Embedding Heterogeneous Data using Statistical Models

Amir Globerson[1] **Gal Chechik**[2] **Fernando Pereira**[3] **Naftali Tishby**[4]
[1] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge MA, 02139
[2] Computer Science Department, Stanford University, Stanford, CA 94305, USA
[3] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA
[4] School of computer Science and Engineering and the Interdisciplinary Center for Neural Computation
The Hebrew University Jerusalem, 91904, Israel

Abstract

Embedding algorithms are a method for revealing low dimensional structure in complex data. Most embedding algorithms are designed to handle objects of a single type for which pairwise distances are specified. Here we describe a method for embedding objects of different types (such as authors and terms) into a single common Euclidean space based on their co-occurrence statistics. The joint distributions of the heterogenous objects are modeled as exponentials of squared Euclidean distances in a low-dimensional embedding space. This construction links the problem to convex optimization over positive semidefinite matrices. We quantify the performance of our method on two text datasets, and show that it consistently and significantly outperforms standard methods of statistical correspondence modeling, such as multidimensional scaling and correspondence analysis.

Introduction

Embeddings of objects in a low-dimensional space are an important tool in unsupervised learning and in preprocessing data for supervised learning algorithms. They are especially valuable for exploratory data analysis and visualization by providing easily interpretable representations of the relationships among objects. Most current embedding techniques build low dimensional mappings that preserve certain relationships among objects and differ in the relationships they choose to preserve, which range from pairwise distances in multidimensional scaling (MDS) (Cox & Cox 1984) to neighborhood structure in locally linear embedding (Roweis & Saul 2000). All these methods operate on objects of a single type endowed with a measure of similarity or dissimilarity.

However, real-world data often involve objects of several very different types without a natural measure of similarity. For example, typical web pages or scientific papers contain several different data types such as text, diagrams, images, and equations. A measure of similarity between words and pictures is difficult to define objectively. Defining a useful measure of similarity is even difficult for some homogeneous data types, such as pictures or sounds, where the physical properties (pitch and frequency in sounds, color and luminosity distribution in images) do not directly reflect the semantic properties we are interested in.

The current paper addresses this problem by creating embeddings from statistical associations. The idea is to find a Euclidean embedding in low dimension that represents the empirical co-occurrence statistics of two variables. Here we focus on modeling the conditional probability of one variable given the other, since in the data we analyze (documents and words, authors and terms) there is a clear asymmetry which suggests a conditional model. Joint models can be constructed similarly , and may be more appropriate for symmetric data. We name our method CODE for *Co-Occurrence Data Embedding*.

Our cognitive notions are often built through statistical associations between different information sources. Here we assume that those associations can be represented in a low-dimensional space. For example, pictures which frequently appear with a given text are expected to have some common, locally low-dimensional characteristic that allows them to be mapped to adjacent points. We can thus rely on co-occurrences to embed different entity types, such as words and pictures, genes and expression arrays, into the same subspace. Once this embedding is achieved it also naturally defines a measure of similarity between entities of the same kind (such as images), induced by their other corresponding modality (such as text), providing a meaningful similarity measure between images.

Embedding of heterogeneous objects is performed in statistics using *correspondence analysis* (CA), a variant of *canonical correlation analysis* for count data (Greenacre 1984). These are related to Euclidean distances when the embeddings are constrained to be normalized. However, as we show below, removing this constraint has great benefits for real data. Statistical embedding of same-type objects was recently studied in (Hinton & Roweis 2002). Their approach is similar to ours in that it assumes that distances induce probabilistic relations between objects. However, we do not assume that distances are given in advance, but instead we derive them from the empirical co-occurrence data.

Problem Formulation

Let X and Y be two categorical variables with an empirical distribution $\bar{p}(x, y)$. No additional assumptions are made on the values of X and Y or their relationships. We wish to

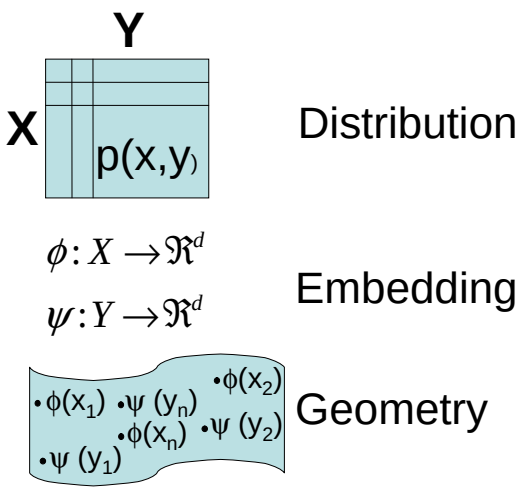

Figure 1: Embedding of X, Y into the same d-dimensional space.

model the statistical dependence between X and Y through an intermediate Euclidean space $\mathbb{R}^d$ and mappings $\vec{\phi} : X \to \mathbb{R}^d$ and $\vec{\psi} : Y \to \mathbb{R}^d$. These mappings should reflect the dependence between X and Y in the sense that the distance between each $\vec{\phi}(x)$ and $\vec{\psi}(y)$ determines their co-occurrence statistics.

