Neural Networks 22 (2009) 766–773

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2009 Special Issue

Predictive learning with structured (grouped) data

Lichen Liang, Feng Cai *, Vladimir Cherkassky

Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA

ARTICLE INFO

Article history:
Received 7 May 2009
Received in revised form 11 June 2009
Accepted 25 June 2009

Keywords:
Heterogeneous data
Learning with structured data
Model selection
Multi-task learning
SVM
SVM-Plus

ABSTRACT

Many applications of machine learning involve sparse and heterogeneous data. For example, estimation of diagnostic models using patients' data from clinical studies requires effective integration of genetic, clinical and demographic data. Typically all heterogeneous inputs are properly encoded and mapped onto a single feature vector, used for estimating a classifier. This approach, known as standard inductive learning, is used in most application studies. Recently, several new learning methodologies have emerged. For instance, when training data can be naturally separated into several groups (or structured), we can view model estimation for each group as a separate task, leading to a Multi-Task Learning framework. Similarly, a setting where the training data are structured, but the objective is to estimate a single predictive model (for all groups), leads to the Learning with Structured Data and SVM+ methodology recently proposed by Vapnik [(2006). *Empirical inference science afterword of* 2006. Springer]. This paper describes a biomedical application of these new data modeling approaches for modeling heterogeneous data using several medical data sets. The characteristics of group variables are analyzed. Our comparisons demonstrate the advantages and limitations of these new approaches, relative to standard inductive SVM classifiers.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction and motivation

Statistical data-driven computer-aided diagnostics have been of growing interest in biomedical applications. Such approaches usually estimate diagnostic models from available (historical) data. Whereas machine learning and statistical approaches often pursue similar goals and use similar techniques, there is a key difference in perspective (Cherkassky & Mulier, 2007). Under predictive learning, the main goal of modeling is good prediction (generalization) for future data. In contrast, statisticians view the probability model as the core of the analysis, with the idea that optimal predictions will arise from this probability model accurately estimated from data. Sometimes machine learning algorithms correspond to statistical models (e.g., mixture models), but other times the predictions feel more like they are coming from 'black boxes' with less statistical interpretation. This distinction is often known as generative (∼statistical) versus discriminative (∼ predictive) modeling. For multivariate sparse data sets common in biomedical applications, the predictive approach is more practical because

(a) there are simply not enough available data samples to estimate the multivariate distributions (this is known as the curse of dimensionality); and

(b) it may be possible to estimate accurate predictive models that reflect *certain properties* of unknown distributions (Cherkassky & Mulier, 2007; Vapnik, 1998, 2006). For example, for classification problems, the goal of estimating a decision boundary (for future predictions) does not require accurate estimation of class distributions. Moreover, Statistical Learning Theory (also known as VC theory) (Vapnik, 1998, 2006, 1982) gives mathematical conditions under which good prediction (generalization) is possible with finite samples, *regardless of dimensionality* (the number of input variables).

The price paid for adopting the predictive approach is that the estimated models may *accurately predict*, but only in a specific well-defined sense (known as 'generalization'). This places an additional burden on a data modeler, who needs to come up with a *meaningful formalization* of an application domain at hand. In particular, this approach requires *close collaboration* between data modelers and clinicians (who provide the data and will use data-driven predictive models). It also implies that medical researchers/clinicians should understand better conceptual aspects of predictive learning. Another important difference is that predictive models may not be easily interpretable, because they do not approximate 'true' distributions, but rather imitate certain properties of unknown distributions.

Future advances in the area of data-driven biomedical applications are limited by two fundamental factors: (a) high dimensionality of the input data (i.e., large number of input variables) and (b) heterogeneous nature of the input data. *High-dimensional, low*

* Corresponding author. Tel.: +1 612 219 3562
E-mail address: caixx043@umn.edu (F. Cai).

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.06.030

sample size (HDLSS) data are common in many biomedical applications, especially studies involving genetic data. For example, a 'typical' clinical study may result in a data set of a few hundred to a couple of thousand patients ('samples'), where each patient has a few hundred genetic predictors (for instance, ∼400 genetic polymorphisms), in addition to a few dozen clinical and demographic inputs. All these heterogeneous inputs may be used as possible predictors for diagnosing a disease or predicting the outcome of a medical treatment procedure.

For such data sets, the dimensionality d of the data vector may be larger than/similar to the sample size n. Such sparse training data sets present new challenges to classification methods that estimate classification decision boundaries from HDLSS data. Note that commonly used discriminative methods (such as neural networks and support vector machines) require significant modifications and/or clever preprocessing in dealing with HDLSS data. *Heterogeneous data* in biomedical applications may include clinical, genomic and demographic data used as input variables for constructing a predictive (diagnostic) model. These inputs can be viewed as several feature sets, and the challenge is to integrate such input data from different modalities into learning with sparse high-dimensional data. There are two principal approaches for dealing with HDLSS and heterogeneous data (Cherkassky & Mulier, 2007).

The *first approach* is to adopt a *standard inductive learning* setting, and to reduce the problem dimensionality via clever preprocessing and feature extraction. That is, the problem of high-dimensional input space is addressed by dimensionality reduction (feature selection, also known as subset selection), and the problem of heterogeneous data is handled by encoding of all inputs into the same type. Then a standard inductive classifier (such as Support Vector Machine (SVM), or a neural network, or logistic regression) is used to estimate a model. This approach has been successfully used in many biomedical and image processing applications (Camps-Valls, Rojo-Alvarez, & Martinez-Ramon, 2007). Commonly used statistical approaches to modeling genetic data for diagnostic and prognostic classification follow feature selection strategy (also known as subset selection) where a few strong informative inputs are selected from a large number of inputs, typically using greedy feature selection. Selection of inputs in the final model is performed via extensive use of resampling (Simon, Radmacher, Dobbin, & McShane, 2003).

