International Journal of Approximate Reasoning 74 (2016) 121–132

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Deep kernel dimensionality reduction for scalable data integration [☆]

Nataliya Sokolovska [a,b,c,*], Karine Clément [a,b,c], Jean-Daniel Zucker [a,b,d]

[a] *Institute of Cardiometabolism and Nutrition, ICAN, Assistance Publique Hôpitaux de Paris, Pitié-Salpêtrière Hospital, Paris, France*
[b] *Sorbonne Universités, UPMC University Paris 6, UMR_S 1166, ICAN, NutriOmics Team, Paris, France*
[c] *INSERM, UMR S U1166, NutriOmics Team, Paris, France*
[d] *Research Institute for Development, UMI 209, UMMISCO, Bondy, France*

ARTICLE INFO

Article history:
Received 10 October 2015
Received in revised form 13 March 2016
Accepted 22 March 2016
Available online 7 April 2016

Keywords:
Dimensionality reduction
Heterogeneous data integration

ABSTRACT

Dimensionality reduction is used to preserve significant properties of data in a low-dimensional space. In particular, data representation in a lower dimension is needed in applications, where information comes from multiple high dimensional sources. Data integration, however, is a challenge in itself.

In this contribution, we consider a general framework to perform dimensionality reduction taking into account that data are heterogeneous. We propose a novel approach, called Deep Kernel Dimensionality Reduction which is designed for learning layers of new compact data representations simultaneously. The method can be also used to learn shared representations between modalities. We show by experiments on standard and on real large-scale biomedical data sets that the proposed method embeds data in a new compact meaningful representation, and leads to a lower classification error compared to the state-of-the-art methods.

1. Introduction

Data integration is a challenging task with an ambitious goal to increase performance of supervised learning, since various sources of data tend to contain different parts of information about the problem.

Structure learning and data integration allow to better understand the properties and content of biological data in general and of “omics” data in particular. Combining complementary pieces issued from different data sources is likely to provide more knowledge, since distinct types of data provide distinct views of the molecular machinery of cells. Medical and biological knowledge can be naturally organized into hierarchies: symptoms of diseases are observed and pathological states on all levels of omics data are hidden. Hierarchical structures and data integration methods reveal dependencies that exist between cellular components and help to understand the biological network structure.

Graphical models follow a natural organization and representation of data, and are a promising method of simultaneous heterogeneous data processing. Hidden variables in a graphical hierarchical model can efficiently agglomerate information of observed instances via dimensionality reduction, since fewer latent variables are able to summarize multiple features. However, integration of latent variables is a crucial step of modeling.

[☆] This paper is part of the virtual special issue on Statistical and Computational methods for genomics and integrative genomics, Edited by Paola Rancoita and Cassio Campos.

[*] Corresponding author.
E-mail address: nataliya.sokolovska@upmc.fr (N. Sokolovska).

http://dx.doi.org/10.1016/j.ijar.2016.03.008
0888-613X/© 2016 Elsevier Inc. All rights reserved.

Multi-modal learning, heterogeneous data fusion, or data integration, involves relating information of different nature. In biological and medical applications, data coming from one source are already high-dimensional. Hence, data integration increases the dimensionality of a problem even more, and some feature selection or dimensionality reduction procedure is absolutely needed both to make the computations tractable and to obtain a model which is compact and easily interpretable.

Our goal is to develop an efficient dimensionality reduction approach which will design a compact model. The method needs to be scalable, to fuse heterogeneous data, and be able to reach a better generalizing performance compared to a full model and to state-of-the-art methods. Another important question is whether introducing data of different nature has a positive effect, and provides additional knowledge.

In this contribution, we propose a deep dimensionality reduction approach which agglomerates original features from a high-dimensional space and creates a hierarchy of new representations. To construct the hidden layers of the proposed deep learning framework, we introduce a deep kernel dimensionality reduction method, and we compare its performance to some standard clustering and dimensionality reduction methods.

