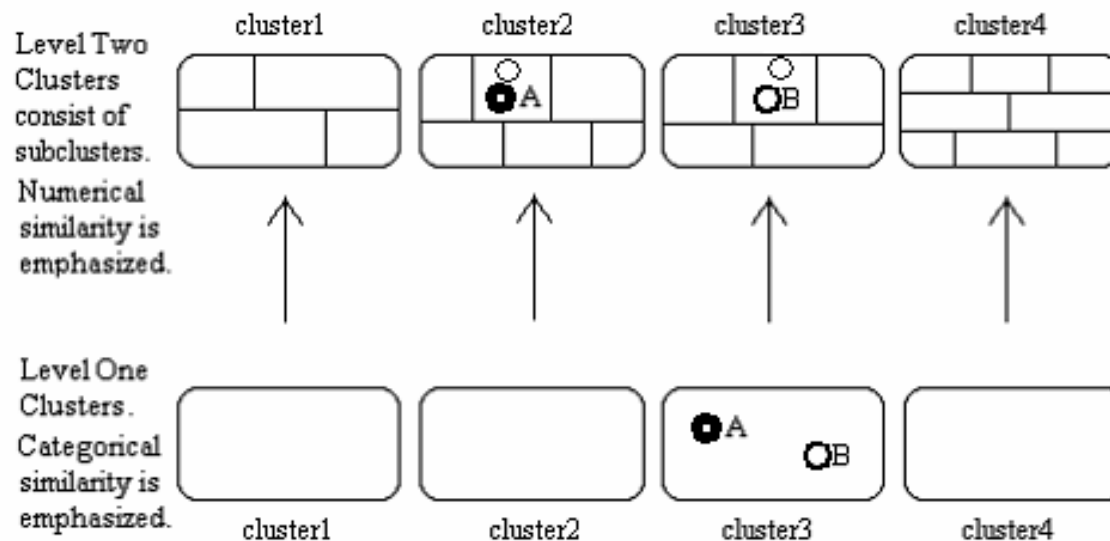
Second level clustering: BILCOM

- *Step 2.* Each *seed* and *center* is inserted in a new *second level subcluster*. The output of this step is a set of *subclusters*, referred to as *seed-containing* or *center-containing* subclusters, whose number equals the number of objects clustered at the first level.



Second level clustering: BILCOM

- *Step 3.* Each object that did not participate at the first level is inserted into the second level subcluster containing the most numerically similar *seed* or *center*. Numerical similarity for Steps 3-5 is determined by the *Pearson correlation coefficient* or the *Shrinkage-based similarity metric* introduced by Cherepinsky et al.



Second level clustering: BILCOM

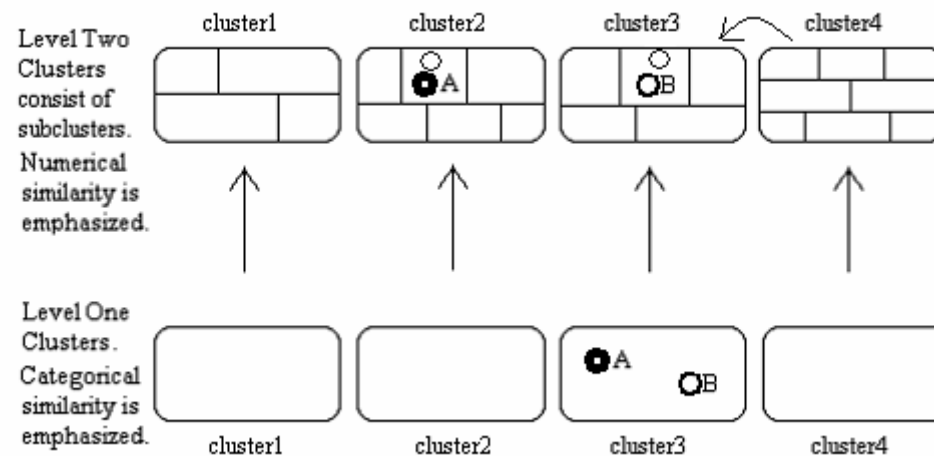
- *Step 4.* Each center-containing subcluster is merged with its most numerically similar seed-containing subcluster. The most numerically similar seed-containing subcluster is found using our version of the ROCK goodness measure [14] that is evaluated between the center-containing subcluster in question and all seed-containing subclusters:

$$G(C_i, C_j) = \frac{\text{link}[C_i, C_j]}{\text{size}(C_i) \times \text{size}(C_j)}$$

- $\text{link}[C_i, C_j]$ stores the number of cross links between subclusters C_i and C_j , by evaluating $\sum(o_q \in C_i, o_r \in C_j) \text{link}(o_q, o_r)$. $\text{link}(o_q, o_r)$ is a boolean value specifying whether a link exists between objects o_q and o_r . A link is set between two objects if the objects' numerical similarity is higher than a value *minimum_numerical_similarity*.

