Multi-view Restricted Boltzmann Machines with Posterior Consistency

Ding Shifei, Zhang Nan(✉), and Zhang Jian

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
{dingsf, chinaxxzhangnan}@cumt.edu.cn,
597409675@qq.com

Abstract. Restricted Boltzmann machines (RBMs) have been proven to be powerful tools in many specific applications, such as representational learning and document modelling. However, the extensions of RBMs are rarely used in the field of multi-view learning. In this paper, we present a new multi-view RBM model, named as the RBM with posterior consistency, for multi-view classification. The RBM with posterior consistency computes multiple representations by regularizing the marginal likelihood function with the consistency among representations from different views. Contrasting with existing multi-view classification methods, such as multi-view Gaussian pro-cess with posterior consistency (MvGP) and consensus and complementarity based maximum entropy discrimination (MED-2C), the RBM with posterior consistency have achieved satisfactory results on two-class and multi-class classification datasets.

Keywords: Restricted Boltzmann machines · Representational learning
Multi-view learning

1 Introduction

Restricted Boltzmann machines (RBMs) are popular probability graph models for representing dependency structure between random variables [1]. It is very known that RBMs are energy-based models and powerful tools for representational learning. By modifying energy functions, RBMs can be widely used in artificial intelligence and machine learning fields [2]. RBMs have been developed for real-valued data modelling [3], sequential data modelling [4], noisy data modelling [5], document modelling [6], multimodal learning [7], and other applications. RBMs also are basic building blocks for creating deep belief networks (DBNs) [1] and deep Boltzmann machines (DBMs) [8]. Contrasting with RBMs, these two deep networks show better representational learning and classification abilities.

The general RBM and many RBM variants are only suitable for addressing the single view data. Actually, there are many data coming from multiple views, where each view may be a feature vector or a domain. Therefore, many researchers focus on the multi-view learning task [9]. Recently, there are many efficient multi-view algorithms for classification, such as multi-view Gaussian process with posterior Consistency (MvGP) [10] and consensus and complementarity based maximum entropy

Z. Shi et al. (Eds.): IIP 2018, IFIP AICT 538, pp. 30–39, 2018.
https://doi.org/10.1007/978-3-030-00828-4_4

discrimination (MED-2C) [11]. The MvGP and the MED-2C are posterior-consistency style and margin-consistency style algorithms, respectively. These two algorithms both balance the relationship between the multi-view data and the model and achieve the state-of-the-art classification accuracy. It is very known that RBMs are powerful tools in machine learning, but RBMs have few applications in multi-view learning. Our work focuses on the consistency among view-specific hidden layers and balances the relationship between the multi-view data and the model for classification.

In this paper, we first propose a new RBM model, named as the RBM with posterior consistency (PCRBM), for multi-view classification. The PCRBM models each separated view as a RBM. The weights of the original RBM are optimized by maximizing the log likelihood function. Unlike the general RBM, the PCRBM updates weights by maximizing the log likelihood function on each view and maximizing the consistency among the hidden layer conditional distributions on each view. In addition, original RBMs only deal with the binary data, so we extend PCRBMs to exponential family RBMs (Exp-RBMs) [12] and propose exponential family RBMs with posterior consistency (Exp-PCRBMs). In the Exp-PCRBM, activation functions in visible or hidden units can be any smooth monotonic non-linearity function, such as Gaussian function and ReLU function.

The remainder of the paper is organized as follows. Section 2 details PCRBMs, including the inference and learning procedures. Section 3 gives extensions of PCRBMs for multi-view data and real data. In Sect. 4, experiment results prove the feasibilities of the proposed methods. Finally, some conclusions and the intending work are given in the last section.

