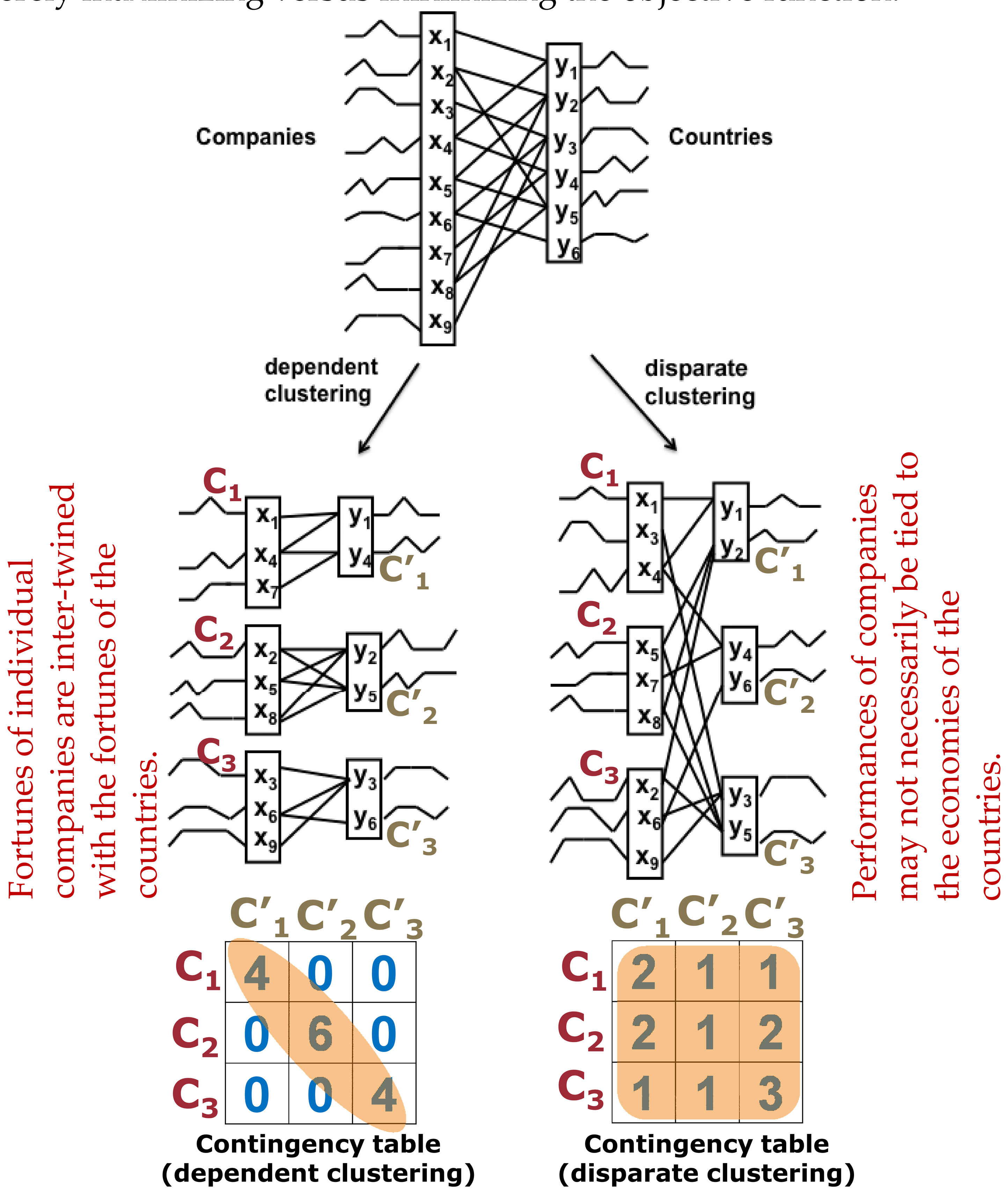
Unifying Dependent Clustering and Disparate Clustering for Non-homogeneous Data

