Sparse Restricted Boltzmann Machine Based on Data Class Entropy

Guobin Zhang1, Xiaoli Li2, Xiaoguang Li3, *
1,2,3School of Liaoning University, Shenyang, Liaoning, China
*Corresponding author e-mail: xgli@lnu.edu.cn

Abstract. The sparse of RBM is a critical technique to prevent the over-fitting of neural networks. Current competition sparse methods suffered from the fix sparse threshold. To this point, an adaptive RBM sparse method based on the complexity of data class distribution is proposed to determine the sparse threshold adaptively. In the method, a hidden-layer based entropy computing is introduced to overcome the curse of dimensionality for the estimation of distribution. Experiments on MNIST and CIFAR-10 are conducted and show that the proposed method is more effective than CaRBM and GC-RBM.

1. Introduction

Restricted Boltzmann machine (RBM) has become an important basic algorithm for deep learning[1]. It is widely used in deep belief network, image processing and so on[2]. However, there is an over-fitting phenomenon in RBM training, which limits RBM application[3]. To prevent over-fitting of neural networks, researchers have proposed many solutions, such as regularization methods and sparse methods[4,5]. The sparse of RBM mainly includes two methods: penalty sparse and competition sparse[6]. Penalty sparse uses sparse constraint as penalty function and integrates the function into the objective optimization[7]. In comparison with penalty sparse, competition sparse is achieved by competition among hidden units with sparse constrains, such as CaRBM and Ga-RBM[8,9]. The problem of competition sparse is that the sparse threshold cannot be adapted to the input data distribution.

The purpose of this article is to configure an appropriate threshold based on the distribution of input data, we propose an adaptive method based on input data, called as Entropy Cardinality Based Restricted Boltzmann Machine (EC-RBM). The sparse threshold is measured by the information entropy of a given data in EC-RBM[10-12]. The basic idea of EC-RBM is that fewer neurons need to be activated in the case that there is little information need to be represented, and vice versa, more activated neurons are required. To overcome the curse of dimensionality when estimating the
distribution of data class, EC-RBM introduces that the information entropy is computed by the hidden layer. Experiments show that EC-RBM is superior to CaRBM and GC-RBM.

2. Related Works

Restricted Boltzmann Machine (RBM) is an energy model of Markov random field with a two-layer structure[13,14]. The $N_v$ dimensional visible layer units $v$ are used to input the original data, and the $N_h$ dimensional hidden layer random units $h$ are used to extract features of the original input data. Visible layer units $v$ are connected to the hidden layer units $h$, and the joint probability distribution of the units $v$ and $h$ are shown in Eq. 1:

$$P(v, h) = \frac{1}{Z} \exp(v^TWh + vb + h^Td)$$ (1)

where $Z$ is the partition function, and the parameters meet the criteria $W \in \mathbb{R}^{N_v \times N_h}, b \in \mathbb{R}^{N_v}, d \in \mathbb{R}^{N_h}$. $W$ represents weight between the visible layer and hidden layer, $b$ and $c$ represent the bias terms of the visible layer units and the hidden layer units, respectively[15-17].

CaRBM proposes an approach to sparse the hidden layer. In CaRBM, the sparse activation threshold is introduced into the joint distribution of visible layer variables and hidden layer variables. However, the activation threshold breaks the assumption of the conditionally independence of hidden layer units implied by RBM. Although CaRBM can sparse RBM and improve the performance in some sense, the sparse threshold is a fixed constant and configured by experience. The sparse does not adapt to the pattern of data. Compared to CaRBM, GC-RBM is a generalized model with adaptive sparsity constraint. The threshold of hidden unit activations is decided by the input data and a given Gaussian distribution in the pre-training phase. Generally, sparse optimization in CaRBM is intuitive, simple, and easy to calculate, but the threshold cannot be adapted to input data. GC-RBM model improves the adaptability, but does not intuitively reflect different input data[18-20].

3. Entropy Cardinality Restricted Boltzmann Machine

This paper proposes an RBM sparse method based on the complexity of data distribution, called Entropy Cardinality Restricted Boltzmann Machine (EC-RBM). EC-RBM uses information entropy to measure the complexity of data distribution for a given class, and gives a probabilistic approach to adaptively configure the activated threshold. Moreover, three kinds of neuron competitions are designed for different distributions of activated units.

3.1 EC-RBM Model

In practice, the complexity of the data distribution is quite different in terms of the classes of data. Figure 1 depicts the information entropy of the number '0' to '9' of MNIST dataset. It shows that the entropy of the number '1' is much larger than the number '2'. It may be out of the usual thought that the pattern of the number '1' should be simple for its simple stroke. However, the actual writing shape of '1' is of more types than other numbers. Naturally, it is expected that the code length of the number '1' is longer than the number '2'. It means that for the more complex data, the more hidden neurons should be activated to encode enough information to distinguish the patterns.