2 Restricted Boltzmann Machines with Posterior Consistency for Two-View Classification

2.1 Restricted Boltzmann Machines with Posterior Consistency for Two-View Data

It well known that the general RBM is only suitable for addressing single view data. We propose a new RBM model to deal with two-view data and call it the RBM with posterior consistency (PCRBM). The PCRBM first makes use of a general RBM to model each view of data. That is, the conditional probabilities of visible or hidden units in the PCRBM are similar to the general RBM. And then, the PCRBM utilizes the stochastic approximation method to update network weights by maximizing the log likelihood function on each view and maximizing the consistency between the conditional probabilities of hidden units given visible data on each view. In the PCRBM, the negative of the distance between two conditional probabilities is used to measure the consistency between two conditional probabilities. The PCRBM is also a generative model, and it contains two layers of visible units $\mathbf{v}^1 = \{v_i^1\}_{i=1}^{D1}, \mathbf{v}^2 = \{v_i^2\}_{i=1}^{D2}$ and two layers of hidden units $\mathbf{h}^1 = \{h_j^1\}_{j=1}^{J}, \mathbf{h}^2 = \{h_j^2\}_{j=1}^{J}$ corresponding to the two-view data with the connection weights $\theta = \{\mathbf{W}^1, \mathbf{b}^1, \mathbf{c}^1, \mathbf{W}^2, \mathbf{b}^2, \mathbf{c}^2\}$. The energy function of the

PCRBM is composed of two general RBM models, then conditional probabilities on two views can be given by:

$$P(h_j^1 = 1|\mathbf{v}^1) = \sigma\left(\sum_i v_i^1 W_{ij}^1 + b_j^1\right), \quad P(v_i^1 = 1|\mathbf{h}^1) = \sigma\left(\sum_j W_{ij}^1 h_j^1 + c_i^1\right), \tag{1}$$

$$P(h_j^2 = 1|\mathbf{v}^2) = \sigma\left(\sum_i v_i^2 W_{ij}^2 + b_j^2\right), P(v_i^2 = 1|\mathbf{h}^2) = \sigma\left(\sum_j W_{ij}^2 h_j^2 + c_i^2\right), \tag{2}$$

where $\sigma(x) = 1/(1 + \exp(-x))$.

Assume that the two views training sample set $\mathbf{X1} = \left\{\mathbf{v}^{1\,(n)}\right\}_{n=1}^N$, $\mathbf{X2} = \left\{\mathbf{v}^{2(n)}\right\}_{n=1}^N$, $\mathbf{Y} = \left\{\mathbf{Y}^{(n)}\right\}_{n=1}^N$, where $\mathbf{X1}$ and $\mathbf{X2}$ are two-view data, and $\mathbf{Y}$ is the corresponding label. In order to maximize the consistency between hidden layer conditional probabilities of two views, the objective of the PCRBM can be expressed as:

$$\begin{aligned} \max_{\theta} \quad & \sum_n \ln P\left(\mathbf{v}^{1\,(n)}; \theta\right) + \sum_n \ln P\left(\mathbf{v}^{2(n)}; \theta\right) \\ & + \lambda consistency\left(\sum_n P\left(\mathbf{h}^{1(n)}|\mathbf{v}^{1(n)}; \theta\right), \sum_n P\left(\mathbf{h}^{2(n)}|\mathbf{v}^{2(n)}; \theta\right)\right), \end{aligned} \tag{3}$$

where λ is a parameter to balance the log likelihood function. We can make use of the stochastic approximation algorithm and the derivation of the posterior consistency to maximize the objective function, and the details is given in next section. After the pre-training, the PCRBM utilizes the data with labels and the gradient descent method to fine-tune the weights for classification. In the general RBM, the weights connecting visible units and hidden units are also fine-tuned. However, in the PCRBM, the weights connecting visible units and hidden units contain the posterior consistency between two views and the conditional probabilities over hidden units given visible units should remain unchanged. Define $\mathbf{H}^{1(n)} = P\left(\mathbf{h}^{1(n)}|\mathbf{v}^{1(n)}; \theta\right)$, and $\mathbf{H}^{2(n)} = P\left(\mathbf{h}^{2(n)}|\mathbf{v}^{2(n)}; \theta\right)$ $(\mathbf{H}^1, \mathbf{H}^2 \in \Re^{N \times J})$, and then the objective function of the classification model can be expressed as:

$$\min_{\theta'} \quad \frac{a}{2}\sum_n \left\|\mathbf{Y}^{(n)} - P\left(\hat{\mathbf{Y}}^{(n)}|\mathbf{H}^{1\,(n)}; \theta'\right)\right\|^2 + \frac{(1-a)}{2}\sum_n \left\|\mathbf{Y}^{(n)} - P\left(\hat{\mathbf{Y}}^{(n)}|\mathbf{H}^{2(n)}; \theta'\right)\right\|^2, \tag{4}$$

