1M-03

スパース中間層を持つ制限ボルツマンマシン分類器

勝亦 利宗 †
山形大学大学院理工学研究科 †

安田 宗樹 ‡
山形大学大学院理工学研究科 ‡

1 はじめに

機械学習の分野において過学習 (overfitting) の問題は避けることのできない問題であり, より高い汎化能力を持つ学習モデルの研究が必要とされている. 本稿では制限ボルツマンマシン (restricted Boltzmann machine; RBM)[1] を分類問題に特化したモデルである制限ボルツマンマシン分類器 (discriminative RBM; DRBM)[2] に着目し, DRBM の中間層を連続値に拡張することによって汎化能力の向上が見られた先行研究 [3] を踏まえ, 新たに中間層にスパース性を持たせた学習モデルを提案する. スパース中間層を持った DRBM は問題に応じて自動的に中間層の素子数を調整できるようになり, 問題に柔軟に対応することが可能となる. また, 提案したモデルで数値実験を行い, 提案モデルの性能を測定する.

2 制限ボルツマンマシン分類器

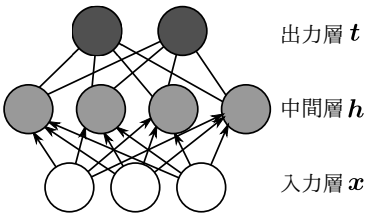

図 1 DRBM のグラフ構造.

DRBM のグラフ構造を図 1 に示す. DRBM は入力層 $\boldsymbol{x} = \{x_i \in (-\infty, \infty) \mid 1 \leq i \leq X\}$, 中間層 $\boldsymbol{h} = \{h_j \in \mathcal{X}_h \mid 1 \leq j \leq H\}$, 出力層 $\boldsymbol{t} = \{t_k \in \{0, 1\} \mid 1 \leq k \leq T\}$ の 3 層からなる構造を持っている. 元論文 [2] では中間層の取る値は 2 値変数であるが, 本稿では文献 [3] を参考にし, $\mathcal{X}_h = [-1, +1]$ とする. また, 出力層 $\boldsymbol{t}$ は 1-of-K 表現ベクトルとする. 1-of-K 表現ベクトルとはベクトルの要素のうち 1 つが 1 を取り, その他の要素は 0 を取るベクトルである. k 番目の要素のみが 1 である 1-of-K ベクトルを $\mathbf{1}_k$ のように表す. このベクトルを使って, 分類問題における教師データを表す.

入力層と出力層はデータの説明変数と目的変数に対応する層であり, その素子数 X と T はそれぞれ説明変数の次元数と目的変数のカテゴリ数によって定まる. また, 中間層はデータに直接対応しない層であり, 素子数 H は問題に応じて適切な数を設定する必要がある. 中間層の素子数が多いほどモデルの表現能力は増し, 入力変数と出力変数の複雑な関係性を表すことができるようになる一方で, より過学習を増長させる原因にもなる.

DRBM の確率分布関数は, エネルギー関数

$$E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{t}; \theta) = -\sum_{j=1}^{H} b_j^{(1)} h_j - \sum_{k=1}^{T} b_k^{(2)} t_k - \sum_{j=1}^{H}\sum_{i=1}^{X} w_{i,j}^{(1)} x_i h_j - \sum_{k=1}^{T}\sum_{j=1}^{H} w_{j,k}^{(2)} h_j t_k \tag{1}$$

を使って,

$$P(\boldsymbol{t}, \boldsymbol{h} \mid \boldsymbol{x}, \theta) = \frac{1}{Z(\theta)} \exp(-E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{t}; \theta)) \tag{2}$$

と定義される. ここで, $\theta = \left\{\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}\right\}$ はモデルのパラメータである. それぞれ, $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}$ はバイアスパラメータ, $\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}$ は層間の結合パラメータを表す. また, $Z(\theta)$ は分配関数である.