Second level clustering: BILCOM

- *Step 5.* The loop below refines the step 4 subcluster merges. All variables take real values in the range 0.0-1.0.
- `repeat {`
- `foreach (center-containing_subcluster)`
- `if (numerical_similarity_of_center_subcluster_to_1st_level_seed_cluster ×`
 `categorical_similarity_of_center_to_seed_of_1st_level_cluster >`
 `numerical_similarity_of_center_subcluster_to_its_numerically_similar_2nd_level_cluster ×`
 `categorical_similarity_of_center_to_seed_of_its_numerically_similar_2nd_level_cluster)`
- `merge center-containing_subcluster to seed-containing_subcluster from 1st_level;`
- `} until (no center-containing_subcluster changes);`



Description of Real Yeast Data Sets

- This algorithm is designed with the goal of applying it to numerical data sets for which some CAs exist and the confidence that the CAs are correct varies.
- We used numerical data derived from gene expression studies on the yeast *Saccharomyces cerevisiae*. These data sets were produced at Stanford to study the yeast cell cycle across time and under various experimental conditions and are available from the SGD database [9, 23]. When clustering this data set, we consider each gene to be a 'data object'. The data set contained 6,200 objects.
- We represented CAs on a gene in terms of Gene Ontology (GO) and GOSlim annotations. GO is a dynamically controlled vocabulary that can be applied to many organisms, even as knowledge of gene and protein roles in cells is changing. GO is organized along the categories of molecular function, biological process and cellular location [12].

Description of Real Yeast Data Sets

- We attached CVs to the CAs to represent the confidence that the corresponding CA is correct.
- CVs are real numbers between 0.0 and 1.0, assigned to the CAs of a gene.
- Besides indicating the confidence that a CA is correct, the CVs on a gene also specify how strongly the gene's CAs should influence the clustering process.
- The CVs are also used in the significance metrics that we define later.
- We determined the CVs by using GO evidence codes. GO evidence codes symbolize the evidence that exists for any particular GO or GOSlim annotation [12].

TAS	-	1.0
IDA	-	1.0
IMP	-	0.8
IGI	-	0.8
IPI	-	0.8
ISS	-	0.5
IEP	-	0.5
NAS	-	0.2
IEA	-	0.1
ND	-	0.0
NR	-	0.0

Figure 5.
GO Evidence
Codes are
mapped to
CVs.

First Significance metric

- Given a resulting cluster, we assigned a P1-value to each CA in the cluster; the term 'P1-value' was derived from the statistical 'P-value'.
- A P1-value measures whether a cluster contains a CA of a particular type more frequently than would be expected by chance [25].
- A P1-value close to 0.0 indicates a frequent occurrence of the CA in the cluster, while a P1-value close to 1.0 its seldom occurrence.
- We multiplied the resulting P1-value with the reciprocal of the average of all CVs assigned to the CA in the cluster, $1/\text{avg}(\text{CV})$, thus resulting in what we call an *M-value*.

Second Significance metric

- This significance metric was inspired by the loop of step 5 that refines the subclusters composing a larger second level cluster.
- Specifically, each subcluster was assigned a significance number by evaluating a formula that considers both categorical (*CAsimilarity*) and numerical (*NAsimilarity*) similarity of the subcluster to the larger second level cluster:
 - $(weight1 * CAsimilarity) + (weight2 * NAsimilarity)$
- The *CAsimilarity* for a subcluster is computed by evaluating a categorical variation of ROCK's goodness measure [16] between the subcluster and its larger cluster and multiplying the result by the percentage of genes in the subcluster that were assigned to it on the basis of categorical similarity.
- The *NAsimilarity* for a subcluster is computed similarly, by evaluating a numerical variation of ROCK's goodness measure [16] between the subcluster and its larger cluster and multiplying the result by the percentage of genes in the subcluster that were assigned to it on the basis of numerical similarity (see step 3).
- We set *weight2* in our trials to be higher than *weight1*, to ensure proper consideration of the numerical similarity of a subcluster.

Experiments on Yeast Data

Table 1. Genes in the data set of Cherepinsky et al. [7] grouped by functions. This is our hypothesis about the correct clustering results.

