Knowledge-Based Systems 000 (2018) 1–9

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

An alternate method between generative objective and discriminative objective in training classification Restricted Boltzmann Machine

Linkai Luo[a,b,*], Songfei Zhang[a], Yudan Wang[a], Hong Peng[a,b]

[a] Department of Automation, Xiamen University, Xiamen, PR China
[b] Xiamen Key Laboratory of Big Data Intelligent Analysis and Decision-marking, PR China

ARTICLE INFO

Article history:
Received 8 May 2017
Revised 20 November 2017
Accepted 27 December 2017
Available online xxx

Keywords:
Restricted Boltzmann Machine
ClassRBM
Alternately training

ABSTRACT

As a derivative of Restricted Boltzmann Machine (RBM), classification RBM (ClassRBM) has been an effective classifier. However, there are still many disadvantages in training ClassRBM. For example, the prediction accuracy with the generative objective function (GenF) is not high, and the training process with the discriminative objective function (DisF) and the hybrid RBM (HDRBM) are time-consuming. In this paper, we propose an alternate method between Generative Objective and Discriminative Objective (ANGD) to train ClassRBM after examining the training process of GenF and DisF. At each iteration step of ANGD, the parameters of ClassRBM are firstly updated by maximizing GenF when the training accuracy can be improved, then modified by maximizing DisF. This process is repeated until some stop criterion is met. ANGD achieves a good prediction accuracy with a relatively less training cost because it utilizes the complementation of GenF and DisF. The comparative experiments on five datasets show that ANGD beats GenF, DisF and HDRBM. As a whole, the accuracy of ANGD is the best and the stability is acceptable, and the training cost of ANGD is also the best on the datasets with a large size. The training efficiency of ANGD is the best among the four methods.

1. Introduction

Restricted Boltzmann Machine (RBM) has attracted a lot of attention recently. As a random neural network, RBM can be used for feature extractor, components of Deep Belief Networks, parameters initialization of neural network, dimensionality reduction, collaborative filtering, etc. [1–3]. Moreover, RBM can be used in classification problems [4–9].

The classification RBM (ClassRBM) is the most popular way of using RBM for classification. There are two commonly used objective functions in training ClassRBM. One is generative function (GenF) while the other is discriminative function (DisF). GenF is the joint distribution of the input variable v and its target classes y, i.e. $p(v, y)$, while DisF is the conditional distribution $p(y|v)$. It is difficult to calculate exactly GenF and its gradient [10] because we need to traverse all possible states and the number of possible states is often huge. In practical training, CD algorithm [1,11] is commonly used to approximate GenF [8,9]. In addition, the information of class labels is submerged by input features because the number of input features is often much more than the number of classes, which results in the information of class labels cannot be fully utilized. Hence the prediction accuracy in the methods based on GenF is usually not high. Compared with GenF, DisF and its gradient can be calculated exactly. In addition, in DisF, there is no phenomenon that the information of class labels is submerged by input features. The prediction accuracy in the methods based on DisF is generally higher than that of GenF. Therefore, most of the researchers are more interested in DisF or the hybrid of GenF and DisF. Larochelle and Bengio [6] evaluated different training objectives and showed that a hybrid (linear combination) of GenF and DisF achieved lower errors. However, the combined coefficient α between GenF and DisF is uneasy to set. Tomczak [8] proposed a sparse learning method for ClassRBM (sparseClassRBM) in which a regularization term is added to DisF. Tomczak showed both DisF and sparse ClassRBM obtain competitive classification results for five medical problems and outperforms well-known strong classifiers, such as AdaBoost, LogitBoost, Random Forest, TreeBagging. But it performs badly on the highly imbalanced data. Tomczak and Zieba [9] further applied the sparse ClassRBM to credit scoring by combining the geometric mean of specificity and sensitivity. Tomczak and Zieba showed that their method can resist to the influence of imbalanced data, and obtains high predictive performance. Yu et al. [12] proposed a learning criterion based on DisF. They extend the conditional log likelihood objective to a general learning

* Corresponding author at: Department of Automation, Xiamen University, Xiamen, PR China.
E-mail address: luolk@xmu.edu.cn (L. Luo).

https://doi.org/10.1016/j.knosys.2017.12.032
0950-7051/

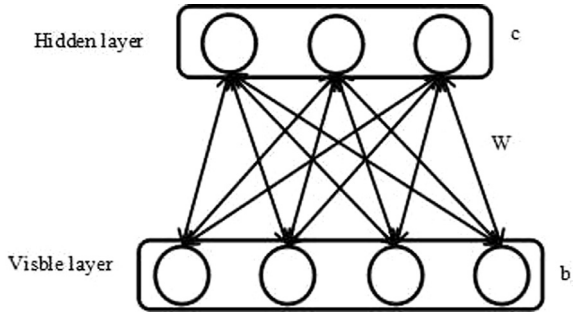

Fig. 1. A frame of RBM with a visible layer and a hidden layer.

criterion by introducing rěnyi divergence, and carry out an effective model selection. Ji et al. [13] proposed a single hidden-layer network TIClassRBM by incorporating linear transformations into DisF. TIClassRBM can extract invariant features and train the classifier at the same time. Yin et al. [14] presented a method based on reconstruction error to improve the classification performance of ClassRBM. In the method of Yin et al., a ClassRBM is firstly trained with GenF or DisF or their hybrid, an additional classifier is then trained by the reconstruction error and the output of hidden layer. The final classification result is a combination of two classifiers. Yin et al. show their method improve the classification performance of ClassRBM.