M. Shahriar Hossain[1], Satish Tadepalli[1], Layne T. Watson[1], Ian Davidson[3], Richard F. Helm[2], Naren Ramakrishnan[1]
[1]Dept. of Computer Science, [2]Dept. of Biochemistry, Virginia Tech, VA 24061, USA
[3]Dept. of Computer Science, UC Davis, CA 95616, USA
Email: {msh, stadepal, ltw}@cs.vt.edu, davidson@cs.ucdavis.edu, helmrf@vt.edu, naren@cs.vt.edu

1. Introduction

Modern data mining settings involve a combination of attribute-valued descriptors over entities as well as specified relationships between these entities. We present an approach to cluster such non-homogeneous datasets by using the relationships to impose either dependent clustering or disparate clustering constraints. We present a formulation that allows constraints to be satisfied or violated in a smooth manner. This enables us to achieve dependent clustering and disparate clustering using the same optimization framework by merely maximizing versus minimizing the objective function.



2. Methodology

The only variables that we adjust during the optimization are the mean prototype vectors. But our optimization criteria are stated in terms of the resulting contingency tables:

Obj = *F* (Contingency table)
= *F* (*n*(Clustering1, Clustering2, Relation))
= *F* (*n*(*v*(Dataset1, Prototypes1), *v*(Dataset2, Prototypes2), Relation))

The formulations to prepare the objective function $\mathcal{F}$ are as follows:

Membership Probability
$$v_i^{(\mathbf{x}_s)} = \frac{\exp(-\frac{\rho}{D}||\mathbf{x}_s - \mathbf{m}_{i,\mathcal{X}}||^2)}{\sum_{i'=1}^{k_x}\exp(-\frac{\rho}{D}||\mathbf{x}_s - \mathbf{m}_{i',\mathcal{X}}||^2)}$$

Contingency Table Counts
$$w_{ij} = \sum_{s=1}^{n_x}\sum_{t=1}^{n_y} B(s,t) v_i^{(\mathbf{x}_s)} v_j^{(\mathbf{y}_t)}$$

Probability Distributions (row and column)
$$p(\alpha_i = j) = p(C_{(y)} = j | C_{(x)} = i) = \frac{w_{ij}}{w_{i.}}$$
$$p(\beta_j = i) = p(C_{(x)} = i | C_{(y)} = j) = \frac{w_{ij}}{w_{.j}}$$

Objective Function
$$\mathcal{F} = \frac{1}{k_x}\sum_{i=1}^{k_x} D_{KL}\Big(\alpha_i || U(\frac{1}{k_y})\Big) + \frac{1}{k_y}\sum_{j=1}^{k_y} D_{KL}\Big(\beta_j || U(\frac{1}{k_x})\Big)$$

An example is shown below with the **disparate clustering** of section 1:

