Engineering Applications of Artificial Intelligence 28 (2014) 174–180

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Combining multiple views: Case studies on protein and arrhythmia features

C. Okan Sakar [a,*], Olcay Kursun [b], Huseyin Seker [c], Fikret Gurgen [d], Nizamettin Aydin [e], Oleg Favorov [f]

[a] Department of Computer Engineering, Bahcesehir University, Istanbul, Turkey
[b] Department of Computer Engineering, Istanbul University, Istanbul, Turkey
[c] Bio-Health Informatics Research Group, Department of Informatics, De Montfort University, UK
[d] Department of Computer Engineering, Bogazici University, Istanbul, Turkey
[e] Department of Computer Engineering, Yildiz Technical University, Istanbul, Turkey
[f] Department of Biomedical Engineering, University of North Carolina, Chapel Hill, NC, USA

ARTICLE INFO

Article history:
Received 25 January 2012
Received in revised form
10 April 2013
Accepted 4 November 2013
Available online 2 December 2013

Keywords:
Multi-view learning
Ensemble methods
Protein structure prediction
Protein sub-nuclear location prediction
Arrhythmia type prediction

ABSTRACT

Computational annotation of protein functions and structures from sequence features, or prediction of certain diseases from gene expression levels are among important applications of computational biology. Developing methods capable of such predictions are not only important in terms of their biological and medical uses but also a very challenging task of pattern recognition due to high input dimensionality and small sample size. Ensemble and multi-view learning has gained popularity due to the rapid rise of such datasets (such as the protein and arrhythmia datasets used in this paper) with large numbers of variables. However, the classical ensemble approach does not take into account conditional interdependences among the views. In this paper, we present a two stage supervised multi-view learning technique called parallel interacting multi-view learning (*PIML*). In the first stage of *PIML*, similar to the ensemble method, the views are individually used by a predictor, and the class posterior probability estimates are obtained. In the second stage, each view is trained using its own features along with the class posterior probability estimates of the other views as the summary information of other views. This is a hybrid way of combining the views in which the views influence each other during training using the predictions of others interdependences. *PIML* is demonstrated and compared with the classical ensemble approach on three real datasets.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Ensemble learning, and more recently, multi-view learning has gained considerable interest in predictive tasks regarding high dimensional biomedical/bioinformatics datasets (Bach et al., 2004; Ruping and Scheffer, 2005; Alpaydin, 2010; Yang et al., 2010). The term "multi-view" is used to refer multiple sets of features about the same underlying phenomenon. The datasets can come as naturally organized into such views (e.g. chemical and biological views of data in drug discovery, or acoustic features and motion of lip region in speech recognition) or can be artificially divided into groups in order for utilizing multiple predictors on each view. These views belong to the same object or class label in supervised settings. The artificial division can be as simple as creating random subsets or applying different feature selection methods on the same data (e.g. applying different sequence driven feature extraction methods on protein sequences yields different groups of features, describing the same sequence in different ways).

In this paper, we propose an ensemble multi-view learning approach in which the curse of dimensionality problem is avoided while the views interact during the training phase. We use three real datasets (two protein datasets and one dataset regarding arrhythmia) to demonstrate the novel ensemble method proposed. For the protein datasets, we have the structure prediction and sub-nuclear location prediction tasks. These protein datasets are split into views using different sequence-driven protein feature extraction methods. For the arrhythmia dataset, we use the random subspace method (Bryll, 2003) to randomly split it into views. The task is to predict the type of arrhythmia. Our experimental results on all three datasets show that the proposed method is the superior to the classical ensemble approach.

The rest of this paper is organized as follows. Section 2 discusses the classical ensemble approach. Section 3 presents the proposed technique called parallel interacting multi-view learning. Section 4 presents the experimental results and we conclude in Section 5.

* Corresponding author. Tel.: +90 212 3810571; fax: +90 212 3810550.
E-mail addresses: okan.sakar@bahcesehir.edu.tr, okan.sakar@boun.edu.tr, okansakars@gmail.com (C.O. Sakar), okursun@istanbul.edu.tr (O. Kursun), hseker@dmu.ac.uk (H. Seker), gurgen@boun.edu.tr (F. Gurgen), naydin@yildiz.edu.tr (N. Aydin), favorov@bme.unc.edu (O. Favorov).

