Combining Disparate Information by Nonmetric Multidimensional Scaling

Brent Castle
Indiana University

This is joint work with Michael Trosset (IU) and in collaboration with Carey E. Priebe, Youngser Park, and others (JHU).

This research was supported by grants from the Office of Naval Research.

Disparate Information Fusion

Disparate Information Fusion (DIF) is the combination of data of heterogeneous type and/or structure, e.g., text and images.

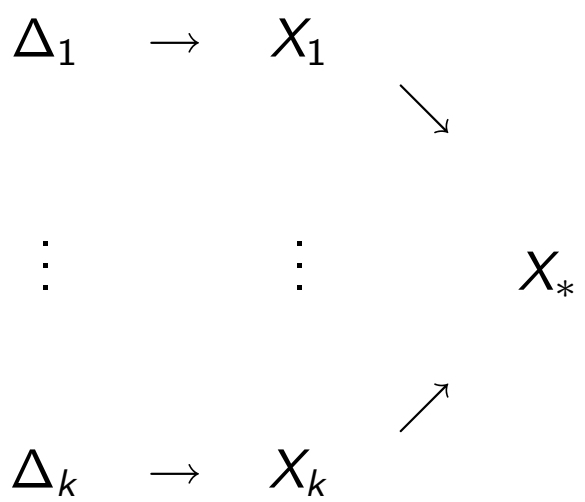
In text mining, image analysis, and many other disciplines, the ambient feature spaces are complicated and high-dimensional. Domain experts have developed specialized measures of (dis)similarity, e.g., for retrieval.

We assume that we have n objects and pairwise dissimilarities for each data type. Our goal is to construct a common representation of the n objects to be used for subsequent analysis. Our present concern is with classification.

Product Embedding Approach

Given: n objects and k nxn pairwise dissimilarity matrices $\Delta_1, \Delta_2, \ldots, \Delta_k$

Goal: construct a Euclidean representation of the objects



1. Construct a representation, X_i, for each of the k disparate measures.
2. Form the product, $X_* = [X_1|\ldots|X_k]$.
3. Apply standard multivariate methods for dimension reduction, classification, etc.

Multidimensional Scaling

Metric multidimensional scaling are techniques for embedding points in Euclidean space such that $d_{ij} \approx \delta_{ij}$.

- Uses actual values of the observed dissimilarities
- Scale equivariant

Nonmetric multidimensional scaling are techniques for embedding points in Euclidean space such that the interpoint distances are monotonically related to the dissimilarities.

- Uses only the rank order of the observed dissimilarities
- Scale invariant

Nonmetric Multidimensional Scaling

Kruskal (1964) formulated nonmetric multidimensional scaling to minimize the normalized stress criterion:

$$\sigma_n(X) = \sqrt{\frac{\sum_{i<j}(\hat{d}_{ij} - d_{ij}(X))^2}{\sum_{i<j} d_{ij}^2(X)}}$$

where d_{ij} is the distance between objects i and j and $\hat{d}_{ij}$ is the monotone regression of $d_{ij}(X)$ on the ranks of the observed dissimilarities.

Normalizing ensures scale invariance and precludes degenerate solutions.

Kruskal proposed minimizing σ_n by a gradient method.

Our Formulation

Our approach to nonmetric MDS is unconventional, motivated by a desire to embed large data sets.

We are less concerned with precise global minimizers than with plausible embeddings of large data sets. For each Δ_r, we seek small values of

$$\sigma(\Delta, X) = \sum_{i<j} [\delta_{ij} - d_{ij}(X)]^2$$

subject to

(1) $\Delta \in M(\Delta_r)$ [monotonicity constraint]

(2) $\sum_{i<j} \delta_{ij}^2 \geq \sum_{\ell=1}^{n(n-1)/2} \ell^2$ [nondegeneracy constraint]

Trosset (1998) proposed (2) as an alternative to normalization.

