The Oretical Analysis of Nearest Neighbor with Feature Transformation in Few-shot Learning

He Lang,Yi Tianlun,Yang Jin

Tsinghua-Berkeley Shenzhen Institute

ABSTRACT

Few-shot learners aim to recognize new object classes based on a small number of labeled training examples.However,a significant problem of few-shot learning is overfitting.To prevent overfitting,a main idea is to extract image features using a convolutional network, and then use a combination of meta-learning and nearest-neighbor to perform recognition.Therefore, since the nearest neighbor method have such a good performance, it's worthwhile to discuss some properties of it.In this paper,we suppose a gaussian distribution as sparse gaussian case and calculate the whole error rate and the asymptotic error rate of the nearset neighbor method.

INTRODUCTION

To preven to verftting,researchers try many methods.A paper[1] states that just using convolutional network and nearest-neighbor without meta-learning can achieve state-of-the-art.Specially,applying simple feature tranformations on the features before nearest-neighbor classifcation leads to very competitive few-shots learning results.

In that paper,the training set with N is denoted by:

$$\mathcal{D}_{base} = \{(\mathbf{I}_1, y_1), \cdots, (\mathbf{I_N}, y_N)\}$$

The convolutional network $f\vartheta(I)$ is trained to minimize the loss function l, which means:

$$arg \min_{\theta, \mathbf{W}} \sum_{(\mathbf{I}, y) \in \mathcal{D}_{base}} l(\mathbf{W}^T f_{\theta(\mathbf{I})}, y)$$

METHOD

In the nearest neighbor rule method,For the classification problem, we use

$$D_{support} = \{(x_1, 1), \cdots, (x_C, C)\}$$

to represent the one-shot setting. So that for the feature x of an test image:

$$y(\mathbf{x}) = arg \min_{c \in \{1, \cdots, C\}} d(\mathbf{x}, \mathbf{x_c})$$

Here we use L2 distance:

$$d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2$$

According to the nearest-neighbor method, the whole error is:

$$P_{error} = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{1}\{\mathbf{x} \notin \mathcal{X}_{y(\boldsymbol{x})}\}$$

RESULT

The whole error rate of the nearset neighbor method is:

$$P_{error} = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X}} P_e(\boldsymbol{x})$$

For x with label c, the unit error rate is:

$$P_e(\boldsymbol{x}) = 1 - \int_{\boldsymbol{x} \in D_c'} \mathcal{N}(\mu_c, \Sigma_c) d\boldsymbol{x}$$

Taking expectation at both sides we get the upper bound for the **overall error rate** of 1NN rule:

$$\begin{aligned} R_{1NN} &= \mathrm{E}_{\boldsymbol{x} \sim f(\boldsymbol{x})}[p_e(\boldsymbol{x})] \\ &= 2\mathrm{E}_{\boldsymbol{x} \sim f(\boldsymbol{x})}[r^*(\boldsymbol{x})] - \frac{M-1}{M} \mathrm{E}_{\boldsymbol{x} \sim f(\boldsymbol{x})}\left[(r^*(\boldsymbol{x}))^2\right] \\ &= 2R^* - \frac{M-1}{M}(R^*)^2 - \frac{M-1}{M}\mathrm{Var}\left[r^*(\boldsymbol{x})\right] \\ &\leqslant 2R^* - \frac{M-1}{M}(R^*)^2 \end{aligned}$$

For the lower bound,we have

$$R^* \leqslant R_{1NN} \leqslant 2R^* - \frac{M-1}{M}(R^*)^2.$$

When we use a prior probability, there is:

$$P_e(\boldsymbol{x}) \doteq e^{-nD(Q^* \| P_0)}$$

$$Q^*(\boldsymbol{x}) = \frac{P_1^s(\boldsymbol{x}) P_0^{1-s}(\boldsymbol{x})}{\sum_{\hat{x}} P_1^s(\hat{x}) P_0^{1-s}(\hat{x})}$$

CONCLUSIONS

Provide a theoretical explanation for the improvement of 1NN rule brought by feature transformations.

MAIN REFERENCES

[1] Y. Wang, W. L. Chao, K. Q. Weinberger, and V. Laurens,"Simpleshot: Revisiting nearest-neighbor classification for few-shot learning," 2019.