Анализ статистической и структурной сложности суперпозиции нейронных сетей

Д. О. Перекрестенко

Научный руководитель:
н.с. ВЦ РАН, к.ф.-м.н. В. В. Стрижов

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Цель работы

- Цель работы — разработать метод нахождения оптимальной структурной сложности универсальной модели нейросети.
- Мотивация — перебор различных структур нейросети связан с вычислительно трудоемкой задачей нахождения оптимальных параметров для каждой из рассматриваемых структур нейросети.
- Предлагается определять оптимальную структуру нейросети без использования переборных методов и многократных процедур оптимизации параметров.
- Идея — оценить структурную сложность одной модели по структурной сложности другой модели. При этом считается, что получение сруктурной сложности второй модели требует меньшего объема вычислений.

Предлагается

1. Задать критерий сложности выборки
2. Задать критерий структурной сложности нейросети
3. Найти связь между геометрической и структурной сложностями на множестве выборок.

Постановка задачи

Задана выборка D из генеральной совокупности D_{gen}:

$$D = \{\mathbf{x}_i, y_i\}_{i=1}^m,$$

где $\mathbf{x}_i \in \mathbb{R}^n$ — вектор, признаковое описание i-го объекта, а $y_i \in \{1, 2, \dots, k\}$ — метка класса из номинальной шкалы. Требуется найти модель:

$$\mathbf{f} : (\mathbf{x}, \mathbf{w}) \to \mathrm{c}$$

$$\mathbf{f} : \mathbb{R}^n \times \mathbb{W} \to \{1, \dots, k\}$$

из множества $\mathbf{f} \in \mathfrak{F}$ нейронных сетей, которая классифицирует генеральную совокупность D_{gen}.

Базовые определения

Определение 1

Назовем функцию $g : \mathbb{W} \to \mathbb{R}$ *структурной сложностью* модели $A = \{\mathbf{f}(\mathbf{x}, \mathbf{w}) | \mathbf{w} \in \mathbb{W}\}$, если $g(\mathbf{w})$ возрастает с ростом N, где N — математическое ожидание числа элементарных шагов алгоритма настройки модели к параметрам $\mathbf{w}$ из случайно заданного начального приближения $\mathbf{w}_0$.

Определение 2

Назовем γ-*геометрической сложностью* выборки D минимальное число радиальных базисных функций $\phi_k(\mathbf{x}) = \exp(-\frac{\rho(\mathbf{c}_k - \mathbf{x})}{a_k^2})$ необходимых для классификации выборки D с точностью γ. В данной работе γ принята равной 0.95.

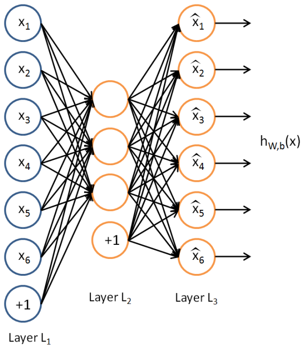
Автокодировщик

Автокодировщик $\mathbf{h}$ это монотонное нелинейное отображение входного вектора свободных переменных $\mathbf{x} \in \mathbb{R}^n$ в скрытое представление $\mathbf{h} \in \mathbb{R}^\nu$ следующего вида:

$$\mathbf{h}(\mathbf{x}) = \boldsymbol{\sigma}(\underset{\nu \times n}{\mathbf{W}}\,\mathbf{x} + \mathbf{b}).$$

Скрытое представление $\mathbf{h}$ создает линейную реконструкцию вектора $\mathbf{x}$:

$$\mathbf{r}(\mathbf{x}) = \underset{n \times \nu}{\mathbf{W}'}\mathbf{h} + \mathbf{b}'.$$



Структура автокодировщика

Настройка параметров

Параметры автокодировщика

$$\boldsymbol{\lambda} = \{\mathbf{W}', \mathbf{W}, \mathbf{b}', \mathbf{b}\}$$

оптимизированы таким образом, чтобы сделать реконструкцию $\mathbf{r}(\mathbf{x})$ максимально близкой к $\mathbf{x}$. Функция ошибки автокодировщика:

$$S(\boldsymbol{\lambda}) = \frac{1}{2m}\sum_{i=1}^{m} \|\mathbf{r}(\mathbf{x}_i) - \mathbf{x}_i\|^2 + \|\mathbf{W}\|_F^2 + \beta \sum_{j=1}^{m} \left[\rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}\right],$$

где m — количество элементов в обучающей выборке, β — вес разреживающего слагаемого, ρ — параметр разреженности, желаемое среднее значение каждой компоненты скрытого представления $\mathbf{h}$, а $\hat{\rho}_j$ — среднее значение j-ой компоненты вектора $\mathbf{h}$.

Базовые определения

Определение 3

Назовем *размерной сложностью* выборки D число нейронов в скрытом слое автокодировщика, задающее минимум его функции ошибки.

$$\mathrm{DimComp}(D) = \arg \min_{\mathrm{size}(\hat{\mathbf{W}},1)} S(\hat{\boldsymbol{\lambda}}|D),$$

где $\hat{\boldsymbol{\lambda}}$ — оптимальные параметры автокодировщика.

Гипотеза

Предполагается что между сложностью выборки и структурной сложностью нейросети есть линейная связь.