We focus in this work on modeling the conditional distribution $p(y|x)$[1], and define a model which relates conditional probabilities to distances by

$$p(y|x) = \frac{\bar{p}(y)}{Z(x)} e^{-d^2_{x,y}} \quad \forall x \in X, \forall y \in Y \tag{1}$$

where $d^2_{x,y} \equiv \|\vec{\phi}(x) - \vec{\psi}(y)\|^2 = \sum_{k=1}^{d} (\phi_k(x) - \psi_k(y))^2$ is the Euclidean distance between $\vec{\phi}(x)$ and $\vec{\psi}(y)$ and $Z(x)$ is the partition function for each value of x. This partition function equals $Z(x) = \sum_y \bar{p}(y) e^{-d^2_{x,y}}$ and is thus the empirical mean of the exponentiated square distances from x (therefore $Z(x) \leq 1$).

This model directly relates the ratio $\frac{p(y|x)}{\bar{p}(y)}$ to the distance between the embedded x and y. The ratio decays exponentially with the distance. Thus for any x, a closer y will have a higher interaction ratio. As a result of the fast decay, the closest objects dominate the distribution. The model of Eq. 1 can also be described as the result of a random walk in the low-dimensional space illustrated in Figure 1. When y has a uniform marginal, the probability $p(y|x)$ corresponds to a random walk from x to y, with transition probability inversely related to distance.

We now turn to the task of learning $\vec{\phi}, \vec{\psi}$ from an empirical distribution $\bar{p}(x, y)$. It is natural in this case to maximize the likelihood (up to constants depending on $\bar{p}(y)$)

$$\max_{\vec{\phi},\vec{\psi}} l(\vec{\phi}, \vec{\psi}) = -\sum_{x,y} \bar{p}(x, y) d^2_{x,y} - \sum_x \bar{p}(x) \log Z(x) \ , \tag{2}$$

[1]We have studied several other models of the joint rather than the conditional distribution. These differ by the way the marginals are modeled and will be described elsewhere

where $\bar{p}(x, y)$ denotes the empirical distribution over X, Y. The likelihood is composed of two terms. The first is (minus) the mean squared distance between x and y. This will be maximized when all distances are zero. This trivial solution is avoided because of the *regularization* term $\sum_x \bar{p}(x) \log Z(x)$, which acts to increase distances between x and y points.

To find the optimal $\vec{\phi}, \vec{\psi}$ for a given embedding dimension d, we use a conjugate gradient ascent algorithm with random restarts. In the "Semidefinite Representation" section we describe a different approach to this optimization problem.

Relation to Other Methods

Embedding the rows and columns of a contingency table into a low dimensional Euclidean space is related to statistical methods for the analysis of heterogeneous data. (Fisher 1940) described a method for mapping X and Y into $\phi(x), \psi(y)$ such that the correlation coefficient between $\phi(x), \psi(y)$ is maximized. His method is in fact the discrete analogue of the *Canonical correlation analysis* (CCA) method (Hotelling 1935) . Another closely related method is *Correspondence analysis* (Greenacre 1984), which uses a different normalization scheme, and aims to model χ^2 distances between rows and columns of $\bar{p}(x, y)$.

The goal of all the above methods is to maximize the correlation coefficient between the embeddings of X and Y. We now discuss their relation to our *distance* based method. It is easy to show that CCA minimizes the following function:

$$\rho(\phi(x), \psi(y)) = -\frac{1}{2} \sum_{x,y} \bar{p}(x, y) d^2_{x,y} + 1$$

under the conditions that $\phi(x), \psi(y)$ have zero mean and identity covariance. Maximizing the correlation is therefore equivalent to minimizing the mean squared distance across all pairs. Thus, both CCA and CODE aim to minimize the average distance between X and Y, but while CCA forces both embeddings to be centered and with a unity covariance matrix, CODE introduces a regularization term related to the partition function.

A well-known geometric oriented embedding method is multidimensional scaling (MDS) (Cox & Cox 1984), whose standard version applies to same-type objects with predefined distances. MDS embedding of heterogeneous entities was studied in the context of modeling ranking data (see (Cox & Cox 1984) section 7.3). These models, however, focus on specific properties of ordinal data and therefore result in optimization principles different from our probabilistic interpretation.