Our Algorithm

1. Construct an inexpensive initial embedding using the method of standards. We embed $d + 1$ anchor points by Classical MDS, then position each remaining point by solving a *dxd* linear system.
2. Decrease σ by several cycles of the following variable alternation scheme:
   a. Fix Δ and modify X by several iterations of a fixed point method, e.g., Guttman majorization or diagonal majorization (Trosset and Groenen 2005).
   b. Fix X and modify Δ by projecting $D(X)$ into the closed and convex set of feasible Δ. This is accomplished by projecting $D(X)$ into $M(\Delta_r)$ by isotonic regression (Grotzinger & Witzgall 1984), followed by rescaling the projected $D(X)$ to satisfy the nondegeneracy constraint (Lemma 2, Trosset 1998).

Two Classification Examples

Example 1: Classify images as one of four classes using two measures

Example 2: Classify yeast genes with functional labels given five measures

For each example we:

1. Construct a Euclidean representation as the product of the separate embeddings constructed by nonmetric MDS.
2. Perform linear discriminant analysis in product representation (or a subspace). Subspaces identified by various methods, e.g., variable selection (McHenry 1978), discriminant coordinates, etc.

Example 1 - Tiger Images

Given $n = 1520$ images that contain the word "tiger" in their caption.

The images range in size from 186×245 pixels to 406×450 pixels. Each pixel is represented as a vector of RGB values.

Each object is labeled by the following:

animal	148	golf	897
baseball	145	rebel	330

Example 1 - Color

1. Transform each image from RGB space to CIE L*a*b* space. [1]

2. For each image, compute pixel representatives $m_1, \ldots, m_8$ by a fast algorithm for k-means clustering. An image's color signature is the discrete measure that weights each m_i by the proportion of pixels associated with m_i.

3. Use $\|m_i - m_j\|_2$ to compute the earth mover's distance between pairs of color signatures. [2]

---

[1] McLaren (1976). The development of the CIE 1976 (L*a*b*) uniform colour-space and colour-difference formula. *Journal of the Society of Dyers and Colourists*, 92:338–341.

[2] Rubner, Tomasi, Guibas (1998). A metric for distributions with applications to image databases. *Proceedings of the IEEE International Conference on Computer Vision*, 59–66.

Example 1 - Texture

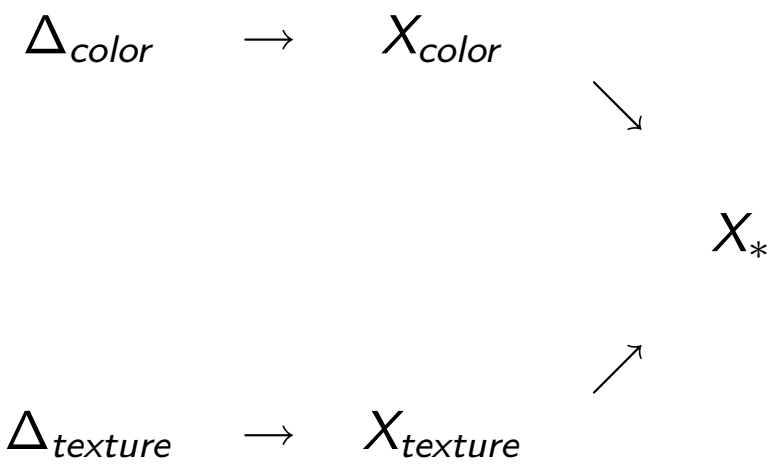
With minor deviations, we followed Chapter 5 of Rubner's 1999 Ph.D. dissertation, Perceptual Metrics for Image Database Navigation, Department of Computer Science, Stanford University.

1. Convolve each image with 24 Gabor filters (6 orientations $\times$ 4 scales), each with a fixed size of 13×13 pixels. This procedure associates a vector of 24 features with each pixel.