Since all methods mentioned above are based on GenF or DisF or the hybrid, there are some inherent disadvantages originated from them. In short, the prediction accuracy for GenF is not high. DisF spends much more training time than GenF and the prediction accuracies appear fluctuations when there are many classes or many hidden units. In the hybrid approach, it is not easy to select the optimal combination coefficient of GenF and DisF.

To overcome the above disadvantages, we propose an alternate method between Generative Objective and Discriminative Objective(ANGD) to train ClassRBM. The training objective of ANGD is alternated between GenF and DisF. At each iteration step, the parameters of ClassRBM are first updated by maximizing GenF when the training accuracy can be improved, then modified by maximizing DisF. This process is repeated until a stop criterion based on DisF is met. ANGD utilizes the complementation of GenF and DisF in training process. It does not have the problem in which the class labels are submerged by input variables and does not need to select the combination coefficient of GenF and DisF. The comparison experiments on five public datasets show that ANGD beat GenF, DisF and HDRBM.

The rest of this paper is organized as follows. Section 2 gives a brief introduction of RBM and ClassRBM. In Section 3, ANGD is proposed for training ClassRBM. The comparison experiments among GenF, DisF, HDRBM, and ANGD are carried out in Section 4. Finally, the conclusions are provided in Section 5.

2. Related works

2.1. Restricted Boltzmann Machine

RBM is an undirected generative model with two layers. One is visible layer that consists of input variables while another is hidden layer that consists of hidden variables. In the same layer, units are independent of each other. Symmetric weights are used for the connections of units in different layers. Fig. 1 shows a frame of RBM [15].

RBM tries to model the probability distribution of input variables in the visible layer by the hidden variables. RBM introduces an energy function to describe the energy of a state. The energy function of RBM with binary variables is defined by

$$E(v,h) = -\sum_{i=1}^{n} b_i v_i - \sum_{j=1}^{m} c_j h_j - \sum_{i=1}^{n}\sum_{j=1}^{m} v_i w_{ij} h_j, \tag{1}$$

where v_i, h_j, respectively, are the values of the ith visible unit and the jth hidden unit, w_{ij} is the connection weight coefficient between them, b_i is the bias of v_i, c_j is the bias of h_j, n and m are the number of visible units and hidden units respectively. According to the energy function, the joint probability of a state (v,h) is

$$p(v,h) = \frac{exp(-E(v,h))}{Z}, \tag{2}$$

where

$$Z = \sum_{v,h} exp(-E(v,h)). \tag{3}$$

Summing $p(v,h)$ over the hidden units, we can obtain the marginal distribution of the visible units by

$$p(v) = \frac{\sum_h exp(-E(v,h))}{Z}. \tag{4}$$

It is almost impossible to calculate $p(v)$ because the number of combinations for (v,h) is usually huge. However, we can calculate the conditional probabilities

$$p(h_j = 1|v) = \sigma\left(c_j + \sum_i v_i w_{ij}\right), \tag{5}$$

and

$$p(v_i = 1|h) = \sigma\left(b_i + \sum_j w_{ij} h_j\right). \tag{6}$$

where

$$\sigma(x) = \frac{1}{1 + exp(-x)} \tag{7}$$

Given a training set T, the objective of RBM is maximizing the marginal distribution of the visible units $p(v)$, i.e.,

$$\max_{\theta} \sum_{v \in T} \ln p(v), \tag{8}$$

where $\theta = \{W, b, c\}$ is the parameter set of RBM,$W = [w_{ij}]$, $b = [b_i]$, $c = [c_j]$ [16].

CD algorithm is commonly used to approximately calculate $p(v)$. If a conventional stochastic gradient ascent is used to solve (8), then the parameters are updated by

$$w_{ij} \leftarrow w_{ij} + \frac{\varepsilon}{bs}(p(h_j = 1|v)v_i - p(h_j = 1|v_{II})v_{IIi}), \tag{9}$$

$$b_i \leftarrow b_i + \frac{\varepsilon}{bs}(v_i - v_{IIi}), \tag{10}$$

$$c_j \leftarrow c_j + \frac{\varepsilon}{bs}(p(h_j = 1|v) - p(h_j = 1|v_{II})), \tag{11}$$

where ε is the learning rate, bs is the size of mini-batch, v_{IIi} is the reconstructed value of v_i.

2.2. Classification Restricted Boltzmann Machine

Different from RBM, the visible units of ClassRBM are composed by feature vector $v = \{v_1, v_2, \ldots, v_n\}$ and the corresponding class label vector y. The length of y is equal to the number of class labels. If a feature vector v belongs to class k, then the corresponding class label vector satisfies that $y(k) = 1$ and $y(i) = 0$ for $i \neq k$. Fig. 2 shows a structure of ClassRBM. The parameter set $\theta = \{W, b, c\} \cup \{U, d\}$ where $U = [u_{kj}]$ are the connection weight

Please cite this article as: L. Luo et al., An alternate method between generative objective and discriminative objective in training classification Restricted Boltzmann Machine, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2017.12.032