Semidefinite Representation

The optimal embeddings $\vec{\phi}, \vec{\psi}$ may be found using unconstrained optimization techniques. However, the Euclidean distances used in the embedding space also allow us to reformulate the problem as constrained convex optimization over the cone of positive semidefinite (PSD) matrices (Weinberger & Saul 2004).

We start by showing that for embeddings with dimension $d = |X| + |Y|$, maximizing (2) is equivalent to minimizing a certain convex non-linear function over PSD matrices. Consider the matrix A whose columns are all the embedded vectors $\vec{\phi}$ and $\vec{\psi}$. The matrix $G \equiv A^T A$ is the Gram matrix of the dot products between embedding vectors. It is thus a symmetric PSD matrix of rank $\leq d$. The converse is also true: any PSD matrix of rank $\leq d$ can be factorized as $A^T A$, where A is an embedding matrix of dimension d. The distance between two columns in A is linearly related to the Gram matrix via $d_{ij}^2 = g_{ii} + g_{jj} - 2g_{ij}$.

Since the likelihood function depends only on the distances between points in X and in Y, we can write the optimization problem in (2) as

$$\begin{array}{ll} \min & \sum_x \bar{p}(x) \log \sum_y \bar{p}(y) e^{-d_{xy}^2} + \sum_{x,y} \bar{p}(x,y) d_{xy}^2 \\ \text{s.t.} & G \succeq 0 \quad , \quad \text{rank}(G) \leq d \end{array} \tag{3}$$

When $\text{rank}(G) = d$ the above problem can be shown to be convex. Thus there are no local minima, and solutions can be found efficiently.

The PSD formulation also allows us to add non-trivial constraints. Consider, for example, constraining the $p(y)$ marginal to its empirical values, i.e. $\sum_x p(y|x)\bar{p}(x) = \bar{p}(y)$. In (Globerson *et al.* 2005) we show how to use the above formulation for solving such constraints, which are not easily incorporated into maximum likelihood optimization. Another interesting model which can be solved this way is $p(x,y) \propto p(x)p(y)\exp(-d_{xy}^2)$ where $p(x), p(y)$ are the marginals of $p(x,y)$ (e.g., $\sum_y p(x,y) = p(x)$).

Embedding into a low dimension requires constraining the rank, but this is difficult since the problem is no longer convex in the general case. Here we penalize high-rank solutions by adding the trace of G weighted by a positive factor, λ, to the objective function in (3). Small values of $\text{Tr}(G)$ are expected to correspond to sparse eigenvalue sets and thus penalize high rank solutions. This approach was tested on subsets of the databases described in the Applications section and yielded similar results to those of the gradient based algorithm. We believe that PSD algorithms may turn out to be more efficient in cases where relatively high dimensional embeddings are sought. Furthermore, under the PSD formulation it is easy to introduce additional constraints, for example on distances between subsets of points (as in (Weinberger & Saul 2004)), and on marginals of the distribution.

Applications

We tested our approach on a variety of applications. Here we present embedding of words and documents. To provide quantitative assessment of the performance of our method, that goes beyond visual inspection, we apply it to problems where some underlying structures are known in advance. The known structures are only used for performance measurement and not during learning.

NIPS Database

To illustrate the use of CODE for studying document databases, we applied it to the NIPS 0-17 database (Chechik 2005). The last three volumes also contain an indicator of the document's topic (AA for algorithms and architecture, LT for learning theory, NS for neuroscience etc.). We used CODE to embed documents and words into $\mathbb{R}^2$. The results are shown in Figure 2. It can be seen that documents with similar topics are mapped next to each other (e.g. AA near LT and NS near Biological Vision). Furthermore, words characterize the topics of their neighboring documents.

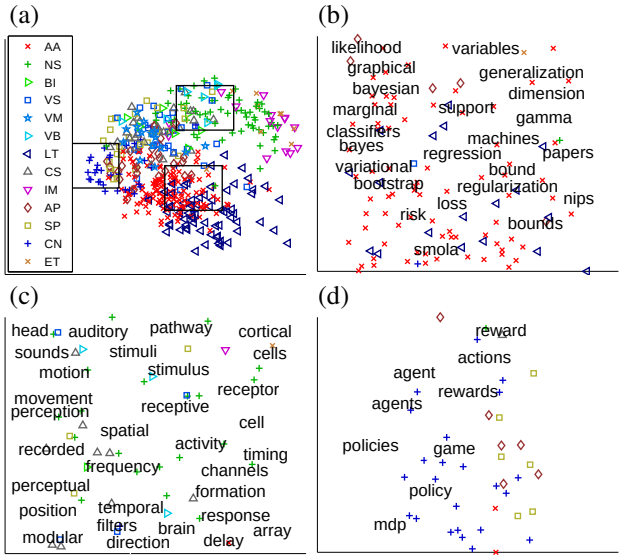

Figure 2: CODE Embedding of 2483 documents and 2000 words from the NIPS database. The left panel shows document embeddings for NIPS 15-17, with colors to indicate the document topic. Other panels show embedded words and documents for the areas specified by rectangles. Figure (b) shows the border region between algorithms and architecture (AA) and learning theory (LT) (bottom rectangle in (a)). Figure (c) shows the border region between neuroscience (NS) and biological vision (VB) (upper rectangle in (a)). Figure (d) shows mainly control (CN) documents (left rectangle in (a)).