For the visible units $v \in \mathbb{R}^{N_v}$ and the hidden units $h \in \mathbb{R}^{N_h}$, the joint probability of EC-RBM with the threshold $k_v$ about the number of activated units is given by Eq. 2.
where for the data \( v \), \( H_c(v) \) is the entropy of the data class \( c \), which represents the information about every uncertain subclass. Then, the specific threshold \( k_{H_c(v)} \) is determined by \( H_c(v) \). Here, we represent \( k_{H_c(v)} \) by \( k_v \) in briefly. Given input data \( v \) and its subclass entropy \( H_c(v) \), we assume that \( k_v \) follows the probability of Eq. 3.

\[
P(k_v|h, H_c(v)) = \frac{P(v|H_c(v)|h)P(h)P(k_v)}{P(v|H_c(v))P(h)} = \frac{P(v|k_v)P(H_c(v)|k_v)P(k_v)}{P(v|k_v)P(k_v|H_c(v))} = \frac{P(k_v|v)P(k_v|H_c(v))}{P(k_v)P(H_c(v))}
\]

Related inference is similar to that in CaRBM and GC-RBM. (2) \( P(k_v) \) is the prior knowledge of \( k_v \), which cannot be obtained directly. So, \( P(k_v) \) is assumed to obey the uniform distribution as GC-RBM did. (3) \( P(k_v|H_c(v)) \) represents the posterior knowledge about \( k_v \) when the subclass entropy introduced. Eq. 6 shows the calculation progress of \( P(k_v|H_c(v)) \), \( P(k_v|v) \) and \( k_v \) can be sampled from specific distribution.

Suppose that \( \mathcal{D} \) is the data space, and \( C \) is the class space, a training data is the set \( \mathcal{D} \subseteq \mathcal{D} \subseteq \mathcal{C} \). For \((v, c) \in \mathcal{D}, v \in \mathcal{R} \) is the input data, \( c \in C \) is the specific class of input data \( v \), \( N_c = |C| \). For the class \( c \), \( D_c \subseteq \mathcal{D} \) is a subset of training data of the class \( c \) in \( D \), \( n_c = |D_c| \). \( p(v, c) \) is defined as the joint probability of data \( (v, c) \). The entropy of the class \( c \) is defined by Eq. 4.

\[
H(c) = -\sum_x p(v, c)logp(v, c) = -p(c)\sum_x p(v|c)logp(v|c)p(c)
\]

The probability of \( p(c) \) can be obtained directly from the training dataset. For \( p(c) \), Parzen Window method is used to estimate the conditional probability \( p(v|c) \). Here, Gaussian kernel is selected as the window function. So, the \( p(v|c) \) is replaced by \( \hat{p}(v|c) \):

\[
\hat{p}(v|c) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{d(v, v_i)}{h_N}\right)
\]

where \( \varphi(\cdot) \) is the Gaussian kernel, \( d(v, v_i) \) is the distance between \( v \) and \( v_i \), \( v_i \) is the sample drawn from the class \( c \), \( h_N \) is the width of window, \( V_N \) is the volume of window, \( N \) is the total number of samples. \( P(k_v|H_c(v)) \) is assumed to follow the Gaussian distribution \( N(\hat{H}_c, \sigma_c) \). \( \hat{H}_c \) is calculated by Eq. 4 and Eq. 5 for each class, and \( \sigma_c \) is set as 1.

The distance between the training data is critical to the class entropy \( \hat{H}_c \). It is well-known that the distance measurements maybe invalid for the high-dimensional data because of the "curse of dimensionality". That is to say, for high dimensional data, it is quite hard to distinguish input data each other. In fact, it is observed that the hidden layer is relevant to the visible layer, and the dimensionality of hidden layer is far lower than the visible layer. In this paper, the entropy of hidden layer substitutes for the entropy of visible layer. According to entropy properties, we have

\[
H(c) = H(X, c) = H(h, c) - log|\text{det}(W)|
\]

\[
H(h, c) = -\sum_{h \in H} p(h|c)logp(h|c)
\]

\[
\hat{p}(h|c) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{d(h, h_i)}{h_N}\right)
\]
where $\text{det}(W)$ is the determinant of weight matrix $W$. $\mathfrak{S}$ is the hidden layer output space of $X$ given $W$ and $b$.

