2J-01

非学習型確率層により多層化された制限ボルツマンマシン分類器の性能の検討

菅野 友理 †　　安田 宗樹 ‡
山形大学大学院理工学研究科 †　山形大学大学院理工学研究科 ‡

1. はじめに

パターン認識問題を解くための確率的なモデルである制限ボルツマンマシン分類器（discriminative restricted Boltzmann machine（DRBM））[1] は，その確率的な構造上，多層化拡張後の学習が容易でなくなるという問題を抱えている．その問題を解決するために，非学習型確率層である確率的極端学習機械（probabilistic extreme leaning machine（PELM））層による多層化が提案されている [2]．非学習型確率層とは，システムの最適化学習前に事前設定する非学習型パラメータを持ち，さらにその出力を確率的に行う層である．PELM 層によって多層化された DRBM を multi-layered DRBM（MDRBM）と呼ぶ．

また，非学習型パラメータの決定にガウシアン・ベルヌーイ制限ボルツマンマシン（Gaussian-Bernoulli restricted Boltzmann machine（GBRBM））[3] を用いた教師なし学習による方法も提案されている [2]．しかし GBRBM の学習には厳密にできない期待値計算が含まれるため，学習状況の可視化が難しいという課題がある．そこで本研究では，GBRBM を用いた MDRBM の性能を複数のデータセットを用いて確認するとともに，より良質な非学習型パラメータを獲得する方法について検討する．

2. 非学習型確率層により多層化された制限ボルツマンマシン分類器

DRBM は制限ボルツマンマシン（restricted Boltzmann machine（RBM））[4][5] と呼ばれる確率モデルを基にした 3 層構造の分類器モデルである．制限ボルツマンマシン分類器は入力層 $\boldsymbol{x} = \{x_i \in (-\infty, +\infty) \mid i = 1, 2, \ldots, n\}$，中間層 $\boldsymbol{h} = \{h_j \in \{-1, +1\} \mid j = 1, 2, \ldots, H\}$，出力層 $\boldsymbol{t} = \{t_k \in \{0, 1\} \mid k = 1, 2, \ldots, K; \sum_{k=1}^{K} t_k = 1\}$ から構成される．これらを用いて DRBM は以下の条件付き確率によってモデル化される．

$$P(\boldsymbol{t}, \boldsymbol{h} \mid \boldsymbol{x}, \theta) := \frac{1}{Z(\boldsymbol{x}, \theta)} \exp\left(\sum_{k=1}^{K} b_k^{(2)} t_k + \sum_{j=1}^{H} b_j^{(1)} h_j + \sum_{k=1}^{K} \sum_{j=1}^{H} w_{k,j}^{(2)} t_k h_j + \sum_{j=1}^{H} \sum_{i=1}^{n} w_{j,i}^{(1)} h_j x_i \right) \quad (1)$$

ここで $Z(\boldsymbol{x}, \theta)$ は規格化定数であり，$\theta = \{\boldsymbol{b^{(1)}}, \boldsymbol{b^{(2)}}, \boldsymbol{w^{(1)}}, \boldsymbol{w^{(2)}}\}$ は学習パラメータの集合である．DRBM のクラス確率は，中間層の周辺化 $P(\boldsymbol{t} \mid \boldsymbol{x}, \theta) = \sum_{\boldsymbol{h}} P(\boldsymbol{t}, \boldsymbol{h} \mid \boldsymbol{x}, \theta)$ を実行して計算する．

PELM 層は，極端学習（extreme leaning machine（ELM））[6] の非学習型パラメータを用いるという考えを基にして設計された確率層である．PELM 層ではまず，非学習型パラメータを用いて素子が取る値の確率を計算し，その確率に従って値を決定する．PELM 層 $\boldsymbol{z} = \{z_j \in \{-1, +1\} \mid j = 1, 2, \ldots, \mathcal{H}\}$ の条件付き確率を以下のベルヌーイ分布として定義する．

$$B(\boldsymbol{z} \mid \boldsymbol{x}, \theta_0) := \prod_{j=1}^{\mathcal{H}} \frac{\exp(b_j^{(0)} + \sum_{i=1}^{n} w_{j,i}^{(0)} x_i z_j)}{2\cosh(b_j^{(0)} + \sum_{i=1}^{n} w_{j,i}^{(0)} x_i)} \quad (2)$$

ここで，$\theta_0 = \{\boldsymbol{b^{(0)}}, \boldsymbol{w^{(0)}}\}$ は非学習型パラメータの集合である．$\boldsymbol{z}$ は式 (2) で計算された確率に従って -1 か $+1$ の値を確率的に取る．図 1 が，PELM 層によって多層化された DRBM である MDRBM のイメージ図である．MDRBM の条件付き確率は期待

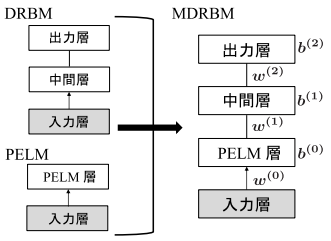

図 1　MDRBM のイメージ図.

