40 Years ICML International Conference On Machine Learning

A General Representation Learning Framework with Generalization Performance Guarantees

Junbiao Cui[1], Jianqing Liang[1], Qin Yue[1], Jiye Liang[1]

Speaker: Junbiao Cui

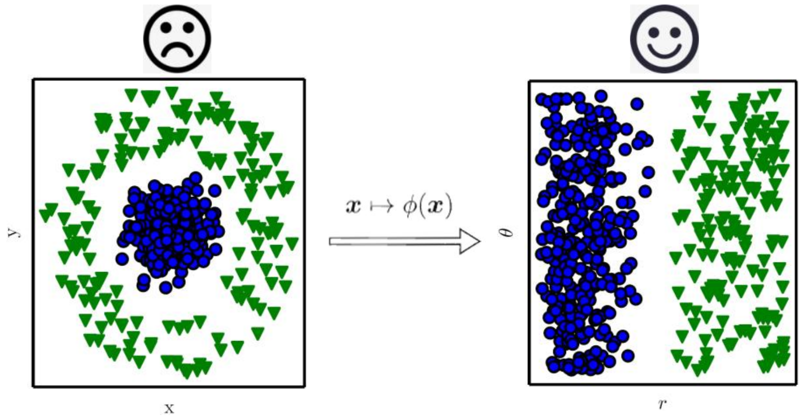
1. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China.

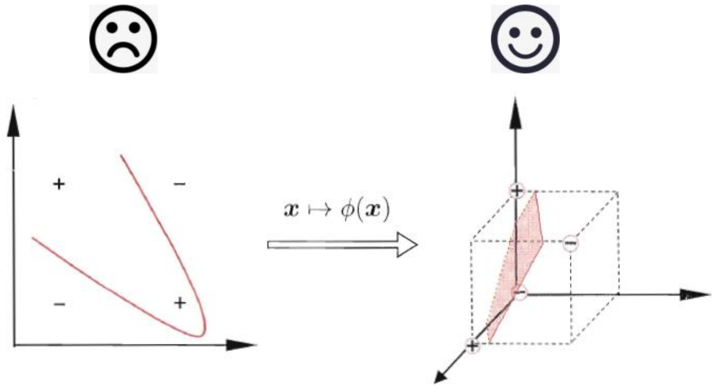
Correspondence to: Jiye Liang <ljy@sxu.edu.cn>

Outline

1. Motivation

■ **Truism** **Good data representation leads to good generalization performance**





■ **However**

General Framework of Machine Learning

$$h^* = \underset{h\in\mathcal{H}}{\arg\min} \sum_{i=1}^{n} \ell\left(y_i, h(x_i)\right) + \mathcal{R}(h), \qquad h^* : \mathcal{X} \to \mathcal{Y}$$

The **relationship between representation learning and generalization performance is not fully considered**

2. Proposed Criterion

(1) Formalize Generalization Error of Representing Learning

$$h^* = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \ell\left(y_i, h(x_i)\right) + \mathcal{R}(h)$$

General Framework of Machine Learning

The **learning process** is decomposed into two processes,

① **Learning classifier**

② **Learning representation**

$$h^* = g^* \circ \varphi^*$$

The final model

$$g^* = \arg\min_{g \in \mathcal{H}\left(\varphi^*(\mathcal{X})\right)} \sum_{i=1}^{n} \ell\left(y_i, g\left(\varphi^*(x_i)\right)\right) + \mathcal{R}\left(g \circ \varphi^*\right)$$

Outer: Learning classifier

Generalization Error of

$$s.t.\ \varphi^* = \arg\min_{\varphi \in \Psi} \boxed{P_{err}\left(\mathcal{H}\left(\varphi(\mathcal{X})\right)\right)}$$

Representing Learning

Inner: Learning representation

$\mathcal{H}\left(\varphi(\mathcal{X})\right)$ is the set of hyperplnes on space $\varphi(\mathcal{X})$

(2) The Upper Bound of Generalization Error

Generalization Error of Representing Learning

$$\min_{\varphi\in\Psi} \boxed{P_{err}\left(\mathcal{H}\left(\varphi\left(\mathcal{X}\right)\right)\right)}$$