where $a \in [0, 1]$ is a parameter to balance two views, and

$$\begin{aligned} P\left(\hat{Y}_l^{(n)}|\mathbf{H}^{1\,(n)}; \theta'\right) &= \exp\left(\sum_j H_j^{1\,(n)} W2_{jl}^1 + b2_l^1\right) \Big/ \sum_l \exp\left(\sum_j H_j^{1\,(n)} W2_{jl}^1 + b2_l^1\right), \\ P\left(\hat{Y}_l^{(n)}|\mathbf{H}^{2(n)}; \theta'\right) &= \exp\left(\sum_j H_j^{2(n)} W2_{jl}^2 + b2_l^2\right) \Big/ \sum_l \exp\left(\sum_j H_j^{2(n)} W2_{jl}^2 + b2_l^2\right). \end{aligned} \tag{5}$$

Therefore, we use the gradient descent method to fine-tune the weights connecting hidden units and label units [13]. The PCRBM is not only suitable for two-class classification data but also for multi-class classification data.

2.2 Inference and Learning Procedure for Two-View Data

For each view, the gradient with respect to a weight can be divided into two parts, the gradient of the posterior consistency and the gradient of the log likelihood function. the consistency between $\mathbf{H}^1$ and $\mathbf{H}^2$ can be defined as the negative of the distance between two conditional probabilities

$$\begin{aligned} consistency\left(\mathbf{H}^1, \mathbf{H}^2\right) &= \frac{1}{N}\sum\nolimits_n \left(-\frac{1}{2}\frac{\left\|\mathbf{H}^{1(n)} - \mathbf{H}^{2(n)}\right\|^2}{\left\|\mathbf{H}^{1(n)}\right\|^2 + \left\|\mathbf{H}^{2(n)}\right\|^2} \right) \\ &= \frac{1}{N}\sum\nolimits_n \left(\frac{\mathbf{H}^{1(n)} \odot \mathbf{H}^{2(n)}}{\left\|\mathbf{H}^{1(n)}\right\|^2 + \left\|\mathbf{H}^{2(n)}\right\|^2} \right) - \frac{1}{2}, \end{aligned} \tag{6}$$

where $\odot$ denotes element-wise multiplication. We have used the mean-field variational inference method to obtain $\mathbf{H}^{1(n)} = P\left(\mathbf{h}^{1(n)}|\mathbf{v}^{1(n)};\theta\right)$ and $\mathbf{H}^{2(n)} = P\left(\mathbf{h}^{2(n)}|\mathbf{v}^{2(n)};\theta\right)$. To compute the gradient of the consistency with respect to a weight, we can compute the gradient of the consistency with respect to $\mathbf{H}^1$ and $\mathbf{H}^2$ and then use backpropagation. Take as an example the gradient with respect to $\mathbf{W}^1$ in the first view. The gradient of the posterior consistency with respect to $\mathbf{W}^1$ can be given by:

$$\Delta\mathbf{W}^1_{consistency} = \frac{1}{N}\mathbf{X}1^T\left(\mathbf{H}^1 \odot \left(1 - \mathbf{H}^1\right) \odot \left(\frac{\mathbf{H}^2}{\left\|\mathbf{H}^1\right\|^2 + \left\|\mathbf{H}^2\right\|^2} - \frac{2\mathbf{H}^1 \odot \left(\mathbf{H}^1 \odot \mathbf{H}^2\right)}{\left\|\left\|\mathbf{H}^1\right\|^2 + \left\|\mathbf{H}^2\right\|^2\right\|^2} \right)\right). \tag{7}$$

In addition, the gradient of the log likelihood function with respect to a weight can be simplified to the difference between data-dependent statistic and model-dependent statistic. Moreover, the CD-k or other stochastic approximation algorithms provide an effective way to estimate the mode-dependent statistic. The gradient of the log likelihood function with respect to $\mathbf{W}^1$ can be given by:

$$\Delta\mathbf{W}^1_{\log-\text{likelihood}} = \left(\mathbf{E}_{P_{\text{data}}}\left[\mathbf{X}1^T\mathbf{H}^1\right] - \mathbf{E}_{P_{\text{model}}}\left[\mathbf{X}1^T\mathbf{H}^1\right]\right)/N, \tag{8}$$

