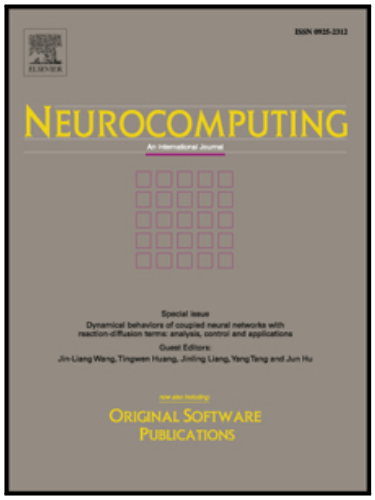
Communicated by Dr. Chenping Hou

Accepted Manuscript



Robust Data Representation using Locally Linear Embedding Guided PCA

Bo Jiang, Chris Ding, Bin Luo

PII: S0925-2312(17)31483-2
DOI: 10.1016/j.neucom.2017.08.053
Reference: NEUCOM 18854

To appear in: *Neurocomputing*

Received date: 8 January 2017
Revised date: 31 July 2017
Accepted date: 29 August 2017

Please cite this article as: Bo Jiang, Chris Ding, Bin Luo, Robust Data Representation using Locally Linear Embedding Guided PCA, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.08.053

Robust Data Representation using Locally Linear Embedding Guided PCA

Bo Jiang[a], Chris Ding[b,a], and Bin Luo[a]
[a]School of Computer Science and Technology, Anhui University, Hefei, China
[b]CSE Department, University of Texas at Arlington, Arlington, TX 76019, USA
jiangbo@ahu.edu.cn, chqding@uta.edu, luobin@ahu.edu.cn

Abstract

Locally Linear Embedding (LLE) is widely used for embedding data on a nonlinear manifold. It aims to preserve the local neighborhood structure on the data manifold. Our work begins with a new observation that LLE has a natural robustness property. Motivated by this observation, we propose to integrate LLE and PCA into a LLE guided PCA model (LLE-PCA) that incorporates both global structure and local neighborhood structure simultaneously while performs robustly to outliers. LLE-PCA has a compact closed-form solution and can be efficiently computed. Extensive experiments on five datasets show promising results on data reconstruction and improvement on data clustering and semi-supervised learning tasks.

Keywords: Principal component analysis, LLE, Semi-supervised learning, Dimension reduction

1. Introduction

Efficient and compact representation of data is a fundamental problem in data mining and machine learning area. In real-world datasets, data usually have high dimensionality and often contain noises and errors. One of the widely used effective approaches is to find a low-dimensional representation for high dimension data based on low-rank matrix factorization [1, 2, 3, 4, 5, 6, 7]. In low-dimensional space, the cluster distribution becomes more apparent and thus can improve the machine learning results [8, 9, 10, 11]. In additional to matrix factorization, trace norm based low-rank and sparse methods have also been widely used to conduct data representation