w

	C′1	C′2	C′3
C1	2	1	1
C2	2	1	2
C3	1	1	3

U

	C′1	C′2	C′3
C1	0.33	0.33	0.33
C2	0.33	0.33	0.33
C3	0.33	0.33	0.33

α — *Row distributions*

	C′1	C′2	C′3
C1	0.5	0.25	0.25
C2	0.4	0.2	0.4
C3	0.2	0.2	0.6

β — *Col. distributions*

	C′1	C′2	C′3
C1	0.4	0.33	0.17
C2	0.4	0.33	0.33
C3	0.2	0.33	0.5

$\mathcal{F}$ = ⅓(0.03+0.023+0.07) + ⅓(0.023+0+0.041) = 0.062

2.1 Optimization

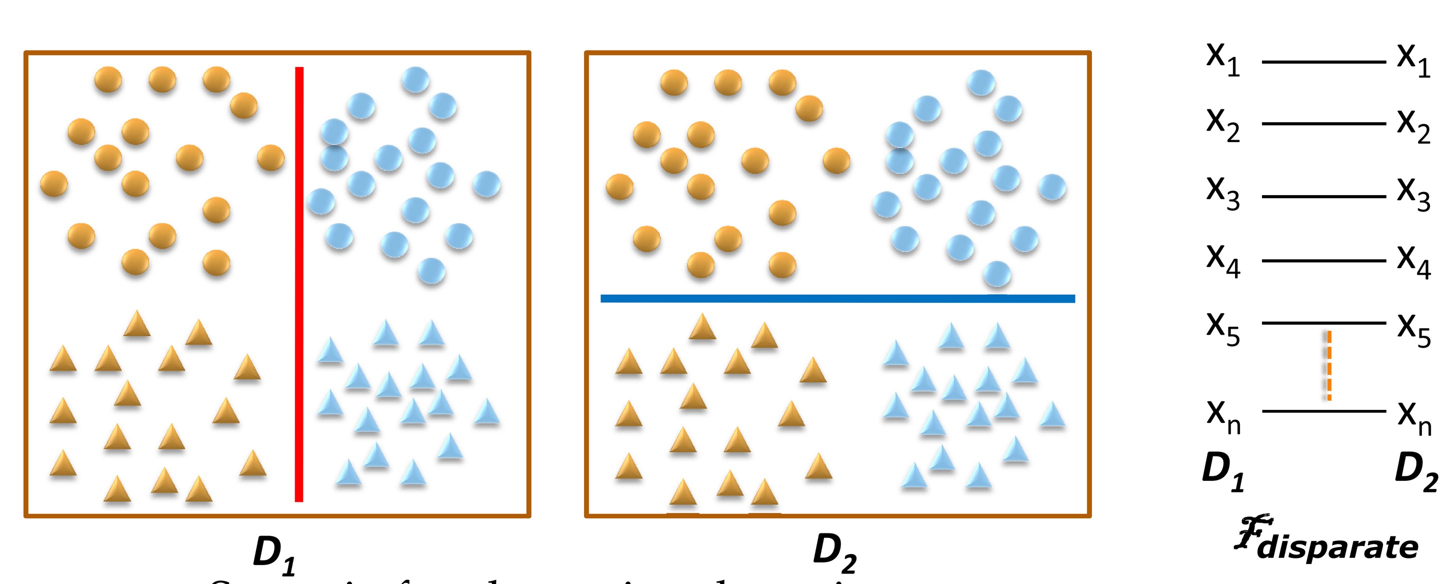
We minimize the objective function for disparate clustering and maximize it for dependent clustering. We use **Quasi-Newton Trust Region Algorithm** for optimization.

3. Experimental Results

We conduct our experiments both for single dataset scenarios and for non-homogeneous datasets.

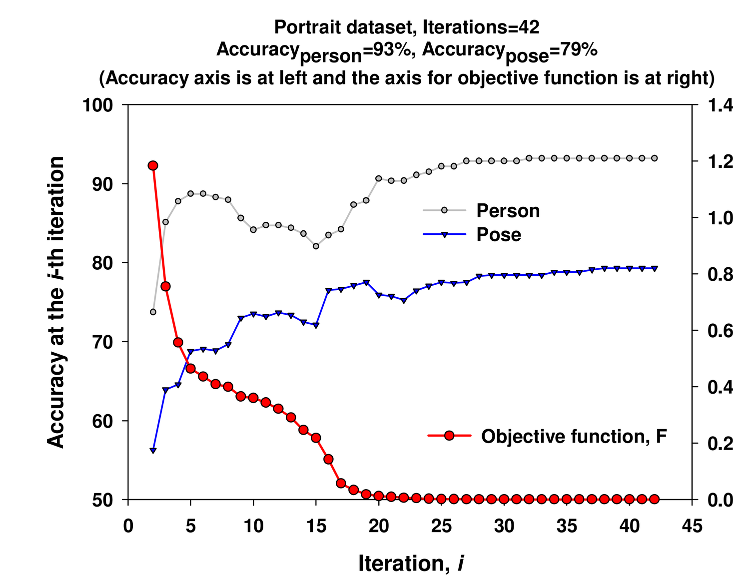
3.1 Alternative Clustering

We use the portrait dataset to conduct an experiment on alternative clustering. The portrait dataset has 324 images of three persons in three different poses (36 illuminations each) and contains 300 features.