3 スパース正則化 制限ボルツマンマシン分類器

スパース正則化 DRBM(Sparse DRBM; SDRBM) を, 式 (3) にスパース正則化項を加え,

$$P(\boldsymbol{t}, \boldsymbol{h} \mid \boldsymbol{x}, \hat{\theta}) = \frac{1}{Z(\hat{\theta})} \exp\left\{-E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{t}; \theta) - \sum_{j=1}^{H} \sigma(\gamma_j)|h_j|\right\} \tag{3}$$

と定義する. ここで, 関数 $\sigma(\gamma_j)$ を

$$\sigma(\gamma_j) := \ln(\exp(\gamma_j) + 1)$$

とする. この関数はソフトプラス関数として知られている関数であり, 常に正の値を取る. また, γ_j はモデルに新たに加わるパラメータである. このパラメータと前節のパラメータ θ を合わせ, $\hat{\theta} = \left\{\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \boldsymbol{\gamma}\right\}$ とする. 新たに加わった項は常に負の値を取り, γ_j が大きいとき j 番目の中間素子の値が確率分布関数の出力に与える影響を減少させる. 適切な γ_j の値は学習によって調節するため, 中間素子の取る値をどれだけ軽視するかの判断をモデル自体が行うことができ, それが不要な素子を自動的に削減する仕組みとなっている. さらに, $\gamma_j \to -\infty$ のとき, 確率分布関数は 2 節で説明した DRBM と同じものとなる.

4 数値実験

4.1 人工データを用いた数値実験

この節ではバイアスパラメータを 0, 結合パラメータを Xavier の方法 [4] で初期化したパラメータ θ_{gen} から定められる 2 節の DRBM $P_{\text{gen}}(\mathbf{1}_k \mid \boldsymbol{x}, \theta_{\text{gen}})$ を生成モデルとし, この生成モデルからサンプリングしたデータを, 元論文 [2] の中間

Discriminative restricted Boltzmann machine with sparse hidden layer
† Tomu Katsumata, Graduate School of Science and Engineering, Yamagata University
‡ Muneki Yasuda, Graduate School of Science and Engineering, Yamagata University

Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.

層が 2 値変数である従来の DRBM, 中間層が連続値である 2 節の DRBM[3], スパース正則化 DRBM の 3 つの学習モデルで学習し, 比較を行う. 学習モデルと生成モデルはどちらも確率分布関数であるので, その差はカルバックライブラー情報量 (Kullback-Leibler divergence; KLD) で表すことができる. 学習モデル $P_{\mathrm{fit}}(\mathbf{1}_k \mid \boldsymbol{x}, \theta_{\mathrm{fit}})$ と生成モデル $P_{\mathrm{gen}}(\mathbf{1}_k \mid \boldsymbol{x}, \theta_{\mathrm{gen}})$ 間の, 入力データ $\boldsymbol{x}$ に対する KLD は

$$K(\boldsymbol{x}) = \sum_{k=1}^{T} P_{\mathrm{gen}}(\mathbf{1}_k \mid \boldsymbol{x}, \theta_{\mathrm{gen}}) \ln \frac{P_{\mathrm{gen}}(\mathbf{1}_k \mid \boldsymbol{x}, \theta_{\mathrm{gen}})}{P_{\mathrm{fit}}(\mathbf{1}_k \mid \boldsymbol{x}, \theta_{\mathrm{fit}})} \quad (4)$$

となる. この実験では $K(\boldsymbol{x})$ の平均を, モンテカルロ積分により近似する. また, 生成モデルの構造を $X = 20, H = 100, T = 10$, 学習モデルの構造を $X = 20, H = 200, T = 10$ とし, より過学習が起こりやすい状況で比較を行う. 学習モデルのパラメータ $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}$ の初期値は生成モデルと同様の方法で決定し, スパース正則化 DRBM のパラメータ $\boldsymbol{\gamma}$ の初期値は 10 とする. $\boldsymbol{\gamma}$ が大きな状態から学習を始めることで, 必要最低限の素子のみが有効になると考えられるためである. また, $\boldsymbol{\gamma}$ 以外のパラメータの更新は Adamax[6] を使用し, $\boldsymbol{\gamma}$ のみ確率的勾配法を使用した. 実験の結果を図 2 に示す.

