Combining Disparate Information by Nonmetric Multidimensional Scaling

Brent Castle and Michael Trosset
Indiana University

Introduction

- **Problem:** How can we embed objects in Euclidean space such that they can be classified, clustered, visualized, etc., given disparate types of measurements?
- Classification and clustering are well understood in Euclidean space.
- Embedding objects with disparate features is nontrivial.
- Examples of objects with disparate features:
  - Websites have content, images, hierarchical structure, design elements, ...
  - People have social networks, physical characteristics, levels of education ...
  - Proteins have sequence information, function, three dimensional structure ...
- Example types of (dis)similarity measurement:
  - Minkowski (city-block, Euclidean, ...), Mahalanobis, and Levenshtein distance
  - Relative graphlet frequency distribution between two networks
  - Jaccard's, Dice's, and Sørenson's similarity coefficient
  - Earth mover's distance between discrete measures

Methodology

- **Goal:** Given n objects and k disparate pairwise measurements, embed each object using nonmetric multidimensional scaling into a representation X.
- What is nonmetric multidimensional scaling (NMDS)?
  A method to embed objects in Euclidean space using only the rank order of the interpoint dissimilarities.
- Why use NMDS?
  By only using rank order of the interpoint dissimilarities we can safely ignore the inherent differences in the distributions of metric information between the disparate measurements.
- An example of our method in the case of classification:
  1. Compute the dissimilarity matrices $\Delta_1, ..., \Delta_k$ in each feature space $Y_1, ..., Y_k$
  2. Embed each object in Euclidean space $(X_1, ..., X_k)$ using NMDS
  3. Form product X such that $X = [X_1|...|X_k]$
  4. Select the subset of dimensions of X that best discriminate the objects in terms of linear discriminant analysis
- Alternative methods may include:
  - Constructing a classifier on the product of the input features
  - Constructing classifiers for each feature space and using these to construct an ensemble classifier
  - Constructing a classifier based on a combination of the dissimilarity measures, e.g., a linear combination such as $\Delta(t) = (1-t)\Delta_1 + t\Delta_2$.

Example

Original Objects

1. Extract Disparate Features

Color Signature

Texture Signature

13.19	3.12	6.03	...	1.03

1.34	0.19	2.87	...	18.65

2. Compute dissimilarities and embed separately

$$Y_{color} \rightarrow \Delta_{color} \rightarrow X_{color}$$
$$Y_{texture} \rightarrow \Delta_{texture} \rightarrow X_{texture}$$

3. Form product

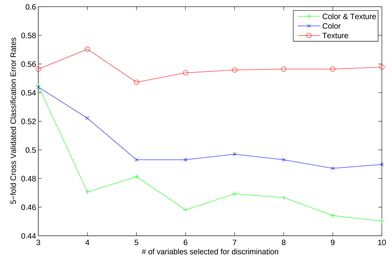
$$X = [X_{color}|X_{texture}]$$

4. Select best discriminating subset

↓ Color 1	Color 2	↓ Color 3	...	Texture 1	↓ Texture 2	...
-0.4326	1.0668	-0.6918	...	0.8156	-1.6041	...
-1.6656	0.0593	0.8580	...	0.7119	0.2573	...
1.1909	-1.3362	0.5711	...	-1.2025	0.5287	...
1.1892	0.7143	-0.3999	...	-0.0198	0.2193	...
-0.0376	1.6236	0.6900	...	-0.1567	-0.9219	...

Example continued ...

- 1520 images with captions containing the word "tiger"
- 4 class labels were applied to the images: animal, golf, baseball, and rebels.
- **Goal:** Construct a classifer to apply a label to each image by using the combination of two measures of dissimilarity. The goal is to combine the data in such a way that it performs better than using the disparate types of data independently.
- **Method:** We used color and texture features to separately embed the images using NMDS. We then formed the product of these embeddings and used McHenry's Best Subset selection method to keep the best discriminating dimensions. We then performed linear discriminant analysis on the combined embedding.



- **Result:** The combination of the color and texture via nonmetric multidimensional scaling outperformed the use of color and texture alone.

Conclusion

There seems to be evidence that the fusion of disparate information is possible via nonmetric multidimensional scaling. Our future work includes theoretical support for our method, testing in several domains, developing a means to quantify how disparate the measurements are, as well as more methods for fusing the data.

Future method for exploration: Nonmetric three-way scaling which uses all k dissimilarity matrices to construct a single Euclidean representation X.

References

McHenry, C.E. *Computation of a Best Subset in Multivariate Analysis.* Applied Statistics 27, No. 3, pp. 291-296. (1978).

Rubner, Y., Tomasi, C., and Guibas, L.J. *The earth mover's distance as a metric for image retrieval.* International Journal of Computer Vision, 40(2):99-121, (November 2000).

Trosset, M.W. *A new formulation of the nonmetric STRAIN problem in multidimensional scaling.* Journal of Classification 15, pp. 15-35. (1993).

Research funded by the Office of Naval Research