0952-1976/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2013.11.004

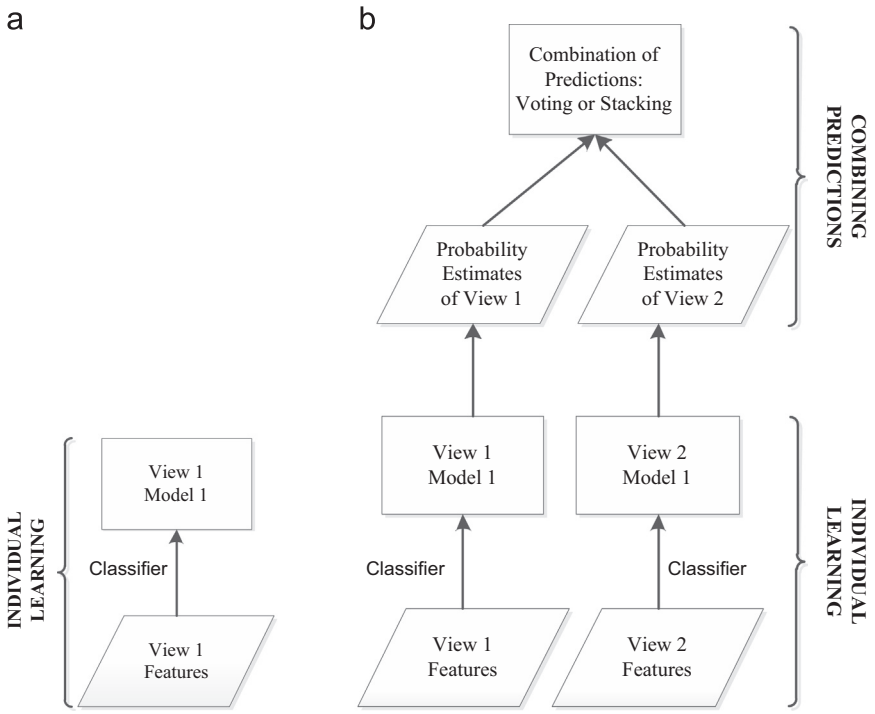

Fig. 1. (a) Single view learning and (b) ensemble multi-view learning.

2. On the ensemble and multi-view learning approaches

The simplest approach to supervised learning from the data with multiple views (i.e. multiple feature sets) is to concatenate all the features and use it as a single view data (Fig. 1). However, single view approach suffers from two main shortcomings: first, it increases the chances of facing the curse of dimensionality and also increases the complexity of the system. Second, it fails to model the individual views of the data sampled from different multivariate statistical distributions, thus achieves lower generalization of the classifiers (Christoudias et al., 2008).

A more sophisticated approach to supervised learning on multi-view data is ensemble learning which is based on employing separate classifiers on each view and combining the predictions of the views using techniques such as voting and stacked generalization (stacking) (Okun and Priisalu, 2005). In ensemble learning, the final prediction of a sample may be obtained using simple majority voting, weighted majority voting, stacked generalization (Alpaydin, 2010), or probabilistic approximation (Kang et al., 1997). The combination scheme is called simple voting if the prediction of each view contributes equally to the final prediction. In stacking, the weights of the view predictions are learned by another learner which does not need to be linear. In classification problems, the information of how much confident the view is for its prediction can be used to obtain the final prediction instead of the hard class labels. However, these ensemble techniques are expected to work well when there are no conditional dependences (given the class-label) among the views since the views do not interact during their individual learning processes (Blum and Mitchell, 1998; Abney, 2002).

3. PIML: parallel interacting multi-view learning

In this study, we propose a novel multi-view learning approach, in which the views interact (in parallel) during the training process, not only for avoiding the curse of dimensionality but also for modeling at least some of the interdependences among the views (Sakar et al., 2009). In other words, *PIML* addresses the following drawbacks of the existing multi-view learning approaches: (1) the lack of training phase interaction problem of the classical ensemble learning approach, (2) the curse of dimensionality problem of the approach which merges the views of the dataset and treats it as if it consists of a single view. In the architecture of *PIML*, the views interact in parallel during the training process for modeling the interdependences among the views, and also the curse of dimensionality is avoided since only the probability estimates of other views as summary information are used instead of high-dimensional original input space. The main idea is as follows: the classifier of each view uses the input variables from its own view along with the predictions (outputs) of the classifiers of the other views. In other words, each view uses the summary of information in the other views and evaluates its own input features (again) this time also by taking into account the predictions obtained from classifiers trained, in a similar fashion, on the other views. This technique increases the individual accuracies (sufficiency) of the views by taking the class posterior probability estimates of the other views during its second training phase, and also aims to preserve the diversity of the views by merging the original features of the individual views only with the summary information of the other views. Therefore, *PIML* approach is expected to reach higher accuracy than its counterparts that merge all the variables of all the views (Fig. 1a) or combine their predictions after the individual learning process, i.e. ensemble methods (Fig. 1b).

If we used the probability estimates of all the views directly, we can only create a single stacking network. Thus, our *PIML* strategy resembles blocking (Bontempi and Blocking, 2007), which is an experimental design strategy which produces similar experimental conditions to compare alternative stochastic configurations in order to be confident that observed differences in accuracy are due to actual differences rather than to fluctuations and noise effects. Using each view's raw features along with the probability estimates of other views at least corresponds to creating multiple versions of stackings (yet possibly still using the same classifier model), thus increasing the number of blocking configurations.

Download English Version:

https://daneshyari.com/en/article/380643

Download Persian Version:

https://daneshyari.com/article/380643

Daneshyari.com