値を用いて

$$P^{\dagger}(\boldsymbol{t} \mid \boldsymbol{x}, \theta, \theta_0) = \sum_{\boldsymbol{z}} P(\boldsymbol{t} \mid \boldsymbol{z}, \theta) B(\boldsymbol{z} \mid \boldsymbol{x}, \theta_0) \quad (3)$$

のように計算される．ここで，式 (3) 中の和は解析的に計算をすることができないため，モンテカルロ積分法を用いて

$$P^{\dagger}(\boldsymbol{t} \mid \boldsymbol{x}, \theta, \theta_0) \approx \frac{1}{S} \sum_{\nu=1}^{S} P\left(\boldsymbol{t} \mid \boldsymbol{z}^{(\nu)}, \theta\right) \quad (4)$$

のように近似計算する．$z^{(1)}, z^{(2)}, \ldots, z^{(S)}$ は $B(\boldsymbol{z} \mid \boldsymbol{x}, \theta_0)$ から生成した S 個のサンプル点であり，このサンプリングは容易に実行することが可能である．

MDRBM において，学習パラメータである θ は最尤法を用いて最適化される．それに対し非学習型パラメータである θ_0 は，何らかの方法で事前決定されるパラメータである．PELM 層は入力データの特徴量変換の役割を果たすため，この非学習型パラメータを良質なものに設定できるかどうかは，最終的な MDRBM の認識性能に大きく関わる．

3. 非学習型パラメータの決定方法

非学習型パラメータは最尤法による最適化以前に決定されるパラメータであり，通常の ELM ではこのパラメータはランダムに決定する．これに対し，ランダムではなく GBRBM によって学習したパラメータを用いることで，ELM がより高い認識精度

Investigation of Multi-layered Discriminative Restricted Boltzmann Machine with Untrained Probabilistic Layer
† Yuri Kanno; Graduate School of Science and Enginnering, Yamagata University
‡ Muneki Yasuda; Graduate School of Science and Enginnering, Yamagata University

表 1　ノイズ加算時のテストデータに対する認識率及び ADR（MNIST）.

	noise level σ						
	0	0.2	0.4	0.6	0.8	1	ADR
MDRBM	**94.5**	**94.4**	**94.3**	**94.0**	**93.5**	**92.9**	**1.8**
DRBM	90.2	90.1	89.2	87.5	85.1	81.6	9.5
RBM-ELM	87.6	87.3	86.6	85.6	83.9	81.7	6.7
4NN	92.4	92.2	91.5	90.2	88.1	85.0	8.0

表 2　ノイズ加算時のテストデータに対する認識率及び ADR（Fashon-MNIST）.

	noise level σ						
	0	0.2	0.4	0.6	0.8	1	ADR
MDRBM	85.9	85.8	**85.5**	**85.0**	**84.3**	**83.3**	**3.1**
DRBM	85.6	85.3	84.0	81.9	79.0	75.9	11.3
RBM-ELM	81.5	81.2	80.3	79.0	77.1	74.7	8.3
4NN	**86.6**	**86.2**	85.2	83.9	81.2	78.3	9.6

表 3　ノイズ加算時のテストデータに対する認識率及び ADR（CIFAR-10）.

	noise level σ						
	0	0.2	0.4	0.6	0.8	1	ADR
MDRBM	30.3	30.2	30.0	30.0	**29.7**	**29.3**	**3.1**
DRBM	27.3	27.0	26.4	25.1	23.9	22.7	16.8
RBM-ELM	22.5	22.2	21.9	21.1	20.4	19.6	12.9
4NN	**33.3**	**32.8**	**31.9**	**30.3**	28.7	27.1	18.6

を持つことが知られている [7]．そのため，MDRBM に対しても GBRBM による非学習型パラメータの決定を行う．GBRBM は可視層と隠れ層の 2 層からなる確率的なモデルである．GBRBM の学習は最尤法を用いて行われるが，最尤法における勾配には厳密計算することができない期待値近似が含まれる．その期待値近似には一般によく contrastive divergence（CD）法 [5] という方法が用いられており，本研究でも CD 法による近似を行う．