Group	Activators	Genes	Functions
1	Swi4, Swi6	CLN1, CLN2, GIC1, MSB2, RSR1, BUD9, MNN1, OCH1, EXG1, KRE6, CWP1	Budding
2	Swi6, Mbp1	CLB5, CLB6, RNR1, RAD27, CDC21, DUN1, RAD51, CDC45, MCM2	DNA replication and repair
3	Swi4, Swi6	HTB1, HTB2, HTA1, HTA2, HTA3, HHO1	Chromatin
4	Fkh1	HHF1, HHT1, TEL2, ARP7	Chromatin
5	Fkh1	TEM1	Mitosis control
6	Ndd1, Fkh2, Mcm1	CLB2, ACE2, SWI5, CDC20	Mitosis control
7	Ace2, Swi5	CTS1, EGT2	Cytokinesis
8	Mcm1	MCM3, MCM6, CDC6, CDC46	Prereplication complex formation
9	Mcm1	STE2, FAR1	Mating

Experiments on Yeast Data

Table 2. Clustering results of AutoClass.			
Cluster	Genes		
1	CLN1, CLN2, GIC1, GIC2, MSB2, RSR1, BUD9, MNN1, OCH1, EXG1, KRE6, CWP1, CLB5, CLB6, RAD51, CDC45, HTB1, HTA2, HHO1, TEL2	{1->{{11,9}}, 2->{{4,16},{5,5}}, 3->{{3,17},{2,8}}, 4->{{1,19},{1,12},{2,8}}, 5->{{1,12}}, 6->{{4,9}}, 7->{{2,11}}, 8->{{4,9}}, 9->{{1,12},{1,9}} }.	FP = 265 FN = 32 Error = 297
2	ARP7, TEM1, CLB2, ACE2, SWI5, CDC20, CTS1, EGT2, MCM3, MCM6, CDC6, CDC46, STE2		
3	RNR1, RAD27, CDC21, DUN1, MCM2, HTB2, HTA1, HHF1, HHT1, FAR1		

Experiments on Yeast Data

Table 3. Clustering results of BILCOM using as numerical similarity metric between objects the Pearson Correlation Coefficient [7] and a max value for ϕ of 7.		<i>FP = 49</i> <i>FN = 47+13+24</i> <i>Error = 133</i>
Cluster	Genes	
1	RSR1, HHT1, ARP7, BUD9, CTS1	<i>1->{{4,4},{1,3},{1,1},{2,1},{2,0}},</i> <i>2->{{3,0},{3,2},{3,5}},</i> <i>3->{{3,0},{2,1}},</i> <i>4->{{2,3},{1,2},{1,4}},</i> <i>5->{{1,4}},</i> <i>6->{{2,3},{2,7}},</i> <i>7->{{1,4},{1,1}},</i> <i>8->{{2,0},{2,3}},</i> <i>9->{{2,3}} }.</i>
2	KRE6, CWP1	
3	RNR1, CDC45, MCM3, CDC46, MCM2	
4	EXG1, EGT2	
5	MCM6, CDC6	
6	HHF1, HTB2, HTA2	
7	HTB1, HTA1, HHO1	
8	GIC1, TEL2, GIC2, MSB2	
9	FAR1, STE2, ACE2, SWI5, TEM1	
10	RAD27, CDC21, DUN1	
11	CLN2, RAD51, MNN1, CLN1, CLB6, OCH1, CLB5, CLB2, CDC20	

Experiments on Yeast Data

Table 4. Clustering results of M-BILCOM using as numerical similarity metric between 2 objects the Pearson Correlation Coefficient [7] and a max value for ϕ of 7.

Cluster	Genes
1	RSR1, BUD9, CTS1
2	KRE6, ARP7, HHT1, CWP1
3	RNR1, CDC45, MCM3, STE2, CDC46, MCM2
4	EXG1, EGT2
5	MCM6, TEM1, CDC6
6	HHF1, HTB2, HTA2
7	HTB1, HHO1, HTA1
8	GIC1, TEL2, GIC2, MSB2
9	FAR1, ACE2, CDC20, SWI5
10	RAD27, CDC21, MNN1, DUN1, RAD51
11	CLN2, CLN1, CLB6, CLB5, OCH1, CLB2

$1 \rightarrow \{\{3,3\}, \{2,2\}, \{2,1\}, \{1,1\}, \{2,2\}, \{1,4\}\},$
 $2 \rightarrow \{\{4,1\}, \{3,3\}\},$
 $3 \rightarrow \{\{3,0\}, \{3,0\}\},$
 $4 \rightarrow \{\{2,2\}, \{1,2\}, \{1,3\}\},$
 $5 \rightarrow \{\{1,2\}\},$
 $6 \rightarrow \{\{3,1\}, \{1,5\}\},$
 $7 \rightarrow \{\{1,2\}, \{1,1\}\},$
 $8 \rightarrow \{\{2,4\}, \{2,1\}\},$
 $9 \rightarrow \{\{1,5\}, \{1,3\}\} \quad \}.$

$FP = 38$
 $FN = 35+49$
 $Error = 122$

Experiments on Yeast Data

Table 5. Comparative error rates of algorithms applied to the “perturbed” yeast data set.