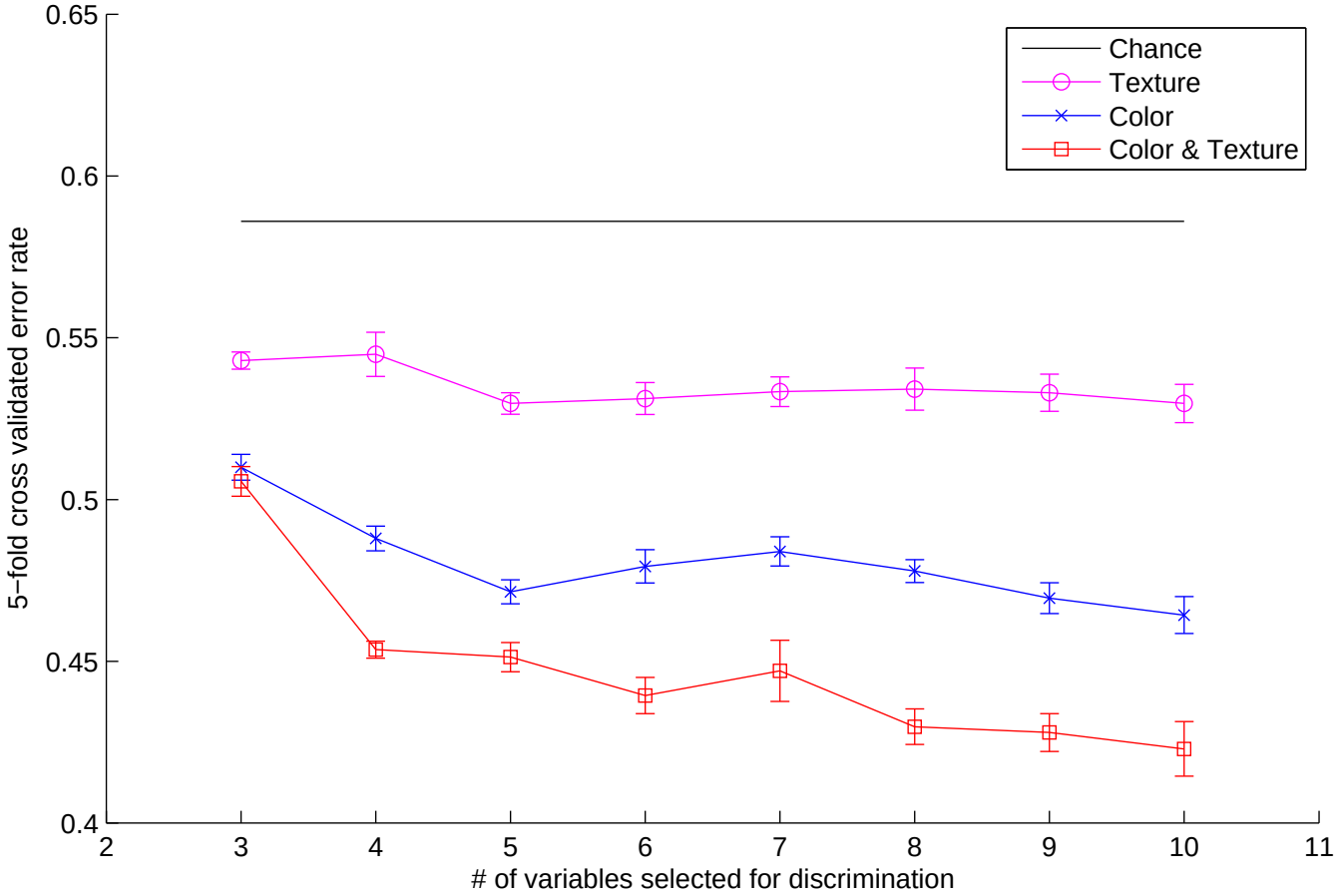
2. For each image, compute pixel representatives $m_1, \ldots, m_8$ by k-means clustering the pixel feature vectors. An image's texture signature is the discrete probability measure that weights each m_i by the proportion of pixels associated with m_i.

3. Use $\|m_i - m_j\|_1$ to compute the earth mover's distance between pairs of texture signatures.

Example 1 - Product Embedding



Example 1 - Results



Example 2 - Yeast gene function

Given $n = 3588$ yeast genes, each with 13 binary labels:

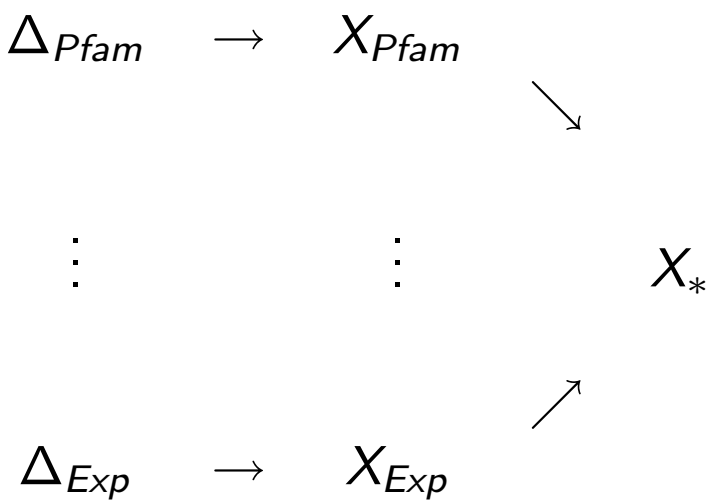
Metabolism	1048	Cell Rescue, Defense & Virulence	264
Energy	242	Interaction w/ cell. env.	193
Cell Cycle & DNA proc.	600	Cell Fate	411
Transcription	753	Control of cellular organization	192
Protein Synthesis	335	Transport Facilitation	306
Protein Fate	578	Others	81
Cellular Transportation	479		

Five measures were taken between each pair of genes:

1. Inner product of binary vectors (presence of Pfam domains) (Pfam)
2. Distance in a graph of genetic interaction information (GI)
3. Distance in a graph of protein-protein interaction information (PPI)
4. Distance in a graph of co-participation in a protein complex (TAP)
5. Dissimilarity measure between expression profiles (Exp)

Deng (2003) used a Markov Random Field to fuse the five measures and predict gene function. Lanckriet et al. (2004) formed an optimal linear combination of the five measures in a kernel representation.