This way, the gradient of the objective function with respect to $\mathbf{W}^1$ can be given by:

$$\begin{aligned} \Delta\mathbf{W}^1 = \Delta\mathbf{W}^1_{\log-\text{likelihood}} + \lambda\Delta\mathbf{W}^1_{\text{correlation}} = \frac{1}{N}\left(\mathbf{E}_{P_{\text{data}}}\left[\mathbf{X}1^T\mathbf{H}^1\right] - \mathbf{E}_{P_{\text{model}}}\left[\mathbf{X}1^T\mathbf{H}^1\right]\right) \\ + \frac{\lambda}{N}\mathbf{X}1^T\left(\mathbf{H}^1 \odot \left(1 - \mathbf{H}^1\right) \odot \left(\frac{\mathbf{H}^2}{\left\|\mathbf{H}^1\right\|^2 + \left\|\mathbf{H}^2\right\|^2} - \frac{2\mathbf{H}^1 \odot \left(\mathbf{H}^1 \odot \mathbf{H}^2\right)}{\left\|\left\|\mathbf{H}^1\right\|^2 + \left\|\mathbf{H}^2\right\|^2\right\|^2} \right)\right). \end{aligned} \tag{9}$$

Likewise, the gradients of the objective function with respect to other weights are computed by using the similar method.

3 Extensions of Restricted Boltzmann Machines with Posterior Consistency

3.1 Extensions for Multi-view Data

By taking two views as an example, we detail the model of restricted Boltzmann machines with posterior consistency (PCRBMs) in the above section. The PCRBM has two objective functions corresponding to two-stage tasks, the objective for maximizing the log likelihood function and the correlation and the objective for classification. The reason the PCRBM can be extend to address multi-view data is that each objective function can be express as an elegant formulation. In the first-stage task, the objective for multiple views also can divided into two parts, the log likelihood function on each view and maximizing the posterior consistency among multiple views. The PCRBM also model each view of data as a general RBM, and the conditional probabilities of hidden units given visible units is easily sampled. Moreover, the posterior consistency between two conditional probabilities can be calculated by the negative of the distance between two conditional probabilities. For a multiple views training set of N samples $\mathbf{X}1 = \{\mathbf{v}^{1\,(n)}\}_{n=1}^{N}, \cdots, \mathbf{X}K = \{\mathbf{v}^{K\,(n)}\}_{n=1}^{N}$, $\mathbf{Y} = \left\{\mathbf{Y}^{(n)}\right\}_{n=1}^{N}$, the objective for maximizing the log likelihood function and the posterior consistency in multiple views can be expressed as:

$$\begin{aligned}\max_{\theta}\ & \sum\nolimits_{k}\sum\nolimits_{n}\ln P\left(\mathbf{v}^{k\,(n)};\theta\right) \\ & + \sum\nolimits_{i=1}^{K}\sum\nolimits_{j>i}^{K}\sum\nolimits_{n}\lambda_{ij}consistency\left(P\left(\mathbf{h}^{i\,(n)}|\mathbf{v}^{i\,(n)};\theta\right),P\left(\mathbf{h}^{j\,(n)}|\mathbf{v}^{j\,(n)};\theta\right)\right).\end{aligned} \tag{10}$$

We can find that the objective in k-view is that

$$\begin{aligned}\max_{\theta}\ & \sum\nolimits_{n}\ln P\left(\mathbf{v}^{k\,(n)};\theta\right) \\ & + \sum\nolimits_{i\neq k}^{K}\sum\nolimits_{n}\lambda_{ik}consistency\left(P\left(\mathbf{h}^{i\,(n)}|\mathbf{v}^{i\,(n)};\theta\right),P\left(\mathbf{h}^{k\,(n)}|\mathbf{v}^{k\,(n)};\theta\right)\right),\end{aligned} \tag{11}$$

and utilize the stochastic approximation algorithm and the derivation of the correlation to maximize the objective function in k-view. In the second-stage task, we utilize the data with labels and the gradient descent method to fine-tune the weights connecting hidden units and label units. The objective function for classification in multiple views can be expressed as:

$$\min_{\theta'}\quad \tfrac{1}{2}\sum\nolimits_{k}a_{k}\sum\nolimits_{n}\left\|\mathbf{Y}^{(n)}-P\left(\hat{\mathbf{Y}}^{(n)}|\mathbf{H}^{k\,(n)};\theta'\right)\right\|^{2}. \tag{12}$$