The *second approach* is to investigate new learning settings for dealing with HDLSS heterogeneous data. This approach is based on the fundamental principle (due to Vapnik) that for finite sample estimation problems one should always use the most appropriate *direct formulation* of the learning problem rather than a more general formulation. It can be argued that most recent advances in statistical learning (i.e., transduction, semi-supervised learning, single-class learning, multi-task learning) reflect an improved understanding of the learning problem setting.

Multi-Task Learning, also known as transfer learning, has had a relatively long history in machine learning. Learning multiple related tasks simultaneously has been empirically (Ando & Zhang, 2005; Bakker & Heskes, 2004; Evgeniou & Pontil, 2004) as well as theoretically (Ando & Zhang, 2005; Ben-David & Schuller, 2003) shown to often significantly improve predictive performance relative to learning each task independently. So MTL approaches can benefit applications using HDLSS heterogeneous data where relatively few data samples per task are available. Most Multi-Task Learning techniques can be broadly grouped into several categories, depending on how task relatedness is modeled:

- methods where multiple tasks share the same internal representation, such as hidden units in neural networks (Ando & Zhang, 2005; Bakker & Heskes, 2004; Caruana, 1997; Liao & Carin, 2005),
- estimating a common set of latent variables consisting of linear combinations of the original input features, as in Partial Least Squares (PLS) statistical approaches (Momma & Bennett, 2006),
- probabilistic methods where task relatedness is modeled by sharing priors (Lawrence & Platt, 2004; Raina, Ng, & Koller, 2006),
- modeling task relatedness via common (shared) features (Argyriou, Evgeniou, & Pontil, 2006; Obozinski, Taskar, & Jordan, 2006),
- kernel methods where different tasks share common part in their decision functions (Evgeniou & Pontil, 2004; Liang & Cherkassky, 2008).

The methods discussed in this paper are most closely related to the last category.

This paper describes application of novel learning methodologies, such as SVM+, and Multi Task Learning (MTL), to classification problems using several medical data sets. The goal is to present several different ways to model heterogeneous data (as discussed in Section 2), and then investigate advantages and limitations of different learning approaches via empirical comparisons, in Sections 3 and 4. Finally, conclusions and discussion are presented in Section 5.

2. Approaches for modeling heterogeneous data

In this paper, we consider supervised learning applications where the training data include additional (group) information about training samples. Examples include: (1) handwritten digit recognition where training examples are provided by several persons, (2) medical diagnosis where a predictive (diagnostic) model, say for lung cancer, is estimated using a training data set of male and female patients, etc. Incorporating this additional information has lead to approaches known as Multi-Task Learning (Ando & Zhang, 2005; Ben-David, Gehrke, & Schuller, 2002; Evgeniou & Pontil, 2004; Liang & Cherkassky, 2008) and, more recently, to Learning with Structured Data (also known as SVM+) (Vapnik, 2006), as briefly discussed next.

Suppose that the training data can be represented as a union of t related groups, i.e. each group $r \in [1, 2, \ldots, t]$ contains n_r samples independently and identically generated from a distribution P_r on $\mathbf{x} \times y$. Therefore, the available data are a union of $t > 1$ groups: $\{\{\mathbf{X}_r, \mathbf{Y}_r\}, r = 1, \ldots, t\}$, $\{\mathbf{X}_r, \mathbf{Y}_r\} = \{\{\mathbf{x}_{r_1}, y_{r_1}\}, \ldots, \{\mathbf{x}_{r_{nr}}, y_{r_{nr}}\}\}$, and it can be thought of as samples identically and independently generated from an unknown distribution $P(\mathbf{x}, y) = \{P_r(\mathbf{x}, y), \text{if } \{\mathbf{x}, y\} \in \{\mathbf{X}_r, \mathbf{Y}_r\}\}$.

If the group labels of future test samples are not given, the appropriate formulation is known as "**L**earning **W**ith **S**tructured **D**ata (LWSD)" (Vapnik, 2006). In this formulation, the goal is to find the best mapping function f such that the expected loss

$$R_{LWSD}(w) = \int L(f(\mathbf{x}, w), y) P(\mathbf{x}, y) \mathrm{d}\mathbf{x}\mathrm{d}y$$

is minimized. Note that even though the expected loss is in the same form as in the supervised learning setting, the difference is that in the supervised learning setting P is unknown, while in LWSD it is known that P is a union of t sub-distributions.

On the other hand, if the group labels of future test samples are given, the problem is formalized as **M**ulti-**T**ask **L**earning (MTL) (Ando & Zhang, 2005; Ben-David et al., 2002; Liang & Cherkassky, 2008; Vapnik, 1998). The goal in multi-task learning is to estimate t related classifiers $\{f_1, f_2, \ldots, f_t\}$ so that the sum of expected losses for each task

$$R_{MTL}(w) = \sum_{r=1}^{t} \left(\int L(f_r(\mathbf{x}, w), y) P_r(\mathbf{x}, y) \mathrm{d}\mathbf{x}\mathrm{d}y \right)$$

is minimized.

Download English Version:

https://daneshyari.com/en/article/404641

Download Persian Version:

https://daneshyari.com/article/404641

Daneshyari.com