Scenario for alternative clustering

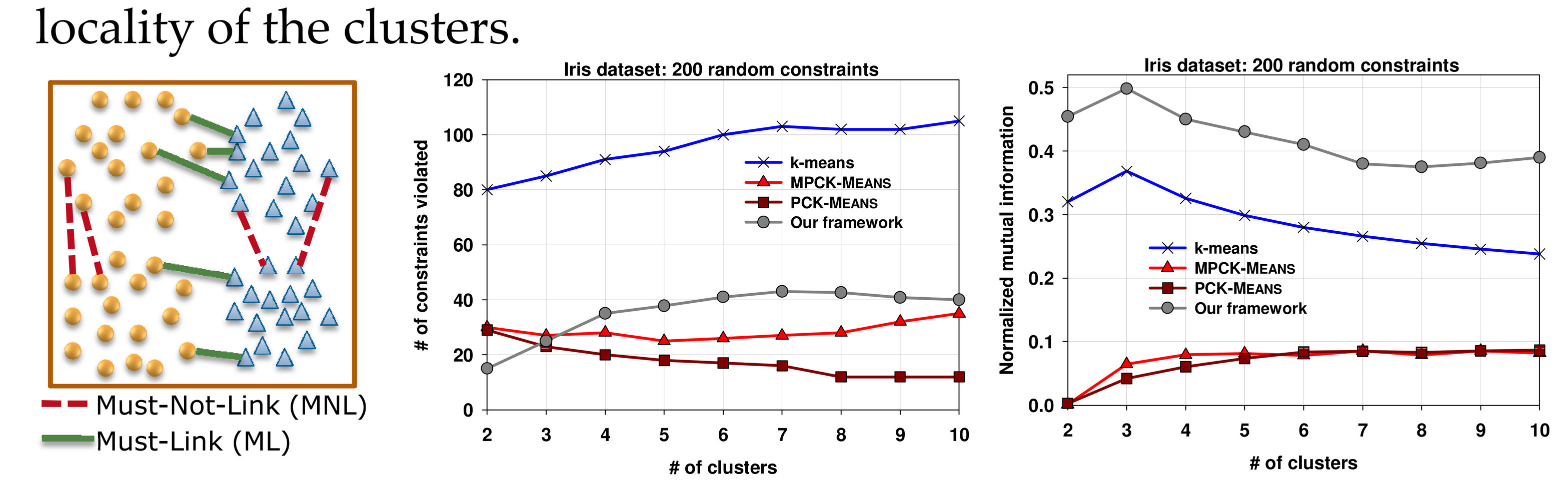
Exp. setup



	Person	Pose
k-means	0.65	0.55
Conv-EM	0.69	0.72
Dec-kmeans	0.84	0.78
Our framework	0.93	0.79

