Appl. Comput. Harmon. Anal. • • • (• • • •) • • •–• • •

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Case Studies

Bigeometric organization of deep nets

Alexander Cloninger [a,*], Ronald R. Coifman [a], Nicholas Downing [b], Harlan M. Krumholz [b]

[a] *Applied Mathematics Program, Yale University, United States*
[b] *Center for Outcomes Research and Evaluation, Yale University, United States*

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form 16 April 2016
Accepted 2 August 2016
Available online xxxx
Communicated by Charles K. Chui

Keywords:
Diffusion embedding
Deep learning
Intrinsic organization
Hospital quality

ABSTRACT

In this paper, we build an organization of high-dimensional datasets that cannot be cleanly embedded into a low-dimensional representation due to missing entries and a subset of the features being irrelevant to modeling functions of interest. Our algorithm begins by defining coarse neighborhoods of the points and defining an expected empirical function value on these neighborhoods. We then generate new non-linear features with deep net representations tuned to model the approximate function, and re-organize the geometry of the points with respect to the new representation. Finally, the points are locally z-scored to create an intrinsic geometric organization which is independent of the parameters of the deep net, a geometry designed to assure smoothness with respect to the empirical function. We examine this approach on data from the Center for Medicare and Medicaid Services Hospital Quality Initiative, and generate an intrinsic low-dimensional organization of the hospitals that is smooth with respect to an expert driven function of quality.

1. Introduction

Finding low dimensional embeddings of high dimensional data is vital in understanding the organization of unsupervised data sets. However, most embedding techniques rely on the assumption that the data set is locally Euclidean [7,14,1]. In the case that features carry implicit weighting, some features are possibly irrelevant, and most points are missing some subset of the features, Euclidean neighborhoods can become spurious and lead to poor low dimensional representations.

In this paper, we develop the method of expert driven functional discovery to deal with the issue of spurious neighborhoods in data sets with high dimensional contrasting features. This allows small amounts of input and ranking from experts to propagate through the data set in a non-linear, smooth fashion. We then build a distance metric based off these opinions that learns the invariant and irrelevant features from

* Corresponding author.
E-mail address: alexander.cloninger@yale.edu (A. Cloninger).

http://dx.doi.org/10.1016/j.acha.2016.08.001
1063-5203/

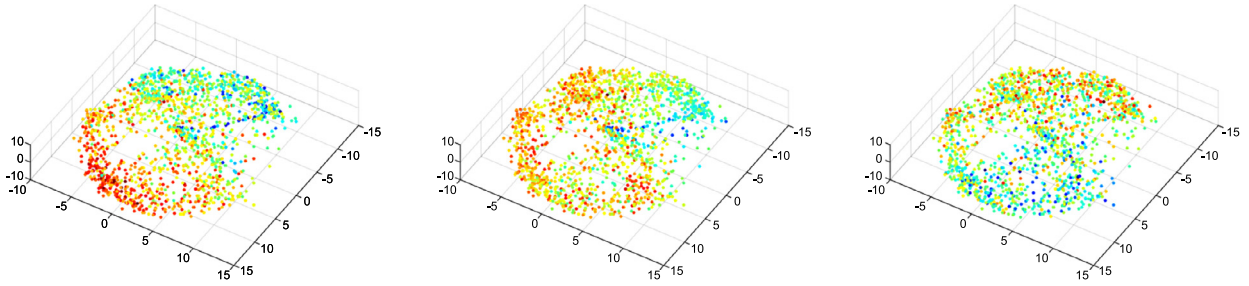

Fig. 1. Organization colored by: (left) risk standardized 30 day hospital wide readmission, (center) percent patients rating overall hospital 9 or 10 out of 10, (right) risk standardized 30 day mortality for heart failure. Embedding generated via bigeometric organization of deep nets. Red is good performance, blue is bad. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

this expert driven function. Finally, we locally normalize this distance metric to generate a global embedding of the data into a homogeneous space.

An example to keep in mind throughout the paper, an idea we expand upon in Section 4, is a data set containing publicly-reported measurements of hospital quality. The Center for Medicare and Medicaid Services Hospital Quality Initiative reports approximately 100 different measures describing various components of the quality of care provided at Medicare-Certified hospitals across the United States. These features range in measuring hospital processes, patient experience, safety, rates of surgical complications, and rates of various types of readmission and mortality. There are more than 5,000 hospitals that reported at least on measure during 2014, but only 1,614 hospitals with 90% measures reported. The measures are computed quarterly, and are publicly available through the Hospital Compare website [8]. The high dimensional nature of these varied measures make comprehensive inferences about hospital quality impossible without summarizing statistics.

Discovering the topology of the hospitals is non-trivial. The features may have significant disagreement, and not be strongly correlated across the population. To examine these relationships, one can consider linear correlations via principal component analysis. The eigenvalues of the correlation matrix do not show the characteristic drop off shown in linear low dimensional data sets. In fact, 76 of the 86 eigenvalues are above 1% the size of the largest eigenvalue. Previous medical literature has also detailed the fact that many of the features don't always correlate [9,4].

For this reason, there does not exist an organization for which all features are smooth and monotonically increasing. This is why the meta-features, and organization, must be driven by minimal external expert opinion. This observation makes the goal of our approach three fold: develop an organization of the data that is smooth with respect to as many features as possible, build a ranking function f that agrees with this organization, and minimize the amount of external input necessary to drive the system.

Our goal is more than just learning a ranking function f on the set of hospitals X. We are trying to characterize the cohort of hospitals and organize the geometry of the data set, and learn a multi-dimensional embedding of the data for which the ranking function is smooth. This gives an understanding of the data that doesn't exist with a one dimensional ranking function. Specifically, we are looking for meta-features of the data in order to build a metric $\rho : X \times X \to \mathbb{R}^+$ that induces a small Lipschitz constant on the function f, as well as on features measured by CMS.

An example of this organization is shown in Fig. 1. The organization is generated via our algorithm of expert driven functional discovery, the details of which are found in Sections 2 and 3. The colors in each image correspond to three notable CMS features: risk standardized 30 day hospital-wide readmission, patient overall rating of the hospital, and risk standardized 30 day mortality for heart failure. This organization successfully separates hospitals into regimes for which each feature is relatively smooth.

Our organization is accomplished via a three step processing of the data. First, we build an initial organization of the data via coupled partition trees [6], and use this partitioning to generate pseudopoints

Please cite this article in press as: A. Cloninger et al., Bigeometric organization of deep nets, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.08.001