4. 数値実験

MNIST，Fashon-MNIST, CIFAR-10 と呼ばれるデータセットを用いて数値実験を行った．本実験では各データセットからからランダムに抽出した訓練データをそれぞれ 6000，6000，3000 個を用いて学習し，テストデータ 10000 個に対して認識率を測定した．また CIFAR-10 は BT.601 によってグレースケール化した．

モデルの設定として，入力層の素子数は，MNIST，Fashon-MNIST, CIFAR-10 ごとに，それぞれに対応した値として $n = 784$，784，1024 とした．また，すべてのデータセットに共通して，中間層の素子数を $H = 500$，PELM の素子数を $\mathcal{H} = 500$，出力層の素子数を $K = 10$ とした．さらに，モンテカルロ積分のサンプル数は訓練時は $S = 5$，推論時は $S = 50$ とした．その他の実験設定として，入力データは前処理として標準化を行い，学習方法として Adam[9] と確率的勾配法を用いた．確率的勾配法におけるミニバッチサイズは 100 とし，学習するパラメータの初期値は Xavier の初期値 [10] で与えた．また，学習アルゴリズムは文献 [2] の方法に従った．

表 1，表 2，表 3 はそれぞれのデータセットに対する，ノイズが加算されたテストデータの認識率を示したものである．ここではノイズとして，テストデータに平均 0, 分散 σ^2 のガウスノイズを加算した．noise level σ はガウスノイズの標準偏差の値であり，$\sigma = 0$ の際はテストデータにノイズは付加されておらず，σ の値が大きくなるにつれてノイズも大きくなる．また，ADR は $\sigma = 0$ と $\sigma = 1$ の認識率の変化率であり，小さいほどノイズ耐性が高いと言える指標である．比較対象として，GBRBM を用いて非学習型パラメータを決めた ELM（RBM-ELM）[7] と，活性化関数に ReLU を使用した 4 層のニューラルネットワーク (4NN) を用いた．表 1，表 2，表 3 において MDRBM は，ノイズが大きくなっても認識率が下がらず，ロバスト性（あるいはノイズに対する頑健性）を獲得していることがわかる．

5. 今後の方針

4 節の実験では GBRBM の期待値近似に CD 法を用いているが，これよりも高精度な近似法として spatial Monte Carlo integration（SMCI）法 [8] という方法が知られている．また，GBRBM の学習の度合いを測定する基準を導入することで，より良質なパラメータを求めることが可能になると考えられる．講演ではこれらの手法についても述べる．

6．まとめ

本稿では PELM 層によって多層化された DRBM の性能を数値実験によって確かめるとともに，より良質な非学習型パラメータを求めるための方針を考えた．GBRBM による非学習型パラメータの決定には改良の余地があり，更なる研究が必要である．

謝辞

本研究は科研費（18K11459，18H03303），JST-CREST（JPMJCE1312）及び JST COI プログラム（JPMJCE1312）の助成を受けたものである．

文献

[1] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio: Learning algorithms for the classification restricted boltzmann machine, The Journal of Machine Learning Research, vol. 13, no. 1, pp. 643–669, 2012.
[2] Yuri Kanno and Muneki Yasuda: Multi-layered Discriminative Restricted Boltzmann Machine with Untrained Probabilistic Layer, In Proc. of the 25th International Conference on Pattern Recognition, 2020.
[3] K. Cho, A. Ilin, and T. Raiko: Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines, In Proc. of the 12th International Conference on Artificial Neural Networks, pp. 10-17, 2011.
[4] P. Smolensky: Information processing in dynamical systems: foundationsof harmony theory, Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1, pp. 194-281, 1986.
[5] G. E. Hinton: Training products of experts by minimizing contrastive divergence, Neural computation, vol. 14, no. 8, pp. 1771-1800, 2002.
[6] Guang-Bin Huang, Qin-Yu Zhu and Chee-Kheong Siew: Extreme Learning Machine: Theory and applications, Neurocomputing, vol. 70, pp. 489-501, vol.70, 2006.
[7] A. G. Pacheco, R. A. Krohling, and C. A. da Silva: Restricted boltzmann machine to determine the input weights for extreme learning machines, Expert Systems with Applications, vol. 96, pp. 77-85, 2018.
[8] M. Yasuda and K. Uchizawa: A generalization of spatial Monte Carlo integration, arXiv:2009.02165, 2020.
[9] Diederik P. Kingma and Jimmy Ba: Adam: A Method for Stochastic Optimization, In Proc. of the 3rd International Conference on Learning Representations, pp. 1-13, 2015.
[10] Glorot, Xavier, and Yoshua Bengio: Understanding the difficulty of training deep feedforward neural networks, In Proc. of the 13th International Conference on Artificial Intelligence and Statistics, vol. 9, pp. 249-256, 2010.