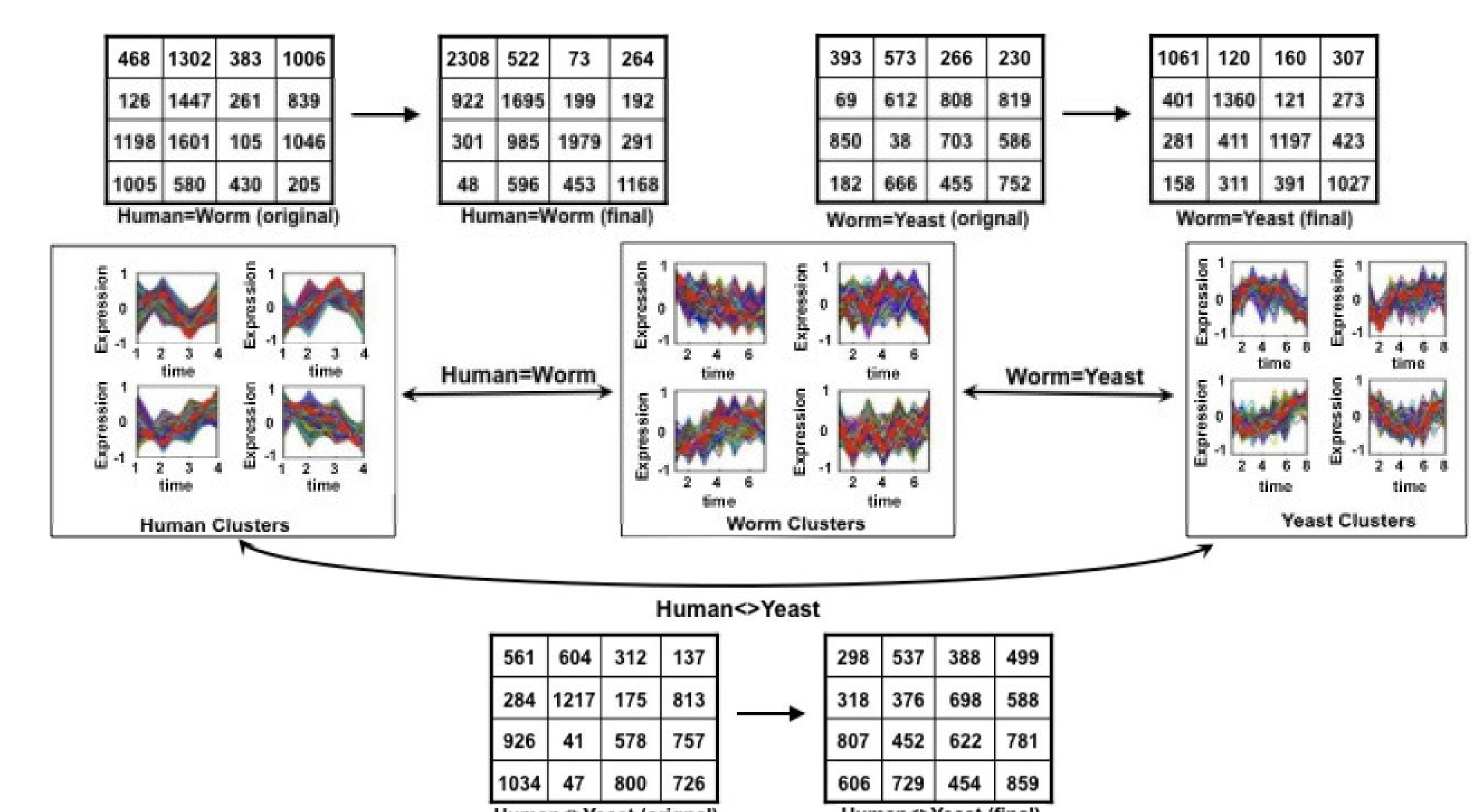
3.2 Constrained Clustering

For a given clustering and a set of instance-level constraints, our framework can perform constrained clustering while maintaining the locality of the clusters.



3.3 Comparing Gene Expression Programs

In this experiment, we use three non-homogeneous datasets of gene expression in human, worm, and yeast.



4. Discussion

1. The framework we developed is a general, expressive framework for clustering non-homogenous datasets.
2. The framework subsumes previously defined formulations.

Acknowledgements: This work is supported in part by US National Science Foundation grants CCF-0937133, CNS-0615181, and the Institute for Critical Technology and Applied Science (ICTAS) at Virginia Tech.