Clustering Tool	Cherepinsky Error rate
M-BILCOM	122
BILCOM	133
AutoClass	297

Experiments on Simulated Data

- We assigned 6 CAs on each gene based on the NAs, representing the genes' action during cell cycle.
- The assignment of the CAs followed a pattern that simulates existing knowledge on the role of genes in the yeast cell cycle.
- The assigned CAs split the objects into a number of well-defined groups, which we attempt to retrieve using clustering; thus, a different set of attribute values had to be used for each group.
- This simulates the results by Spellman et al, who showed that in each cluster there is a consistent pattern of NAs that appear frequently and that different CAs are characteristic of different clusters [23].
- We assigned the 6 CAs using the following strategy: (1) The first CA split the genes into cell cycle phases and has the values G1, S, G2, M, M/G1, or unknown. This CA was set for each gene based on the experimental point at which the gene reaches its peak expression level, indicating what cell cycle phase it is likely to be involved in.
- Etc.

Experiments on Simulated Data

- Furthermore, we perturbed the CAs to simulate noise in the resulting data set. Our aim was to use M-BILCOM to retrieve the known underlying cluster structure effectively.
- A significant outcome of our experiments was to show that given the genes whose CAs were not perturbed in the simulation (most of which are likely to have high CVs) a fair number of genes were assigned to the correct clusters to which they were categorically similar and were not assigned to the incorrect clusters to which they might be numerically similar. The basis for this is that most of these genes had a high confidence overall.
- Another significant outcome of our experiments was to show that given the genes whose CAs were perturbed in the simulation (most of which are likely to have low CVs) a fair number of genes were assigned to the correct clusters to which they were likely to be numerically similar and were not assigned to the incorrect clusters to which they were categorically similar. The basis for this is that most of these genes had a low confidence overall.

Experiments on Simulated Data

Table 6. Results for clustering the data set into 20 clusters. We do not show results for clusters whose size was too small.

1 - Cluster #

2 - Number of objects in the cluster

3 - Most common values {A,B,C} on the objects' first 3 CAs

4 - Ratio X of objects in the cluster that had CAs modified during the simulation

5 - Ratio of X that had an original CA very close to {A,B,C}

6 - Ratio P of objects in the cluster that had an original CA very close to {A,B,C}