① The upper bound of **Generalization error** is dominated by **VC**

② The **VC dimension** is dominated by the **ratio between Radius and Margin**

$$\min_{\varphi\in\Psi} \frac{R^2(\varphi(X))}{M^2(\varphi(X_+),\varphi(X_-))}$$

$$s.t. \begin{cases} \frac{n_{err}}{n} \leq \varepsilon \\ M\left(\varphi\left(X_+\right),\varphi\left(X_-\right)\right) = \sup\limits_{h\in\mathcal{H}(\varphi(X))} g_m(h) \end{cases}$$

Not Executable

Geometric Meaning

Numerator: Radius of training set

Denominator: Margin of hyperplne

① **Theorem 2.4** (Corollary in Chapter 5.4 of (Vapnik, 1999)). *With probability $1-\eta$ one can assert that the probability that a test sample will not be separated correctly by the M-margin hyperplane has the bound $P_{err} \leq \frac{n_{err}}{n} + B_2\left(n, n_{err}, \eta, d_{VC}\right)$, where $B_2\left(n, n_{err}, \eta, d_{VC}\right) = \frac{\mathcal{E}}{2}\left(1+\sqrt{1+\frac{4n_{err}}{n\mathcal{E}}}\right)$, $\mathcal{E} = 4\frac{d_{VC}\left(\ln\frac{2n}{d_{VC}}+1\right)-\ln\frac{\eta}{4}}{n}$, n is the number of training samples, n_{err} is the number of training samples that are not separated correctly by this M-margin hyperplane, and d_{VC} is the VC dimension in **Theorem** 2.3.*

② **Theorem 2.3** (Theorem 5.1 of (Vapnik, 1999)). *Let vectors $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ belong to a sphere of radius R. Then the set of M-margin separating hyperplanes has VC dimension d_{VC} bounded by the inequality*

$$d_{VC} \leq B_1(d, R, M) = \min\left(\left[\frac{R^2}{M^2}\right], d\right) + 1.$$

[1] Vapnik, V. The Nature of Statistical Learning Theory. Springer, 2 edition, 1999.

(3) Making Margin and Radius Executable

ICML
40 Years
International Conference On Machine Learning

$$\min_{\varphi\in\Psi} \frac{R^2(\varphi(X))}{M^2(\varphi(X_+),\varphi(X_-))}$$

$$s.t. \begin{cases} \frac{n_{err}}{n} \le \varepsilon \\ M(\varphi(X_+),\varphi(X_-)) = \sup_{h\in\mathcal{H}(\varphi(X))} g_m(h) \end{cases}$$

Not Executable

① **Theorem 3.1 Margin** can be bounded by a **convex optimization problem**

② **Theorem 3.2 Radius** can be solved by a **convex optimization problem**

③ **Remark A.5** $M > 0 \Rightarrow n_{err}/n = 0$

$$\min_{\varphi\in\Psi} \frac{g_1(\varphi)}{g_2(\varphi)}$$

$$s.t. \begin{cases} g_1(\varphi) = \min_{c\in ch(\varphi(X))} \max_{p\in\varphi(X)} \|c-p\|_2^2 \\ g_2(\varphi) = \min_{p\in ch(\varphi(X_+)),\, q\in ch(\varphi(X_-))} \|p-q\|_2^2 \end{cases}$$

Executable

① **Theorem 3.1.** *The optimization problem (8) can be bounded by the following convex optimization problem. See **Appendix** A.1 for proof.*

$$M^2(X_+,X_-) \ge \frac{1}{4} \min_{\boldsymbol{\alpha}\in\triangle^{n_+},\, \boldsymbol{\beta}\in\triangle^{n_-}} \|\boldsymbol{X}_+\boldsymbol{\alpha} - \boldsymbol{X}_-\boldsymbol{\beta}\|_2^2$$

② **Theorem 3.2.** *Given a set $X \subset \mathbb{R}^d$, the squared radius of X can be computed by the following convex optimization problem. See **Appendix** A.2 for proof.*

$$R^2(X) = \inf_{\boldsymbol{\theta}\in\triangle^{|X|}} \sup_{\boldsymbol{x}\in X} \|\boldsymbol{x} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$$

Can be Calculated Effectively and Efficiently

[1] Boyd, S. and Vandenberghe, L. Convex Optimization. Cambridge University Press, 2004.