4. Experiment

4.1 Setting

The experiments are conducted on the datasets of MNIST and CIFAR-10. EC-RBM is compared to typical sparse methods including CaRBM and GC-RBM. In the experiment, the parameters of the models are set as follows: for gauss kernel of Parzen window, $\mu = 0$, $\sigma = 1$, $N = 0.5*|V|$ [21]; for $P(k_v|v)$, $\mu = 20.0$, $\sigma = 3.0$. The $k$ in CaRBM is 20.

4.2 Experiment Evaluation

Figure 1 of MNIST dataset shows that in comparison with CaRBM and GC-RBM, EC-RBM has the lowest loss, which is reduced by 20% and 12%. At the same time, EC-RBM achieves better performance with the accuracy increasing 10% and 5%. The loss and accuracy of EC-RBM computed by the hidden layer are not greatly improved compared with the input layer. The reason is that the dimension of MNIST data is 28*28, which is much lower than CIFAR-10, so the curse of dimensionality has less impact on the Parzen window. For CIFAR-10, EC-RBM has a significant improvement compared with CaRBM and GC-RBM, with the loss reduced by 2.0% and 1.5%, and the accuracy improved by 9% and 5%. The variance of EC-RBM after 150 epochs is much smaller than the one of CaRBM and GC-RBM. Furthermore, the loss of EC-RBM hidden is lower than EC-RBM visual thorough the whole epochs, because the dimension of CIFAR-10 data is about 4 times of MNIST data and the distance measure in Parzen window becomes more valid for the hidden layer.

5. Conclusion

In this paper, we propose a sparse model EC-RBM based on data class entropy. Within the EC-RBM, the size of activated neurons is dynamically determined by entropy. EC-RBM also introduces the hidden layer-based entropy to overcome the curse of dimensionality. Experiments show that EC-RBM has better performance than CaRBM and GC-RBM.
References

[1] Abadi, Martín, et al. Tensorflow: a system for large-scale machine learning. OSDI. Vol. 16. 2016.

[2] Bengio, Yoshua, et al. Greedy layer-wise training of deep networks. Advances in neural information processing systems. 2007.

[3] Bengio, Yoshua. Learning deep architectures for AI. Foundations and trends® in Machine Learning 2.1 (2009): 1-127.

[4] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. science313.5786 (2006): 504-507.

[5] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. Neural computation 18.7 (2006): 1527-1554.

[6] Lee, Honglak, Chaitanya Ekanadham, and Andrew Y. Ng. Sparse deep belief net model for visual area V2. Advances in neural information processing systems. 2008.

[7] Luo, Heng, et al. Sparse Group Restricted Boltzmann Machines. AAAI. 2011.

[8] Swersky, Kevin, et al. Cardinality restricted boltzmann machines. Advances in neural information processing systems. 2012.

[9] Wan, Cheng, et al. Gaussian Cardinality Restricted Boltzmann Machines. AAAI. 2015.

[10] Zhang, Xiao, et al. Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. Pattern Recognition 56 (2016): 1-15.

[11] Wang, Handing, and Xin Yao. Objective reduction based on nonlinear correlation information entropy. Soft Computing 20.6 (2016): 2393-2407.

[12] Tan, Yong-zhong, and Ci-fang Wu. The laws of the information entropy values of land use composition. Journal of natural resources 1 (2003): 017.

[13] Tarlow, Daniel, et al. Fast exact inference for recursive cardinality models. arXiv preprint arXiv:1210.4899 (2012).

[14] Hinton, Geoffrey E. Training products of experts by minimizing contrastive divergence. Neural computation 14.8 (2002): 1771-1800.

[15] Tarlow, Daniel, Inmar Givoni, and Richard Zemel. Hop-map: Efficient message passing with high order potentials. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010.

[16] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. Replicated softmax: an undirected topic model. Advances in neural information processing systems. 2009.

[17] Krizhevsky, Alex, and Geoffrey Hinton. Learning multiple layers of features from tiny images. Vol. 1. No. 4. Technical report, University of Toronto, 2009.

[18] Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. Restricted Boltzmann machines for collaborative filtering. Proceedings of the 24th international conference on Machine learning. ACM, 2007.

[19] Srivastava, Nitish, Ruslan R. Salakhutdinov, and Geoffrey E. Hinton. Modeling documents with deep boltzmann machines. arXiv preprint arXiv:1309.6865 (2013).

[20] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature 521.7553 (2015): 436.

[21] YAN, Xiaobo, Shitong WANG, and Huiling GUO. Feature reduction of high-order statistics based On Parzen window. CAAI Transactions on Intelligent Systems 1 (2013): 002.