SCIENCE CHINA
Information Sciences

• LETTER •

March 2023, Vol. 66 139101:1–139101:2
https://doi.org/10.1007/s11432-020-3113-4

Neural network equivalent model for highly efficient massive data classification

Siquan YU[1,2,3], Zhi HAN[2,3]*, Yandong TANG[2,3] & Chengdong WU[4]

[1]*School of Information Science and Engineering, Northeastern University, Shenyang* 110819*, China;*
[2]*Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang* 110016*, China;*
[3]*State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang* 110016*, China;*
[4]*Faculty of Robot Science and Engineering, Northeastern University, Shenyang* 110819*, China*

Received 2 May 2020/Revised 30 June 2020/Accepted 29 September 2020/Published online 8 December 2022

Citation Yu S Q, Han Z, Tang Y D, et al. Neural network equivalent model for highly efficient massive data classification. Sci China Inf Sci, 2023, 66(3): 139101, https://doi.org/10.1007/s11432-020-3113-4

Dear editor,

Recently, neural networks (NNs) have been successfully applied to massive data classification tasks [1, 2]. However, there is an inconsistent problem in neural network structure and convergence [3]. On one hand, under the over-parameterize assumption, it requires to design a complex network to guarantee the convergence [4]. On the other hand, over-parameterization makes the neural networks flexible. Therefore, the networks are difficult to obtain the global optimal solution, and may lead to over-fitting [5].

To overcome the aforementioned problems, we establish a neural network equivalent model and apply it to massive data classification. We regard the structure of neural networks as a model of feature extraction and classification hyperplane solution. More specifically, our basic idea is to use the hidden layer of the neural network as a feature extractor. The hidden parameters of our model are directly generated by the minimal Riesz energy points on the sphere and the equidistant points on an interval [6]. The purpose of generating hidden parameters instead of training them is to reduce the computation of network learning, and overcome the inconsistency problem between optimization and learning. For solving classification hyperplane, we transform a maximal margin principle problem to a non-smooth convex optimization and adopt the ADMM algorithm to solve this problem. The reason for using ADMM is that it can guarantee the convergence of the optimization problem. Theoretically, we first analyze that ADMM can converge to the global minimum of the non-smooth convex optimization problem. Second, we justify that our model can achieve an almost optimal generalization error bound. Extensive experiments verify the effectiveness of the proposed model for massive data classification.

The proposed equivalent model. Given the training samples $D := \{(x_i, y_i)\}_{i=1}^m$, the feature space is

$$\mathcal{H}_{l,n,\sigma} := \left\{ \sum_{j=1}^{n} \sum_{k=1}^{l} a_{jk}\sigma(\alpha_j x - b_k) : a_{jk} \in \mathbb{R}^1 \right\}, \qquad (1)$$

where σ is a nonlinear function, $n \times l$ denotes splitting of N. The inner weights $\{\alpha_j\}_{j=1}^n$ is obtained by minimizing the Riesz τ-energy point of $\mathbb{S}^{d-1}$ with $\tau \geqslant d-1$, and $\{b_k\}_{k=1}^l$ is the ESPs in an interval. Then, the solution of classification hyperplane can be transformed to an empirical risk minimization problem:

$$f_D = \arg \min_{f \in \mathcal{H}_{l,n,\sigma}} \left\{ \frac{1}{m} \sum_{i=1}^{m} (1 - y_i f(x_i))_+ \right\}, \qquad (2)$$

where t_+ is the hinge loss function which is defined as $\max\{t, 0\}$ for $t \in \mathbb{R}$.

In our model, we adopt ADMM to solve the unregularized optimization problem (2). First, we reformulate the problem (2) as an unconstrained form and introduce a new variable ν:

$$\min_{\nu \in \mathbb{R}^n, \mu \in \mathbb{R}^m} f(\nu) \quad \text{s.t. } A\mu - \nu = 0, \qquad (3)$$

where $f(\nu) := \frac{1}{m}\sum_{i=1}^m (1 - y_i\nu_i)_+$, and $A \in \mathbb{R}^{m\times n}$ is the feature matrix with $A_{ij} = \sigma(\alpha_j x_i - b_{kj})$. The augmented Lagrangian function of (3) is as follows:

$$L_\beta(\mu, \nu, w) = f(\nu) + \langle \omega, A\mu - \nu \rangle + \frac{\beta}{2}||A\mu - \nu||_2^2, \qquad (4)$$

where $\beta > 0$ is the augmented Lagrangian coefficient, and $\omega \in \mathbb{R}^m$ is a multiplier variable. According to [7], given $\alpha > 0$, the closed-form solutions of ADMM of (4) are

$$\mu^{t+1} = (\beta A^{\mathrm{T}} A + \alpha I_n)^{-1}(\alpha\mu^t + \beta A^{\mathrm{T}}\nu^t - A^{\mathrm{T}}\omega^t), \qquad (5)$$

$$\nu^{t+1} = \mathrm{Hinge}_{m\beta}(y, A\mu^{t+1} + \beta^{-1}\omega^t), \qquad (6)$$

$$w^{t+1} = \omega^t + \beta(A\mu^{t+1} - v^{t+1}), \qquad (7)$$

where $\mathrm{Hinge}_\gamma(\zeta, \xi) = (\mathrm{hinge}_\gamma(\zeta(1), \xi(1)), \ldots, \mathrm{hinge}_\gamma(\zeta(\mathrm{m}), \xi(m)))$, $\zeta = (\zeta(1), \ldots, \zeta(m))^{\mathrm{T}}$ for $\zeta \in \mathbb{R}^m$, $\gamma > 0$ and

$$\mathrm{hinge}_\gamma(a, b) = \begin{cases} b, & \text{if } a = 0, \\ b + \gamma^{-1}a, & \text{if } a \neq 0 \text{ and } ab \leqslant 1 - \gamma^{-1}a^2, \\ a^{-1}, & \text{if } a \neq 0 \text{ and } 1 - \gamma^{-1}a^2 < ab < 1, \\ b, & \text{if } a \neq 0 \text{ and } ab > 1. \end{cases}$$

* Corresponding author (email: hanzhi@sia.cn)

Theoretical behaviors. The theoretical behaviors include the convergence issue of ADMM and the generalization performance analysis. We will analyze the convergence issue of ADMM in Appendix B. In the following, we mainly study the generalization error of the proposed model under the learning theory framework [8].

Consider the samples $D = \{(x_i, y_i)\}_{i=1}^{m}$ with the range of $x \in X = [0,1]^d$ and $y \in Y = \{-1, 1\}$. These samples are generated independently according to ρ and with the range of $Z := X \times Y$. We measure generalization ability by the misclassification error which is defined as follows:

$$\mathcal{R}(\mathcal{C}) = \boldsymbol{P}[\mathcal{C}(x) \neq y] = \int_X \boldsymbol{P}[y \neq \mathcal{C}(x)|x] \mathrm{d}\rho_X,$$

where the conditional probability is expressed as $\boldsymbol{P}[y|x]$ with $x \in X$. Thus, we can use the well known Bayes rule which is defined as

$$g_c(x) = \begin{cases} 1, & \xi(x) \geqslant 1/2, \\ -1, & \xi(x) < 1/2 \end{cases}$$

to minimize $\mathcal{R}(\mathcal{C})$, where $\xi(x)$ is the decision function $\boldsymbol{P}[y = 1|x]$. Define the classical Sobolev class by

$$W_p^r := W_p^r(\mathbb{B}^d) := \left\{ g : \mathbb{B}^d \to \mathbb{R}^1 : \max_{0 \leqslant |\boldsymbol{k}| \leqslant r} \|D^{\boldsymbol{k}} g\|_p < \infty \right\},$$

where $\mathbb{B}^d$ denotes a d-dimensional unit ball. Let J be the identity mapping:

$$L^2(\mathbb{B}^d) \xrightarrow{J} L^1_{\rho_X},$$

and $D_{\rho_X} = \|J\|$. Like [9], we assume $D_{\rho_X} < \infty$ and $f_\rho \in W_2^r$. We say σ is a sigmoid function, if σ satisfies

$$\lim_{z \to -\infty} \sigma(z) = 0 \quad \text{and} \quad \lim_{z \to \infty} \sigma(z) = 1.$$

Thus, for any σ, we can find a positive constant L which satisfy

$$\begin{cases} |\sigma(t) - 1| < m^{-\frac{2}{2r+d}}, & t \geqslant L, \\ |\sigma(t)| < m^{-\frac{2}{2r+d}}, & t \leqslant -L. \end{cases} \tag{8}$$

Let

$$\sigma_K(z) := \sigma(Kz),$$

for any

$$K \geqslant \ell L, \tag{9}$$

where ℓ denotes the number of different thresholds in the LtDaHP scheme. We further support that for an arbitrary closed set G in $\mathbb{R}^1$, σ is a local square integrable function, which is defined as $\sigma \in L^2_{\mathrm{Loc}}(\mathbb{R}^1)$.