The biomedical problem of our interest is a real problem which is a binary classification of obese patients. The aim is to stratify patients in order to choose an efficient appropriate personalized medical treatment. The task is motivated by a recent French study [1] of gene-environment interactions carried out to understand the development of obesity. It was reported that the gut microbial gene richness can influence the outcome of a dietary intervention. A quantitative metagenomic analysis stratified patients into two groups: group with low gene gut flora count (LGC) and high gene gut flora count (HGC) group. The LGC individuals have a higher insulin-resistance and low-grade inflammation, and therefore the gene richness is strongly associated with obesity-driven diseases. The individuals from a low gene count group seemed to have an increased risk to develop obesity-related cardiometabolic risk compared to the patients from the high gene count group. It was shown [1] that a particular diet is able to increase the gene richness: an increase of genes was observed with the LGC patients after a 6-weeks energy-restricted diet. A similar study with Dutch individuals was conducted by [2], and made a similar conclusion: there is a hope that a diet can be used to induce a permanent change of gut flora, and that treatment should be phenotype-specific. There is therefore a need to go deeper into these biomedical results and to identify candidate biomarkers associated with cardiometabolic disease (CMD) risk factors and with different stages of CMD evolution.

Our contribution is multi-fold:

- we introduce a novel kernel-based deep dimensionality reduction method which constructs layers of a deep structure simultaneously,
- we illustrate that the proposed framework is efficient on standard data sets and on a real original rich heterogeneous MicrObese data set [1], which contains meta-data, i.e., clinical parameters and alimentary patterns of patients, gene expressions of adipose tissue, and gene abundance of gut flora. We efficiently learn new data representations structured into a multi-level hierarchy. We evaluate the prediction power of the models with the reduced dimensionality showing that the proposed approach outperforms the state-of-the-art dimensionality reduction methods.

The paper is organized as follows. Section 2 considers the related work and the state-of-the-art data integration and dimensionality reduction methods. We introduce our approach in Section 3. We show the results of our experiments in Sections 4 and 5. Concluding remarks and perspectives close the paper.

2. Related work

We tackle a complex problem which consists of a data integration task and a dimensionality reduction procedure. In this section, we consider some state-of-the-art data fusion methods, dimensionality reduction approaches, and some recent attempts to combine both within a hierarchical model. The literature on clustering and dimensionality reduction is quite rich; publications on heterogeneous data integration, on the contrary, are not so numerous.

The state-of-the-art data integration methods are traditionally divided into four categories: functional linkage networks, vector subspace integration, kernel fusion methods, and ensemble methods. Graphical models (functional linkage networks) are based on graphical representation of nodes and relations between variables of interest, e.g., Bayesian networks. Vector space integration is a method where data from various sources are concatenated in a vector. Kernel methods for data integration are motivated by the fact that variables with similar functions share expression patterns. Kernel functions are used to define similarities between the variables of interest. Recently Ref. [3] reported that ensemble methods, that have been ignored for a long time, are a competitive data integration approach. Ensemble methods combine outputs produced by different classifiers trained on various data sets, or data views; they are known to be scalable, and data of different formats can be easily integrated, since the data integration is done at the decision level. Our framework, introduced in the next section, is a graphical framework and incorporates a vector concatenation of heterogeneous data, a similarity matrix base on a kernel function, and can also embed an ensemble method.

The idea to use a hierarchy for biomedical data is not new. So, Bayesian networks are still often used in systems biology. They model a joint probability distribution, parameterized by a parameter θ over all nodes. More specifically, the Bayesian networks define a joint probability distribution $P(x;\theta) = \prod_{i=1}^{n} P(x_i|x_{PA_i})$, where PA stands for "parent". E.g., Ref. [4] considers a linear model, where observed and hidden variables follow $x_i = \sum_{j\in PA_i} \alpha_{ij} h_j + \epsilon_i$, where x are observed and h are hidden. To estimate the vector of parameters α of the model, the hidden variables are integrated out. The problem that is