7 - Ratio of P that had its CAs modified during the simulation

8 - Number of merged second level subclusters

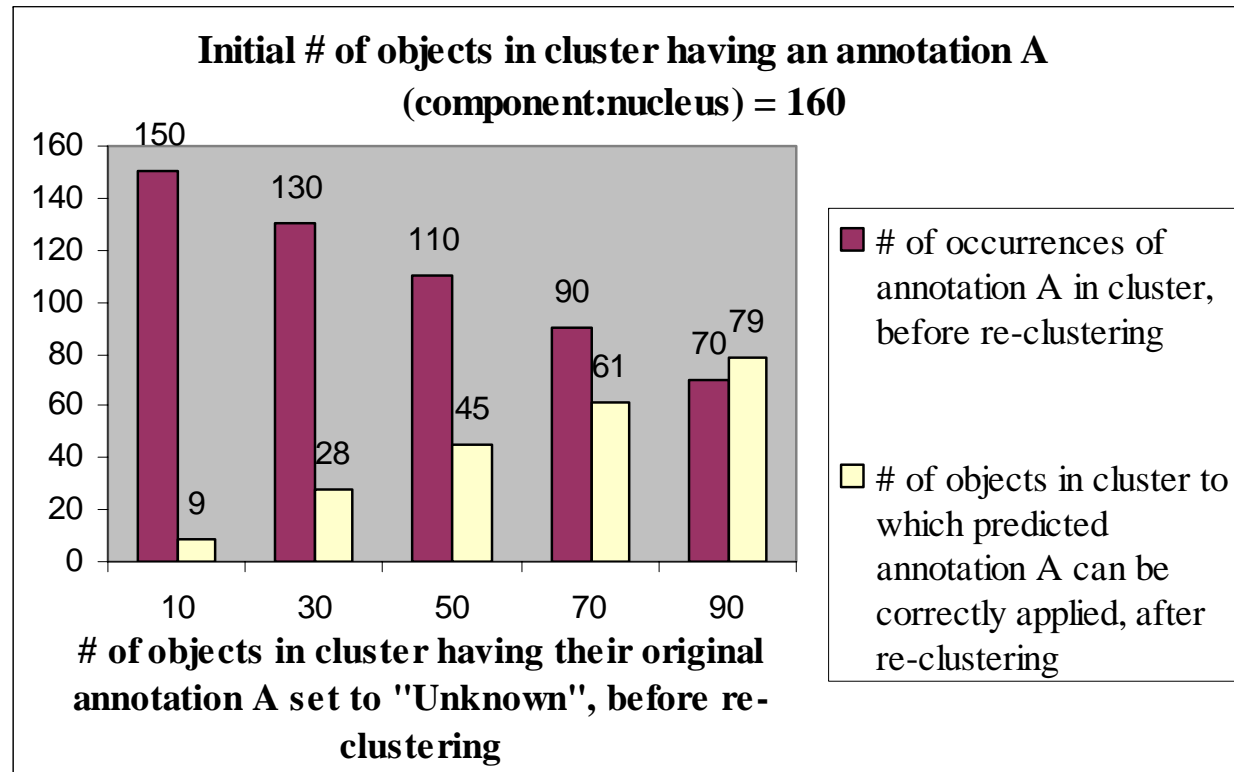
1	2	3	4	5	6	7	8
1	2032	{M,D,L}	616/ 2032	217/ 616	1047/ 2032	217/ 1047	537
2	1186	{MG1,E,N}	305/ 1186	202/ 305	1102/ 1186	202/ 1102	180
3	724	{G2,C,J}	177/ 724	111/ 177	672/ 724	111/ 672	121
4	317	{G1,A,F}	83/ 317	48/ 83	302/ 317	48/ 302	44
5	709	{S,B,H}	94/ 709	24/ 94	684/ 709	24/ 684	183
6	218	{M,D,M}	50/ 218	26/ 50	198/ 218	26/ 198	59
8	66	{MG1,E,N}	66/66	22/ 66	22/ 66	22/22	9
11	71	{MG1,E,N}	71/71	27/ 71	27/ 71	27/27	3
15	74	{MG1,E,N}	74/74	24/ 74	24/ 74	24/24	2
20	333	{S,B,H} and {S,B,I}	180/ 333	148/ 180	260/ 333	148/ 260	13

Applications of Significance Metrics

Table 7. CAs pointed out in 5 clusters as the most significant. The CAs pointed out in clusters 1-5 as having the lowest M-values - the most representative ones for the cluster - correlated with the CAs in the original cluster that were set to 'Unknown'.

Cluster	Some of the CAs pointed out in each cluster as having low M-values (meaning they occurred frequently and had high avg(CV)) after the CAs with the highest avg(CV) in each cluster were set to 'Unknown' and the set was clustered.
1	vacuolar membrane, ubiquitin-specific protease, small nuclear ribonucleoprotein complex, glycolysis, 3'-5' exoribonuclease, cytosolic small ribosomal subunit, lipid particle, cytosolic large ribosomal subunit, tricarboxylic acid cycle
2	rRNA modification, ATP dependent RNA helicase, nuclear pore, structural molecule, small nucleolar ribonucleoprotein complex, snoRNA binding, mediator complex
3	cytosol, proteasome endopeptidase, non-selective vesicle fusion, translation initiation factor
4	transcription initiation from Pol II promoter, general RNA polymerase II transcription factor, nucleus
5	endoplasmic reticulum membrane, component:endoplasmic reticulum

Applications of Significance Metrics



- **Figure 6. Results for different trials of setting occurrences of 'component:nucleus' to 'Unknown' and re-clustering the set.**

Discussion: Using Significance Metrics for Deriving Potential Gene Functions

- Biologists will find this method useful for deriving hints about potential functions of genes or proteins.
- The hints that are derived as to a gene's function can later be validated experimentally.
- This will save time and money from the experimentalists' side.
- In our experiments with the yeast cell cycle data set, the utility of the significance metrics (SMs) is especially evident from the fact that the vast majority of genes in each cluster or subcluster analyzed had all CAs set to 'Unknown' meaning that no knowledge exists.

Thank You!

billa@cs.yorku.ca