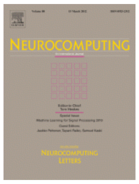
Archetypal analysis for machine learning and data mining

E sciencedirect.com/science/article/abs/pii/S0925231211006060



,

Abstract

Archetypal analysis (aa) proposed by Cutler and Breiman (1994) [7] estimates the *principal convex hull* (pch) of a data set. As such aa favors features that constitute representative ‘corners’ of the data, i.e., distinct aspects or archetypes. We currently show that aa enjoys the interpretability of clustering – without being limited to hard assignment and the uniqueness of svd – without being limited to orthogonal representations. In order to do large scale aa, we derive an efficient algorithm based on projected gradient as well as an initialization procedure we denote FurthestSum that is inspired by the FurthestFirst approach widely used for k-means (Hochbaum and Shmoys, 1985 [14]). We generalize the aa procedure to kernel-aa in order to extract the principal convex hull in potential infinite Hilbert spaces and derive a relaxation of aa when the archetypes cannot be represented as convex combinations of the observed data. We further demonstrate that the aa model is relevant for feature extraction and dimensionality reduction for a large variety of machine learning problems taken from computer vision, neuroimaging, chemistry, text mining and collaborative filtering leading to highly interpretable representations of the dynamics in the data. Matlab code for the derived algorithms is available for download from www.mortenmorup.dk↗.

Introduction

Decomposition approaches have become a key tool for the analysis of a wide variety of massive data from modeling the Internet, such as term-document matrices of word occurrences, bio-informatics data such as micro-array data of gene expressions, neuroimaging data such as neural activity measured over space and time to collaborative filtering such as the celebrated Netflix problem to mention but a few. The conventional approaches range from low rank approximations such as singular value decomposition (svd), principal component analysis (pca) [4], independent component analysis (ica) [6], sparse coding (sc) [24] also denoted dictionary learning [8] and non-negative matrix factorization (nmf) [18] and convex nmf [10] to soft clustering approaches such as fuzzy k-means [3] and the EM-algorithm for clustering [4] to hard assignment clustering methods such as k-means

and k-medoids. Common to these approaches is that they can be understood as a linear mixture or factor analysis type representation of data with various constraints. Thus data $x_{m,n}$, where $m=1,\ldots,M$ is the feature index and ·· $n=1,\cdot\cdot,N$ is the sample index, is written in terms of hidden variables $s_{d,n}$ and projections $a_{m,d}$ with $d=1,\ldots,D$ where *D* denotes the number of factors, typically with a Gaussian noise model, i.e.,
$x_{m,n}=\sum_d a_{m,d}s_{d,n}+e_{m,n}, e_{m,n}\sim N(0,\sigma^2)$.

svd/pca requires A and S be orthogonal, in ica statistical independence is assumed for S and in sc a penalty term is introduced to promote sparsity of S, while in nmf all variables are constrained non-negative. In soft clustering S is also non-negative but the columns constrained to sum to one whereas in hard clustering by k-means S is constrained to be a binary assignment matrix such that $A=XS^\top(SS^\top)^{-1}$ represents the Euclidean centers of each cluster while for k-medoids $a_d=x_n$ for some *n*, i.e., the cluster centers are actual data points.

Despite the similarities of the above approaches their internal representations of the data differ greatly and thus the nature of the data analyses they offer. In svd/pca the features constitute the directions of maximal variation, i.e., so-called eigenmaps, for nmf the features are constituent parts, while k-means and k-medoids find the most representative prototype objects.

A benefit of clustering approaches is that features are similar to measured data making the results easier to interpret, however, the binary assignments reduce flexibility. Also, clustering typically involves complex combinatorial optimization leading to a plethora of heuristics. On the other hand low rank approximations based on svd/pca/nmf have a great degree of flexibility but the features can be harder to interpret. Invariance to rotation of the extracted features can lead to lack of uniqueness, i.e.,
$X\approx AS=AQQ^{-1}S=\tilde{A}\tilde{S}$. In addition, svd/pca/ica/sc are prone to cancelation effects in which two components both lose meaning because they locally become highly correlated taking positive and negative near-canceling values (while still being globally orthogonal).

In conclusion, clustering approaches give easy interpretable features but pay a price in terms of modeling flexibility due to the binary assignment of data objects. Approaches such as svd/pca/ica/nmf/sc have added model flexibility and as such can be more efficient in capturing, e.g., variance, however, this efficiency can lead to complex representations from which we learn relatively little.