We also used the data to generate an authors-words matrix. We could now embed authors and words into $\mathbb{R}^2$, by using CODE to model $p(word|author)$. The results are given in (Globerson *et al.* 2005) and show that authors are indeed mapped next to terms relevant to their work, and that authors dealing with similar domains are also mapped together.

Taken together, these results illustrate how co-occurrence of words and authors/documents may be used to obtain meaningful geometric maps of the data.

Information Retrieval

To obtain a more quantitative estimate of performance, we applied CODE to the 20 newsgroups corpus, preprocessed as described in (Chechik & Tishby 2002). The resulting words and documents were embedded with CODE, Correspondence Analysis (CA), SVD, IsoMap and classical MDS [2]. CODE was used to model the distribution of words given documents $p(word|doc)$. All methods were tested under several normalization schemes, including document sum

[2] See (Globerson *et al.* 2005) for evaluation details.

normalization and TFIDF. Results were consistent across all normalization schemes.

An embedding of words and documents is expected to map documents with similar semantics together, *and* to map words close to documents which are related to the meaning of the word. We next test how our embeddings performs with respect to these requirements. To represent the *meaning* of a document we use its corresponding newsgroup. Note that this information is used only for evaluation and not in constructing the embedding itself.

To measure how well similar documents are mapped together we devised a purity measure, which we denote *doc-doc*. Briefly, *doc-doc* measures if documents with similar topics appear as nearest neighbors of each other. To measure how documents are related to their neighboring words, we devised a measure denoted by *doc-word*, which quantifies how likely it is for a document to be mapped near a word which characterizes its topic.

Results for both measures are given in (Globerson *et al.* 2005) and illustrate that CODE is significantly better than all other methods on both measures, implying that it achieves a better geometric model of both same and different object relations.

Discussion

We presented a method for embedding objects of different types into the same low dimension Euclidean space. This embedding can be used to reveal low dimensional structures when distance measures between objects are unknown. Furthermore, the embedding induces a meaningful metric also between objects of the same type, which could be used, for example, to embed images based on accompanying text, and derive the semantic distance between images.

Co-occurrence embedding should not be restricted to pairs of variables, but can be extended to multivariate joint distributions. For example Markov Random Fields may be extended such that interaction potentials are modeled as Euclidean distances. One implementation of this idea would be to model interaction potentials between variables in a given clique as the mean distance between their embeddings. Another option is to model the potential as the mean distance of embedded points from their centroid.

Another interesting extension would be to use distances between same-type objects when these are known. The semidefinite formulation could prove useful in that respect, since it allows incoporation of equality constraints.

We focused here on geometric models for *conditional* distributions. While in some cases, such a modeling choice is more natural in others joint models may be more appropriate. In this context it will be interesting to consider models of the form $p(x, y) \propto p(x)p(y)e^{-d^2_{x,y}}$ where $p(x), p(y)$ are the marginals of $p(x, y)$. Maximum likelihood in these models is a non-trivial constrained optimization problem, and may be approached using the semidefinite representation outlined above.

References

Chechik, G., and Tishby, N. 2002. Extracting relevant structures with side information. In Becker, S.; Thrun, S.; and Obermayer, K., eds., *NIPS 15*.

Chechik, G. 2005. The NIPS 0-17 database. http://robotics.stanford.edu/∼gal/.

Cox, T., and Cox, M. 1984. *Multidimensional Scaling*. London: Chapman and Hall.

Fisher, R. 1940. The percision of discriminant functions. *Ann. Eugen. Lond.* 10:422–429.

Globerson, A.; Chechik, G.; Pereira, F.; and N.Tishby. 2005. Euclidean embedding of co-occurrence data. In Saul, L. K.; Weiss, Y.; and Bottou, L., eds., *NIPS 17*.

Greenacre, M. 1984. *Theory and applications of correspondence analysis*. Academic Press.

Hinton, G., and Roweis, S. 2002. Stochastic neighbor embedding. In *NIPS 15*.

Hotelling, H. 1935. The most predictable criterion. *Journal of Educational Psych.* 26:139–142.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323–2326.

Weinberger, K., and Saul, L. 2004. Unsupervised learning of image manifolds by semidefinite programming. In *CVPR*.