Our main result on generalization error analysis for the algorithm is the following Theorem 1, of which we will provide the proof in Appendix C.

Theorem 1. Let $0 < \theta < 1$ and $d \geqslant 2$. Assume $0 < r \leqslant \frac{d+1}{2}$, $\sigma \in L^2_{\mathrm{Loc}}(\mathbb{R}^1)$ is a bounded sigmoid function. If $f_\rho \in W_2^r$, $\ell = [m^{\frac{1}{d+2r}}]$, $n \sim \ell^{d-1}$ and K satisfies (9), then with confidence at least $1 - \theta$, there holds

$$\mathcal{R}(\mathrm{sign}(f_{D,n,\ell,\sigma})) - \mathcal{R}(f_c) \leqslant C D_{\rho_X} (m/\log m)^{-\frac{r}{2r+d}} \log \frac{4}{\theta}, \tag{10}$$

where C denotes a positive constant independent of ℓ, m, n, K or θ.

Experimental results. Our experiments mainly include the toy simulations, the real world data experiments and the massive data experiment. The experimental results are provided in Appendixes D and E. In the massive data experiment, we conduct an experiment on the super symmetry particles (SUSY) dataset. We report the test accuracy (ACC) and training time (Time) in Table 1. Our method achieves the highest classification accuracy compared to the other methods. It should be noted that the first 10 algorithms are implemented on GPU while FPC and our model are implemented on a single CPU. Considering the hardware factors, our algorithm is also superior to the first 11 algorithms in training time.

Table 1 The classification accuracy and training time on real world massive dataset[a)]

Method	ACC (%)	Time (s)	Method	ACC (%)	Time (s)
kNN	67.5	1464	SGD-F	77.7	118
SGD	76.5	25	LB-SGD	77.7	1022.4
HT	78.2	45	HT-SGD	78.4	154.3
LB-HT	78.7	530	NN	82.7	1250.44
HT-kNN	77.2	1428	FPC	79.0	732.8
kNN-F	71.2	4714	Ours	**83.2**	604.2

a) The best result is highlighted in bold.

Conclusion. This study proposes a neural network equivalent model and applies it to massive data classification. In our model, we first generate the hidden parameters by using the minimal Riesz energy points on a sphere and equally spaced points in an interval. Then, we obtain a classifier via exploiting ADMM to optimize a non-smooth convex optimization problem. The effectiveness of our model is verified by both theoretical analysis and real world experiments.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61773367, 61903358, 61821005), Youth Innovation Promotion Association of the Chinese Academy of Sciences (Grant No. 2016183), and China Postdoctoral Science Foundation (Grant No. 2019M661154).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhou L, Yan Y, Qin X, et al. Deep learning-based classification of massive electrocardiography data. In: Proceedings of IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, 2016. 780–785
2 Zhang L, Zhang Y. Big data analysis by infinite deep neural networks. J Comput Res Dev, 2016, 53: 68–79
3 Wang D, Zeng J, Lin S B. Random sketching for neural networks with ReLU. IEEE Trans Neural Netw Learn Syst, 2021, 32: 748–762
4 Oymak S, Soltanolkotabi M. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. 2019. ArXiv:1902.04674
5 Kawaguchi K. Deep learning without poor local minima. Adv Neural Inf Process Syst, 2016, 5: 586–594
6 Fang J, Lin S, Xu Z. Learning through deterministic assignment of hidden parameters. IEEE Trans Cybern, 2020, 50: 2321–2334
7 Zeng J, Wu M, Lin S, et al. Fast polynomial kernel classification for massive data. 2019. ArXiv:1911.10558
8 Steinwart I, Christmann A. Support vector machines. Inf Sci Stat, 2008, 158: 1–28
9 Lin S B, Zhou D X. Distributed kernel-based gradient descent algorithms. Constr Approx, 2018, 47: 249–276