Автоматическое построение нейросети оптимальной сложности.

Забазнов А. Г.[1], *Бахтеев О. Ю.*[1], *Стрижов В. В.*[1,2]
antoniozabaznov@yandex.ru; bakhteev@phystech.edu; strijov@phystech.edu
Московский физико-технический институт[1];
Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН[1,2]

В данной работе рассматривается задача выбора оптимальной модели нейросети и оптимизация её параметров. В общем случае нейросеть представляется графом, ребрами которого являются нелинейные операции, а вершины – промежуточные представления выборки, полученные под действием этих операций. Параметры сети можно разделить на три типа: параметры, отвечающие за итоговое качество классификации; гиперпараметры, отвечающие за процесс обучения и предотвращение переобучения; структурные параметры, отвечающие непосредственно за структуру сети, такие как количество слоев и тип нелинейных операций. Предлагается подход выбора структуры нейросети на основе вариационного вывода и алгоритма выбора оптимальных значений гиперапараметров с использованием релаксации, учитывающий неточности при оптимизации параметров и позволяющий находить наиболее устойчивые модели.

Ключевые слова: *нейронные сети, автоматическое построение нейронных сетей, оптимальная структура нейронной сети*

1 Введение

При решении задачи классификации или регрессии в машинном обучении выбранная модель зачастую оказывается неоптимальной. Под оптимальной моделью понимается структура обучаемой сети и совокупность её гиперпараметров, которая даёт приемлемое качество классификации или регрессии при небольшом количестве параметров. В данной работе в качестве критерия выбора модели предлагается сложность модели, то есть величина, учитывающая сложность описания совокупности выборки и модели. Под описанием выборки понимается приближенная оценка сложности модели, основанная на связи с её правдоподобием [1]

Существует несколько подходов выбора модели оптимальной сложности. В работе [2] используется метод прореживания модели. Он заключается в построении заведомо переусложнённой модели с дальнейшим удалением параметров, не влияющих на качество классификации, таким образом получается сеть наименьшего размера. Ещё одиним способом, предложенным в работе [3], являются байесовские методы оптимизации параметров нейронных сетей. В работе [4] для оптимизации модели предлагается использовать метод градиентного спуска.

Одна из проблем оптимизации моделей глубокого обучение – большое количество параметров и гиперпараметров, которое может достигать миллионов. Кроме того, сам процесс оптимизации становится ресурсоёмким. Задача выбора модели глубокого включает в себя выбор стратегии построения модели, эффективной по вычислительным ресурсам. Существуют методы градиентной оптимизации совокупности параметров и гиперпараметров.

В данной работе построение модели оптимальной сложности происходит в процессе самого обучения. В основе разработанного метода лежит алгоритм DARTS, предложенный в работе [5]. Для выбора оптимального набора гиперпараметров предлагается параметризовать структуру модели некотором действительным вектором, путём перехода от дискретного множества возможных значений гиперпараметров к непрерывному множетсву их комбинаций.

Проверка и анализ метода проводится на выборке Boston Housing [6], MNIST [7] и CIFAR-10 [8] и синтетических данных. Проводится сравнеине представленного метода с эвристическими алгоритмами выбора модели, а также с алгоритмом DARTS.

2 Постановка задачи

Пусть заданы обучающая и вылидационная выборки:

$$\mathfrak{D}^{\text{train}} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \ldots, m^{\text{train}},$$

$$\mathfrak{D}^{\text{valid}} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \ldots, m^{\text{valid}},$$

состоящие из множеств пар объект-метка,

$$\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^{\text{n}}, \quad y_i \in \mathbf{Y} \subset \mathbb{R}.$$

Метка y объекта $\mathbf{x}$ принадлежит множеству $y \in \mathbf{Y} = \{1, \ldots, Z\}$, где Z - количество классов.

Модель задаётся ориентированным графом $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, где для каждого ребра (i, j) заданы базовые функции $\mathbf{g}^{i,j}, |\mathbf{g}^{i,j}| = K^{i,j}$ и их веса $\boldsymbol{\gamma}^{i,j}$. Требуется построить такую модель $\mathbf{f}$ с параметрами $\mathbf{W} \in \mathbb{R}^{\text{n}}$:

$$\mathbf{f}(\mathbf{x}, \mathbf{W}) = \{\mathbf{f}_i(\mathbf{x}, \mathbf{w}_i)\}_{i=1}^{|\mathbf{V}|}$$

где $\mathbf{f_i}(\mathbf{x}, \mathbf{w_i})$ - подмодель с параметрами $\mathbf{w}_i$ задаётся как:

$$\mathbf{f}_i(\mathbf{x}, \mathbf{w}_i) \ = \sum_{j \in adj(i)} \left\langle \boldsymbol{\gamma}^{i,j}, \mathbf{g}^{i,j} \right\rangle \mathbf{f}_j(\mathbf{x}, \mathbf{w}_j)$$

.

Тогда параметры модели определяются как конкатенация всех параметров каждой подмодели: $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{|\mathbf{V}|}]$, а структура модели $\boldsymbol{\Gamma}$ задаётся вектором $\{\boldsymbol{\gamma}^{i,j}\}_{\mathbf{E}}$.