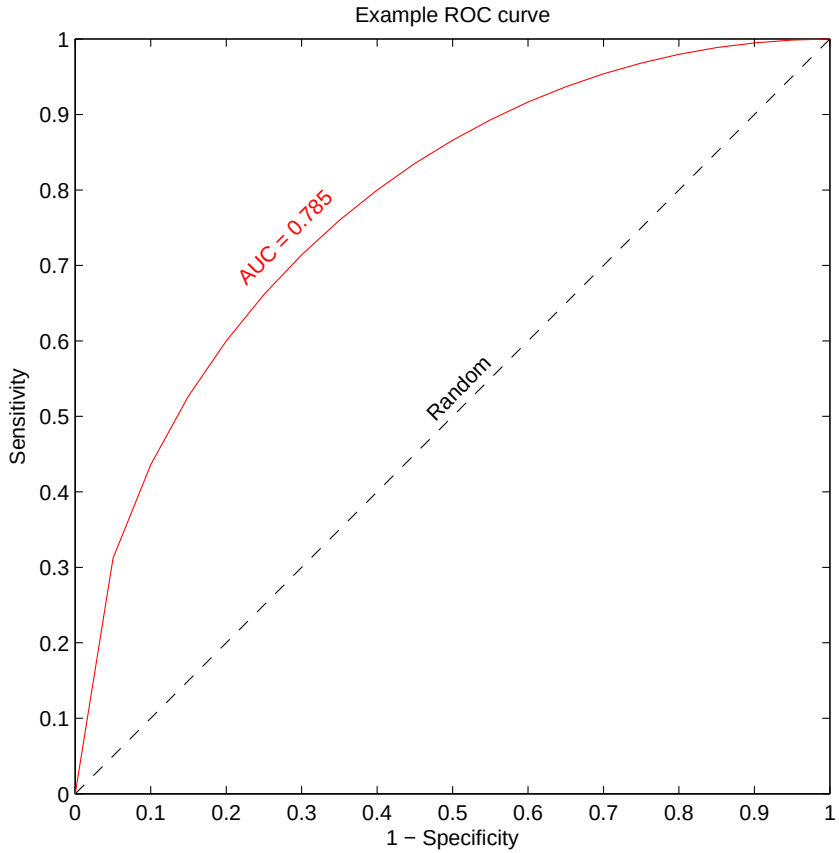
Example 2 - Product Embedding



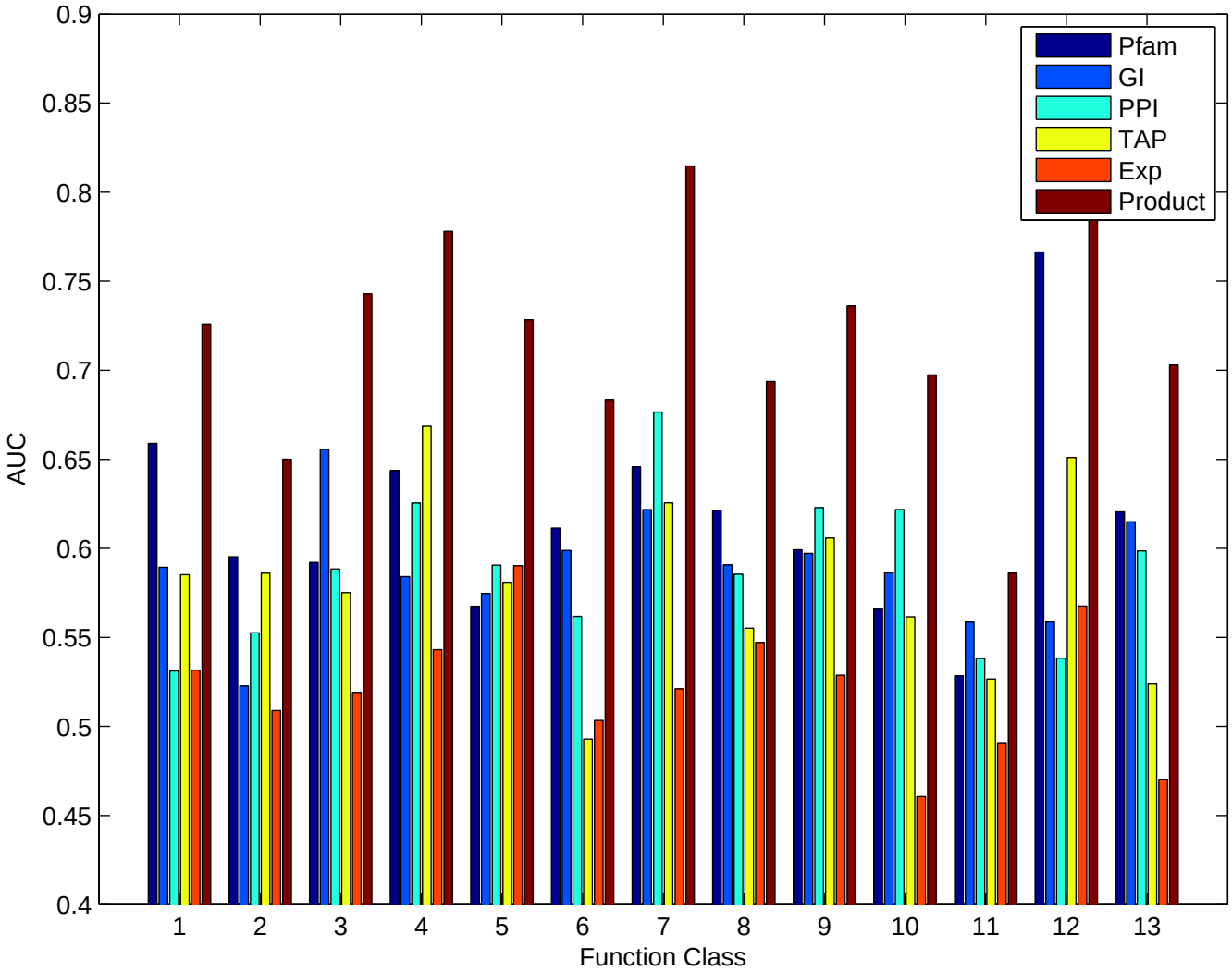
Example 2 - Results

Following the methods in Deng (2003) and Lanckriet (2004) we present the results as the area under the receiver operating characteristic (ROC) curve (AUC).

The ROC curve is a plot of the true positive rate (sensitivity) vs. the false positive rate (1-specificity) for various discrimination thresholds. The AUC statistic is a means of summarizing the curve in a single number.



Example 2 - Results



Summary

- Motivated the use of nonmetric MDS in the product embedding
- Formulated a scalable implementation of nonmetric MDS
- Results from image classification and gene function classification show the technique works on complex data

Thank you!