Accepted Manuscript

Adaptive non-negative projective semi-supervised learning for inductive classification

Zhao Zhang, Lei Jia, Mingbo Zhao, Qiaolin Ye, Min Zhang, Meng Wang

PII: S0893-6080(18)30216-8
DOI: https://doi.org/10.1016/j.neunet.2018.07.017
Reference: NN 4000

To appear in: *Neural Networks*

Received date : 27 December 2016
Revised date : 19 March 2018
Accepted date : 25 July 2018

Please cite this article as: Zhang, Z., Jia, L., Zhao, M., Ye, Q., Zhang, M., Wang, M., et al., Adaptive non-negative projective semi-supervised learning for inductive classification. *Neural Networks* (2018), https://doi.org/10.1016/j.neunet.2018.07.017

Adaptive Non-Negative Projective Semi-Supervised Learning for Inductive Classification

Zhao Zhang[#,]*, Lei Jia[#], Mingbo Zhao[ᶴ], Qiaolin Ye[☆, §], Min Zhang [#], and Meng Wang[※]
[#] School of Computer Science and Technology & Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China
[☆]Department of Computer Science and Technology, Nanjing Forestry University, Nanjing 210037, China
[ᶴ] Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong
[§]Department of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
※School of Artificial Intelligence & School of Computer and Information, Hefei University of Technology, Hefei, China
*Correspondence author: E-mail: cszzhang@gmail.com

Abstract— We discuss the inductive classification problem by proposing a joint framework termed *Adaptive Non-negative Projective Semi-Supervised Learning* (ANP-SSL). Specifically, ANP-SSL integrates the adaptive inductive label propagation, adaptive reconstruction weights learning and the neighborhood preserving projective nonnegative matrix factorization (PNMF) explicitly. To make the label prediction results more accurate, ANP-SSL incorporates the semi-supervised data representation and classification errors into regular PNMF for minimization, which can enable our ANP-SSL to perform the adaptive weights learning and label propagation over the spatially local and part-based data representations, which differs from most existing work that usually assign weights and predict labels based on the original data that often has noise and corruptions. Moreover, existing methods usually pre-assign weights before the process of label estimation, but such operation cannot ensure the learnt weights by independent step to be optimal for the subsequent classification. The combined representation error can also make the learnt reduced part-based representations of neighborhood preserving PNMF, which can potentially enhance the prediction results. By minimizing the classification error jointly over the neighborhood preserving nonnegative representation can make the embedding based classification efficient. Extensive results on several public image databases verified the effectiveness of our ANP-SSL, compared with other state-of-the-art methods.

***Index Terms*— Adaptive projective semi-supervised learning; Inductive label propagation; non-negative matrix factorization; representation and classification**

I. INTRODUCTION

Representing and classifying real data by performing semi-supervised learning (SSL) has been an important topic in the fields of neural networks and data mining [1-4][10][31-34][49-58]. SSL methods can obtain knowledge and valuable information from the setting equipped with a small number of labeled data and a large amount of unlabeled data, which well suits the characteristics of real application data, since most real data are unlabeled, huge in volume and also difficult to distinguish in practical applications. Towards handling these issues, semi-supervised data representation and classification methods can effectively deal with the unlabeled data by fully using class formation of the small number of labeled data and building the connections between labeled and unlabeled data via the pairwise similarity measure. It is worth noticing that the SSL methods have been widely and successfully used to various real applications, e.g., data classification, clustering, regression analysis and information retrieval, etc.

In recent years, *Label Propagation* (LP) [5][7][11][15][21-22][27], as a representative SSL classification algorithm, has aroused considerable attention and interests in academia due to its elegant formulation in efficiency and effectiveness, and the requirements in real applications. The learning process of LP is to propagate label information of the labeled data to the unlabeled data based on trading-off the manifold smoothness and label fitness [5][7][11][15]. That is, label formation of each sample is partly received from its initial state encoded by the label fitness term and is partly from its neighborhoods encoded by the manifold smoothness term. Based on whether outside new data can be involved efficiently, existing models of LP can be roughly divided into inductive and transductive settings. Tho transductive learning methods aim to estimate the unknown labels of inside unlabeled data, but they cannot predict the unknown labels of outside unlabeled data. Several representative transductive LP learning algorithms consist of *SSL using Gaussian Fields and Harmonic Functions* (GFHF) [9], *Learning with Local and Global Consistency* (LLGC) [16], *Linear Neighborhood Propagation* (LNP) [5], *Special Label Propagation* (SLP) [28], *Projective Label Propagation* (ProjLP) [38], *Class Dissimilarity based LNP* (CD-LNP) [18], *Robust Linear Neighborhood Propagation* (R-LNP) [39], and *Sparse Neighborhood Propagation* (SparseNP) [37], etc. It is worth noting that several researchers have also incorporated the idea of semi-supervised label propagation learning into the *Non-Negative Matrix Factorization* (NMF) [6] and the *Projective NMF* (PNMF) frameworks [13], termed *Semi-Supervised NMF* (SSNMF) [47] and *Semi-Supervised PNMF* (Semi-PNMF) [48]. Note that both Semi-PNMF and SSNMF are also the transductive methods as aforementioned models. To enable transductive methods to handle outside new data, the authors of LNP [5] have suggested to involve new data by reconstructing the label of each new data using the predicted soft labels of its neighbors from the training set, which is not straightforward, and time-consuming especially for large-scale testing set [23], because this approach needs to find the neighbors of each new test data by performing the nearest-neighbor search firstly. To handle the out-of-sample problem efficiently, several extensions by direct embedding have been