3. Application I: Kernel Selection

ICML 40 Years International Conference On Machine Learning

■ Kernel Selection Framework

$$\min_{\varphi \in \Psi_{\text{kernel}}} f(\varphi) = 4\frac{g_1(\varphi)}{g_2(\varphi)}$$

Generalization Error Criterion

$$s.t. \begin{cases} g_1(\varphi) = \min_{\mathbf{u}} \max_{\mathbf{x}_i \in X} \\ \sigma(\mathbf{u})^T \mathbf{K}^{\varphi} \sigma(\mathbf{u}) - 2\mathbf{K}^{\varphi}_{[i,:]} \sigma(\mathbf{u}) + k^{\varphi}_{ii} \\ g_2(\varphi) = \min_{\mathbf{v},\mathbf{w}} \begin{pmatrix} \sigma(\mathbf{v}) \\ \sigma(\mathbf{w}) \end{pmatrix}^T \hat{\mathbf{K}}^{\varphi} \begin{pmatrix} \sigma(\mathbf{v}) \\ \sigma(\mathbf{w}) \end{pmatrix} \end{cases}$$

Candidate set of kernel functions

- **kernel type**
- **kernel parameter**

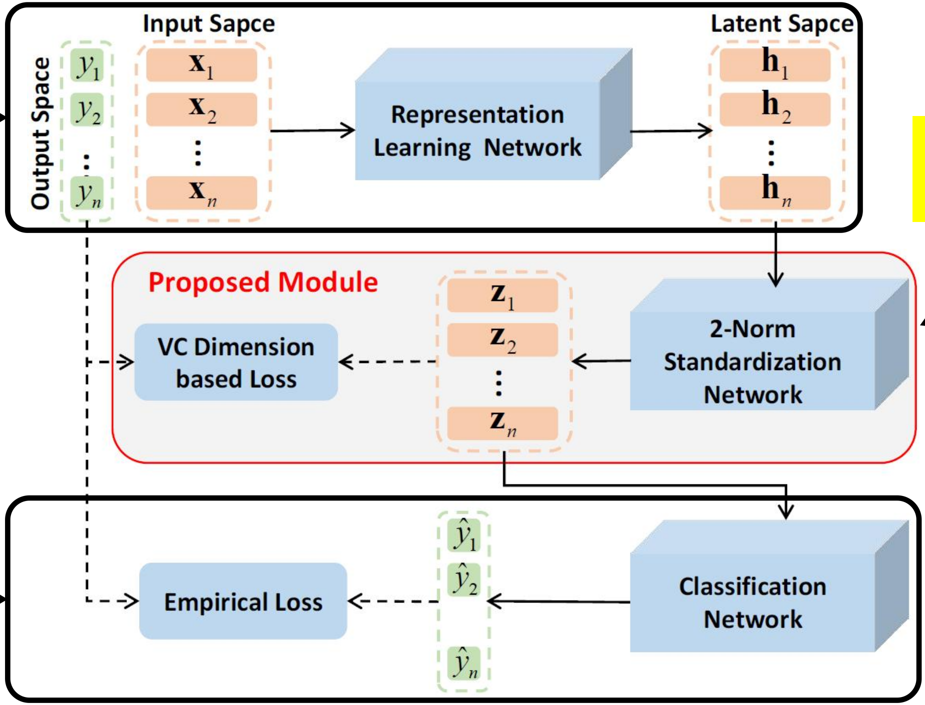
Select the kernel function with the **smallest generalization error**

4. Application II: DNN Boosting

■ DNN Boosting Framework

General DNN

- ➢ Architecture
- ➢ Loss



Generalization Error Criterion

Using the proposed criteria to train the parameters

5. Conclusion and Outlook

1. **A criterion** can measure the **generalization error of representation learning** and can be **calculated Effectively and Efficiently** **(Have completed)**
2. **Successful application** in kernel selection **(Have completed)**
3. **Successful application** in boosting DNN **(Have completed)**
4. **A Powerful tool** for analyzing other methods **(Be going to)**
5. **Provide Guidance** for designing new methods **(Be going to)**

Thanks!

School of Computer and Information Technology (School of Big Data), Shanxi University http://cs.sxu.edu.cn/index.html

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, http://cicip.sxu.edu.cn/index.htm

Correspondence to: Jiye Liang <ljy@sxu.edu.cn>