Функция потерь на обучении L и функция потерь на валидации Q задаются как:

$$L(\mathbf{W}, \boldsymbol{\Gamma}) = \log p(\mathbf{Y}^{\text{train}} | \mathbf{X}^{\text{train}}, \mathbf{W}, \boldsymbol{\Gamma}) + \boldsymbol{e}^{\mathbf{A}} ||\mathbf{W}||^2,$$

$$Q(\mathbf{W}, \boldsymbol{\Gamma}) = \log p(\mathbf{Y}^{\text{valid}} | \mathbf{X}^{\text{valid}}, \mathbf{W}, \boldsymbol{\Gamma}) + \lambda p(\boldsymbol{\Gamma}),$$

где $\mathbf{A}$ и λ - регуляризационные слагаемые, $p(\boldsymbol{\Gamma})$ - произведение всех произведение вероятностей всех $\boldsymbol{\gamma}^{i,j} \in \boldsymbol{\Gamma}$. Перед подсчётом значения функции потерь на валидации делается априорное предположение о распределении вектора $\boldsymbol{\Gamma} = \{\boldsymbol{\gamma}^{i,j}\}$: вектор структуры модели имеет распределение либо Дирихле [9] либо Gumbel-Softmax [10].

Вектор $\{\gamma^{i,j}\}$ имеет распределение Дирихле с параметром α, если:

$$f(\gamma) = f(\gamma_1, \ldots, \gamma_K) = \begin{cases} \frac{\boldsymbol{F}(K \times \alpha)}{\boldsymbol{F}(\alpha)^K} \prod\limits_{i=1}^{K} \gamma_i, \gamma \in \boldsymbol{S} \\ 0, \gamma \notin \boldsymbol{S} \end{cases}$$

, где $\boldsymbol{F}$ - гамма-функция, $\boldsymbol{S}$ - симплекс: $\{\gamma \in \mathbb{R}^K : \sum_{i=1}^K \gamma_i = 1, \gamma_i \geqslant 0\}$.

Вектор $\{\gamma^{i,j}\}$ имеет распределение Gumbal-Softmax с параметром α и параметром τ, если:

$$f(\gamma_1, \ldots, \gamma_K) = (K-1)!\tau^{K-1}\alpha^K \prod_{i=1}^{K} \frac{\gamma_i^{-\tau-1}}{\alpha \sum_{j=1}^K \gamma_j^{-\tau}}$$

При $\tau \to \inf$ распределение Gumbal-Softmax эквивалентно многомерному нормальному распределению.

Требуется решить задачу двухуровневой оптимизации, оптимизируя параметры модели по обучающей выборке, а структуру модели по валидационной:

$$\mathbf{W}^*(\mathbf{\Gamma}) = \arg\min_{\mathbf{W}} L(\mathbf{W}, \mathbf{\Gamma})$$

$$\mathbf{\Gamma}, \mathbf{A} = \min_{\mathbf{\Gamma}} Q(\mathbf{W}^*(\mathbf{\Gamma}), \mathbf{\Gamma})$$

3 Релаксация

Известно множество всех возможных операций $\mathbf{g}^{i,j} \in \mathbf{G}$. Для перехода к непрерывному пространству таких функций проводится релаксация каждой операции:

$$\overline{\mathbf{g}(\mathbf{x})} = \sum_{\gamma \in \mathbf{\Gamma}} \frac{\boldsymbol{e}^{\gamma}}{\sum_{\gamma' \in \mathbf{\Gamma}} \boldsymbol{e}^{\gamma'}} \mathbf{g}(\mathbf{x})$$

После релаксации необходимо совместное исследование $\mathbf{\Gamma}$ и весов $\boldsymbol{w}$ всех смешанных операциях $\overline{\mathbf{g}^{i,j}}$.

Литература

[1] Grunwald P. A tutorial introduction to the minimum description length principle. 2005.

[2] John S. Denker Yann Le Cun and Sara A. Solla. Optimal brain damage. 1989.

[3] A. Neal and M. Radfor. Bayesian learning for neural networks. 1995.

[4] T. Raiko J. Luketina, M. Berglund and K. Gref. Scalable gradient-based tuning of continuous regularization hyperparameters. 2016.

[5] Yang .Y Hanxiao L., Simonyan K. Darts: Differentiable architecture search. 2018.

[6] Daniel L. Harrison Jr., Rubinfeld D. Hedonic housing prices and the demand for clean air. 1978.

[7] Christopher J.C. Burges Yann LeCun, Corinna Cortes. The mnist database of handwritten digits. 1998.

[8] G. Hinton A. Krizhevsky, V. Nair. The cifar-10 dataset. 2009.

[9] Tommi S. Jaakkol Harald Steck. On the dirichlet prior and bayesian regularization.

[10] Yee Whye Tehl Chris J. Maddison, Andriy Mnih. The concreate relaxation: A continues relaxation of discrete random variables.

Поступила в редакцию