3.2 Exponential Family Restricted Boltzmann Machines with Posterior Consistency for Real Data

Ravanbakhsh et al. proposed exponential family RBMs (Exp-RBMs) where each unit can choose any smooth monotonic non-linearity function as the activation function. Regardless of the activation function, each visible (hidden) unit receives an input $\upsilon_i = \sum_j W_{ij} h_j + c_i$ ($\eta_j = \sum_i v_i W_{ij} + b_j$). Consider an Exp-RBM with variables $\{\mathbf{v}, \mathbf{h}\}$, and the energy function is defined as:

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i=1}^{D}\sum_{j=1}^{J} v_i W_{ij} h_j - \sum_{j=1}^{J} b_j h_j - \sum_{i=1}^{D} c_i v_i + \sum_{j=1}^{J} \left(R^*(h_j) + s(h_j)\right) + \sum_{i=1}^{D} \left(F^*(v_i) + g(v_i)\right), \tag{13}$$

where F^* and g are functions of v_i, the derivative of F^* is f^{-1} (f^{-1} is the inverse function of f and the anti-derivative of f is F), and similarly R^* and s are functions of h_j.

Like the general RBM, the proposed PCRBM is also only suitable for binary data. Each unit of the Exp-RBM can choose any smooth monotonic non-linearity function as the activation function. Therefore, we propose the exponential family restricted Boltzmann machines with posterior consistency (Exp-PCRBM) for multi-view learning. The proposed Exp-PCRBM is suitable for binary and real-valued data, where the activation function of each hidden unit can choose any smooth monotonic non-linearity function not just the sigmoid function. in this paper, we choose the sigmoid function as the activation function of each hidden unit in the Exp-PCRBM.

Assume that all the hidden units of the Exp-PCRBM are binary, the solution of two objective functions in the Exp-PCRBM is similar to the PCRBM. For each binary visible unit, the conditional probability is strictly the sigmoid function, and then we have $F(\upsilon_i) = \log(1 + \exp(\upsilon_i))$, $F^*(v_i) = (1 - v_i)\log(1 - v_i) + v_i \log(v_i) = 0$, and $g(v_i)$ is a constant. Thus, if each visible unit is binary, then the energy of the Exp-PCRBM is same as the PCRBM. For each visible unit obeying Gaussian conditional distribution, this distribution can be expressed as a Gaussian approximation $(f(\upsilon_i), f'(\upsilon_i))$, where $f(\upsilon_i) = \sigma_i^2 \upsilon_i$ is the mean and $f'(\upsilon_i) = \sigma_i^2$ is the variance. Then, $F(\upsilon_i) = \left(\sigma_i^2 \upsilon_i^2\right)/2$, $F^*(v_i) = v_i^2/\left(2\sigma_i^2\right)$, and $g(v_i)$ is a constant. Thus, if each visible unit obeying Gaussian conditional distribution, then the Exp-PCRBM is same as the PCRBM except conditional distributions over visible units. Therefore, in this paper, we choose the activation function of each hidden unit in the Exp-PCRBM according to the input data from each view.

4 Experiments

In order to test the performance of the algorithms, the proposed algorithms are compared with state-of-the-art classification algorithms, the multi-view Gaussian process with posterior consistency (MvGP) and consensus and complementarity based maxi-

mum entropy discrimination (MED-2C). All these algorithms are carried out in a work station with a core i7 DMI2-Intel 3.6 GHz processor and 18 GB RAM running MATLAB 2017a.

4.1 Learning Results on Two-Class Data Sets

Advertisement: The Advertisement is a binary data set, and it contains 3279 examples (459 ads and 2820 non-ads). The first view describes the image itself, while the other view contains all other features [11]. The dimensions of the two views are 587 and 967, respectively.

WDBC: The WDBC contains 569 examples (357 benign and 212 malignant). The first view contains 10 features which are computed for each cell nucleus, while the other view contains all other 20 features which is the mean and the standard error of the first view.

Z-Alizadeh sani: The Z-Alizadeh sani contains 303 examples (216 cad and 87 normal). The first view contains the patients' demographic characteristics and symptoms, while the other view contains the results of physical examinations, electrocardiography, echocardiography, and laboratory tests. The dimensions of the two views are 31 and 24, respectively.