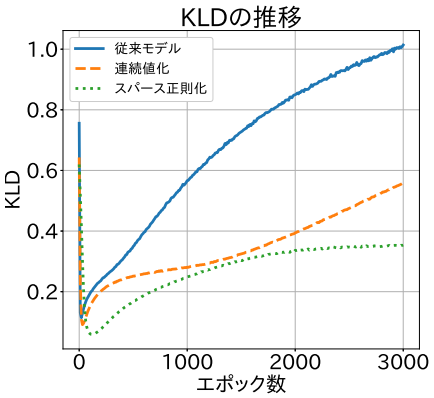

図 2 DRBM(実線)[2], 連続値化 DRBM(破線)[3], SDRBM(点線) の KLD の推移. プロットは 100 回の試行の平均である.

図 2 から, 提案モデルは最も KLD の小さい学習モデルであり, 従来のモデルよりも汎化能力が向上していると言える.

4.2 実データを用いた数値実験

この節では, MNIST の手書き数字画像のデータセットを使用し, 実データに対する数値実験を行う. MNIST は 28×28 のサイズを持つ 256 階調の画像データであり, 学習用のデータが 60,000 個, テスト用のデータが 10,000 個用意されている. この実験では過学習の傾向がより顕著に現れるよう, 学習データを 1,000 個に減らし, さらにテストデータに $\mathcal{N}(0, 120)$ のガウスノイズを加算した後に標準化し学習を行う. また, 中間層の素子数は 200 とし, 学習モデルのパラメータは前節と同様に初期化し, 更新する. 実験では学習データに対する誤認識率を訓練誤差, テストデータに対する誤認識率をテスト誤差とし, この 2 つの値を比較する. 実験の結果を図 3 に示す.

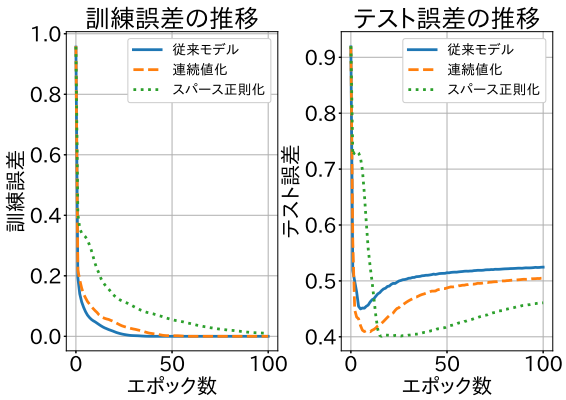

図 3 DRBM(実線)[2], 連続値化 DRBM(破線)[3], SDRBM(点線) の訓練誤差・テスト誤差の推移. プロットは 100 回の試行の平均である.

図 3 より, SDRBM は訓練誤差が上昇し, テスト誤差は減少していることが分かる. 過学習が起こるほど訓練誤差は減少し, テスト誤差は増加するため, SDRBM は過学習を抑制できていると考えられる.

5 まとめ

本稿では DRBM の確率分布にスパース正則化項を加えた新たな学習モデルを提案し, 従来の DRBM との比較を行った. 数値実験の結果, SDRBM は総じて従来の DRBM よりも優れた性能を示した. 今回の手法は隠れ素子を持つボルツマンマシンであればいずれも適用可能であると考えられるため, より隠れ層の数を増やしたモデルであるディープボルツマンマシン [5] 等においても, 同様に性能の向上が示されるかどうかが今後の課題として挙げられる.

参考文献

[1] D. Ackley, G. Hinton and T. Sejnowski: A learning algorithm for boltzmann machines, Cognitive science, 9(1):147–169, 1985.

[2] H. Larochelle and Y. Bengio: Classification using discriminative restricted boltzmannmachines, International Conference on Machine Learning, pp.536-543, 2008.

[3] Y. Yokoyama, T. Katsumata and M. Yasuda: Restricted Boltzmann Machine with Multivalued Hidden Variables: a model suppressing over-fitting, The Review of Socionetwork Strategies, DOI:10.1007/s12626-019-00042-4, pp.1-14, 2018.

[4] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp.249–256, 2010.

[5] R. Salakhutdinov and G. Hinton: Deep Boltzmann Machines, Artificial intelligence and statistics, pp.448-455, 2009.

[6] D. Kingma and J. Ba: Adam: a method for stochastic optimization, Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), 2015.

Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.