Archetypal analysis (aa) proposed by Cutler and Breiman [7] directly combines the virtues of clustering and the flexibility of matrix factorization. In the original paper on aa [7] the method was demonstrated useful in the analysis of air pollution and head shape and later also for tracking spatio-temporal dynamics [27]. Recently, archetypal analysis has found use in benchmarking and market research identifying typically extreme practices, rather than just good practices [26] as well as in the analysis of astronomy spectra [5] as an approach for the end-member extraction problem [25].

In this paper we demonstrate the following important theoretical properties of aa

- •
  The archetypal analysis representation is unique in the sense that a solution to aa in general does not suffer from rotational ambiguity.
- •
  Archetypal analysis can be initialized efficiently through the proposed FurthestSum method.
- •
  Archetypal analysis can be efficiently computed using a simple projected gradient method.
- •
  Archetypal analysis can be generalized to kernel-aa in order to extract the principal convex hull in a Hilbert space.

We further demonstrate that aa is useful for a wide variety of important machine learning and data mining problem domains resulting in easy interpretable features that well account for the inherent dynamics in data. This paper is an extended version of Mørup and Hansen [23]. Apart from elaborating more on the aspects of aa given above we have included an additional application within chemistry as well as a relaxation of aa that can address the issue when “pure” archetypes cannot be represented by convex combinations of the observed data. We further provide Matlab code for the derived algorithms available for download from www.mortenmorup.dk.

Section snippets

Archetypal analysis and the principal convex hull

The convex hull or envelope of a data matrix X is the minimal convex set containing X. Informally it can be described as a rubber band wrapped around the data points, see also Fig. 1. While the problem of finding the convex hull is solvable in linear time (i.e., $O(N)$) [21] the size of the convex set increases dramatically with the dimensionality of the data. The expected size of the convex set for *N* points in general position in *M* dimensional space grows exponentially with dimension as $O(\log(N)M$

Data mining and machine learning applications

We demonstrate the usefulness of the aa model on five data sets taken from a variety of important machine learning problem domains.

Discussion

We demonstrated how the archetypal analysis model of Cutler and Breiman [7] is useful for a broad variety of machine learning problems. A simple algorithm for fitting the aa/pch model was derived as well as the FurthestSum initialization procedure to extract end-members for initial archetypes. The utility of aa/pch over clustering methods is that it focuses more on distinct or discriminative aspects yet has additional modeling flexibility by using soft assignment. We saw examples of improved

Acknowledgement

We thank Gitte Moos Knudsen and Claus Svarer for providing the PET data set. This work is supported in part by the Danish Lundbeck Foundation through CIMBI, Center for Integrated Molecular Brain Imaging.

Morten Mørup received his Ph.D. degree in Applied Mathematics at the Technical University of Denmark in 2008. He is currently an Assistant Professor at the Section for Cognitive Systems at the Technical University of Denmark and a Current Member of the Technical Committee for the IEEE International Workshops on Machine Learning for Signal Processing (MLSP). His research interest is machine learning and biomedical data analysis and a major focus of his research has been dedicated to unsupervised

Cited by (158)

- **Recovering Gene Interactions from Single-Cell Data Using Data Diffusion**

  2018, Cell

- **Tumour heterogeneity and the evolutionary trade-offs of cancer**

  2020, Nature Reviews Cancer

- **Evolutionary highways to persistent bacterial infection**

  2019, Nature Communications

- **An Introduction to Systems Biology: Design Principles of Biological Circuits, Second Edition**

  2019, An Introduction to Systems Biology: Design Principles of Biological Circuits, Second Edition

- **Perspective: Sloppiness and emergent theories in physics, biology, and beyond↗**

  2015, Journal of Chemical Physics

- **Inferring biological tasks using Pareto analysis of high-dimensional data↗**

  2015, Nature Methods

> View all citing articles on Scopus↗

Morten Mørup received his Ph.D. degree in Applied Mathematics at the Technical University of Denmark in 2008. He is currently an Assistant Professor at the Section for Cognitive Systems at the Technical University of Denmark and a Current Member of the Technical Committee for the IEEE International Workshops on Machine Learning for Signal Processing (MLSP). His research interest is machine learning and biomedical data analysis and a major focus of his research has been dedicated to unsupervised learning with application to the analysis of neuroimaging data such as EEG and fMRI.

Lars Kai Hansen is a Professor of Digital Signal Processing at the Technical University of Denmark, Lyngby, where he also heads the Cognitive Systems Section. His research concerns adaptive signal processing and machine learning with applications in biomedicine and digital media. He has published more than 225 contributions on these subjects in journals, conferences, and books.

View full text