We make use of the 5-fold cross-validation method to evaluate the proposed methods on two-class data sets, where three folds are used for training and the rest two folds for testing. In addition, we also divide the above training set into a training set and a validation set, where each of the folds is used as the validation set once (10-fold cross-validation). In the MvGP, the value of parameters a and b is determined by cross-validation from $\{0, 0.1, \ldots, 1\}$ and $\{2^{-18}, 2^{-12}, 2^{-8}, 2, 2^3, 2^8\}$, respectively [10]. In the MED-2C, the value of parameter c is determined by cross-validation from $\{2^{-5}, 2^{-4}, \ldots, 2^5\}$ [11]. Therefore, in the Exp-PCRBM, the value of parameters a and λ is determined by cross-validation from $\{0, 0.1, \ldots, 1\}$ and $\{2^{-18}, 2^{-12}, 2^{-8}, 2, 2^3, 2^8\}$, respectively. In the Exp-PCRBM, the number of hidden layer units corresponding to each view is set to 100. We also run the Exp-RBM for each view, and Exp-RBM1 and Exp-RBM2 correspond to the first view and the second view, respectively. Moreover, the Exp-RBM1, the Exp-RBM2 and the Exp-PCRBM use mini-batch learning, and 100 samples are randomly selected in every iteration.

The average accuracies and standard deviations of all the algorithms are given in Table 1. We can see that the Exp-PCRBM outperforms the other algorithms on all the data sets. From Table 1, we can also find that: (1) the Exp-PCRBM outperforms the MvGP and the MED-2C on all the data sets, which demonstrates the effectiveness of the Exp-PCRBM; (2) the MvGP performs worst on all the data sets, this is because that the point selection scheme is not used and this scheme can also be used in other algorithms; (3) the Exp-PCRBM outperforms the Exp-RBM1 and the Exp-RBM2 on all the data sets, which demonstrates that the representations from two views are perfectly used for classification in the Exp-PCRBM. We can make conclusion that the Exp-PCRBM is an effective classification method for multi-view two-class data sets.

Table 1. Performance comparison of proposed algorithms on two-class data sets

Data sets	Exp-RBM1	Exp-RBM2	MvGP	MED-2C	Exp-PCRBM
Advertisement	95.61 ± 0.39%	**96.58 ± 0.65%**	95.70 ± 1.06%	**96.68 ± 0.45%**	**96.84 ± 0.51%**
WDBC	95.87 ± 1.41%	**98.07 ± 0.50%**	96.13 ± 1.82%	96.92 ± 1.02%	**98.28 ± 0.64%**
Z-Alizadeh sani	86.80 ± 2.69%	76.74 ± 3.24%	83.98 ± 4.15%	86.47 ± 2.11%	**89.61 ± 2.14%**

4.2 Results and Evaluation

The multi-class data sets used in this paper are two UCI data sets including Dermatology and ForestTypes.

Dermatology: The Dermatology contains 358 examples (111 psoriasis, 60 seboreic dermatitis, 71 lichen planus, 48 pityriasis rosea, 48 cronic dermatitis, and 20 pityriasis rubra pilaris). The first view describes clinical features, while the other view contains histopathological features. The dimensions of the two views are 12 and 22, respectively.

ForestTypes: The ForestTypes contains 523 examples (195 Sugi, 83 Hinoki, 159 Mixed deciduous, and 86 Other). The first view describes ASTER image bands, while the other view contains all other features. The dimensions of the two views are 9 and 18, respectively.

We make use of the 5-fold cross-validation method to evaluate the proposed methods on multi-class data sets, too. Like one-versus-rest support vector machines (OvR SVMs) [14], we extend the MvGP and the MED-2C to deal with multi-class data, and name they as the one-versus-rest MvGP (OvR MvGP) and one-versus-rest MED-2C (OvR MED-2C). The parameters of the OvR MvGP, the OvR MED-2C, the Exp-RBM1, Exp-RBM2, and the Exp-PCRBM are determined by cross-validation from, too.

Table 2 shows the average accuracies and standard deviations of all the algorithms on the multi-class data sets. We can see that the Exp-PCRBM outperforms the other algorithms on all the data sets. From Table 2, we can also find that: (1) the Exp-PCRBM outperforms the MvGP and the MED-2C on all the data sets, which demonstrates the effectiveness of the Exp-CRBM on multi-class data sets; (2) the Exp-PCRBM also outperforms the Exp-RBM1 and the Exp-RBM2 on all the data sets, which demonstrates that the representations from two views are perfectly used for classification in the Exp-PCRBM. We can make conclusion that the Exp-PCRBM is an effective classification method for multi-view multi-class data sets.

Table 2. Performance comparison of proposed algorithms on multi-class data sets

Data sets	Exp-RBM1	Exp-RBM2	OvR MvGP	OvR MED-2C	Exp-PCRBM
Advertisement	86.45 ± 1.58%	94.97 ± 1.75%	95.53 ± 2.50%	97.21 ± 1.71%	**98.32 ± 1.06%**
Z-Alizadeh sani	**89.48 ± 0.88%**	88.91 ± 1.36%	87.86 ± 1.45%	88.14 ± 1.05%	**89.77 ± 1.70%**

5 Conclusions

Restricted Boltzmann Machines (RBMs) are effectively probability graph models for representational learning. On this basis, this paper extends RBMs to deal with multi-view learning and names it as RBMs with posterior consistency (PCRBMs). PCRBMs utilize the negative of the distance between two conditional probabilities to measure the posterior consistency between two views and maximize this posterior consistency. Then, this paper proposes correlation RBMs with posterior consistency (Exp-PCRBMs), which are suitable for binary and real-valued data. In addition, activation functions of Exp-PCRBMs can be any smooth monotonic non-linearity function. Finally, experimental results show that Exp-PCRBM is effective multi-view classification method for two-class and multi-class data.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant no. 61672522 and no. 61379101.

References

1. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
2. Zhang, N., Ding, S., Zhang, J., Xue, Y.: An overview on restricted Boltzmann machines. Neurocomputing **275**, 1186–1199 (2018)
3. Courville, A., Desjardins, G., Bergstra, J., Bengio, Y.: The spike-and-slab RBM and extensions to discrete and sparse data distributions. IEEE Trans. Pattern Anal. Mach. Intell. **36**(9), 1874–1887 (2014)
4. Mittelman, R., Kuipers, B., Savarese, S., Lee, H.: Structured recurrent temporal restricted Boltzmann machines. In: Proceedings of International Conference on Machine Learning, ICML 2014, Beijing, China, pp. 1647–1655, 21–26 June 2014
5. Zhang, N., Ding, S., Zhang, J., Xue, Y.: Research on point-wise gated deep networks. Appl. Soft Comput. **52**, 1210–1221 (2017)
6. Nguyen, T.D., Tran, T., Phung, D., Venkatesh, S.: Graph-induced restricted Boltzmann machines for document modelling. Inf. Sci. **328**, 60–75 (2016)
7. Amer, M.R., Shields, T., Siddiquie, B., Tamrakar, A., Divakaran, A., Chai, S.: Deep multimodal fusion: a hybrid approach. Int. J. Comput. Vis. **126**(2–4), 440–456 (2018)
8. Salakhutdinov, R.R., Hinton, G.E.: Deep Boltzmann machines. In: Proceedings of International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, pp. 448–455, 16–18 April 2009
9. Zhao, J., Xie, X., Xu, X., Sun, S.: Multi-view learning overview: recent progress and new challenges. Inf. Fusion **38**, 43–54 (2017)
10. Liu, Q., Sun, S.: Multi-view regularized gaussian processes. In: Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (eds.) PAKDD 2017, Part II. LNCS (LNAI), vol. 10235, pp. 655–667. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57529-2_51
11. Chao, G., Sun, S.: Consensus and complementarity based maximum entropy discrimination for multi-view classification. Inf. Sci. **367**, 296–310 (2016)

12. Ravanbakhsh, S., Póczos, B., Schneider, J., Schuurmans, D., Greiner, R.: Stochastic neural net-works with monotonic activation functions. In: Proceedings of International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, pp. 809–818, 9–11 May 2016
13. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Comput. **14**(8), 1711–1800 (2002)
14. Ding, S., Zhang, X., An, Y., Xue, Y.: Weighted linear loss multiple birth support vector machine based on information granulation for multi-class classification. Pattern Recogn. **67**, 32–46 (2017)