Прогнозирование структурной сложности

Пусть задано M выборок $\{D_1, \ldots, D_M\}$, где $D_i = \{\mathbf{x}^i_m, y^i_m\}_{m=1}^n$, таких что для каждой из них известна оптимальная структурная сложность классифицирующей их нейронной сети. Будем восстанавливать по этим выборкам регрессию структурной сложности модели StrComp по сложности выборки Comp:

$$\hat{\chi} = \arg\min_{\chi} \sum_{i=1}^{M} \left(\chi_0 + \chi_1 \mathrm{Comp}_i - \mathrm{StrComp}_i\right)^2,$$

получив модель регрессии мы можем получать суб-оптимальные значения структурной сложности сети для заданной геометрической сложности выборки.

$$\mathrm{StrComp}_{\mathrm{subopt}}(\mathrm{Comp}) = \hat{\chi}_0 + \hat{\chi}_1 \mathrm{Comp}.$$

Выбор модели

В данной работе модель $\mathbf{f}$ представлена в виде суперпозиции блоков:

$$\mathbf{f} = \mathbf{a}(\mathbf{h}_N(\ldots \mathbf{h}_1(\mathbf{x}))),$$

где $\mathbf{h}_k$ — блоки-автокодировщики, вида

$$\mathbf{h}_k(\mathbf{x}) = \boldsymbol{\sigma}(\mathbf{W}_k \mathbf{x} + \mathbf{b}_k),$$

а блок $\mathbf{a}$ — классификатор мультиномиальной логистической регрессии вида

$$\mathbf{a}(\mathbf{x}) = \arg\max_l \left(\frac{1}{\sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^\top \mathbf{x}}} \begin{bmatrix} e^{\boldsymbol{\theta}_1^\top \mathbf{x}} \\ e^{\boldsymbol{\theta}_2^\top \mathbf{x}} \\ \vdots \\ e^{\boldsymbol{\theta}_k^\top \mathbf{x}} \end{bmatrix}^\top \cdot \mathbf{e}_l \right),$$

где $\mathbf{e}_l$ — l-й столбец единичной матрицы $\mathbf{E}_k$.

Функция ошибки

Функция ошибки модели $\mathbf{f}$:

$$S(\boldsymbol{\alpha}) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k}[y_i = j]\log p(\hat{y}_i = j|\mathbf{x}_i; \boldsymbol{\alpha}),$$

где $\boldsymbol{\alpha} = \{\mathbf{W}_1, \ldots, \mathbf{W}_N, \mathbf{b}_1, \ldots, \mathbf{b}_N, \boldsymbol{\theta}\}$ — вектор состоящий из параметров всех блоков модели $\mathbf{f}$, а

$$p(\hat{y}_i = j|\mathbf{x}_i; \boldsymbol{\alpha}) = \frac{e^{\boldsymbol{\theta}_i^T \mathbf{h}_N(\ldots\mathbf{h}_1(\mathbf{x}_i))}}{\sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^T \mathbf{h}_N(\ldots\mathbf{h}_1(\mathbf{x}_i))}}.$$

Требуется найти вектор параметров $\boldsymbol{\alpha}_{\text{opt}} \in \mathbb{W}$, который минимизирует функцию ошибки на заданной выборке D:

$$\boldsymbol{\alpha}_{\text{opt}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{W}} S(\boldsymbol{\alpha}|D).$$

Структурные сложности

1. Количественная сложность:

$$\mathrm{Comp}_1(\mathbf{f}) = k\nu_N + \sum_{i=1}^{N} \nu_i(\nu_{i-1} + 1),$$

   где k — количество классов, а ν_i — размер i-го блока(скрытого слоя) модели $\mathbf{f}$. Размерная сложность это количество параметров модели $\mathbf{f}$.

2. Графовая сложность:

$$\mathrm{Comp}_3(\mathbf{f}) = \sum_{i=1}^{M} \sum_{(k,j)\in V^i} \omega_{kj},$$

   где V^i — i-й подграф, ω_{ij} — индикатор существования ребра (k, j), M — число вершин графа.

Структурные сложности

3 Взвешенная количественая сложность:

$$\mathrm{Comp}_2(\mathbf{f}) = \|\hat{\boldsymbol{\theta}}\|_F^2 + \sum_{i=1}^{N} \left(\|\hat{\mathbf{W}}_i\|_F^2 + \|\hat{\mathbf{b}}_i\|_F^2 \right),$$

где $\hat{\boldsymbol{\theta}}$ — матрица настроенных параметров классификатора, а $\hat{\mathbf{W}}_i, \hat{\mathbf{b}}_i$ — настроенные параметры i-го блока-автокодировщика. Взвешенная размерная сложность это сумма квадратов значений всех параметров модели.

4 Взвешенная графовая сложность:

$$\mathrm{Comp}_4(\mathbf{f}) = \sum_{i=1}^{M} \sum_{(k,j) \in V^i} w_{kj},$$

где V^i — i-й подграф, w_{kj} — вес ребра (k, j), M — число вершин графа.

Структурные сложности

Утверждение 1

Количественная Comp_1 и графовая Comp_3 сложности являются *структурными сложностями* для слоевых нейронных сетей настраиваемых алгоритмом глубокого обучения.

Утверждение 2

Взвешенные количественная Comp_2 и графовая Comp_4 сложности являются *структурными сложностями* для слоевых нейронных сетей настраеваемых алгоритмом глубокого обучения, для которых алгоритм сошелся к глобальному минимуму.

Вычислительный эксперимент

Для вычислительного эксперимента использовалось 6 выборок:

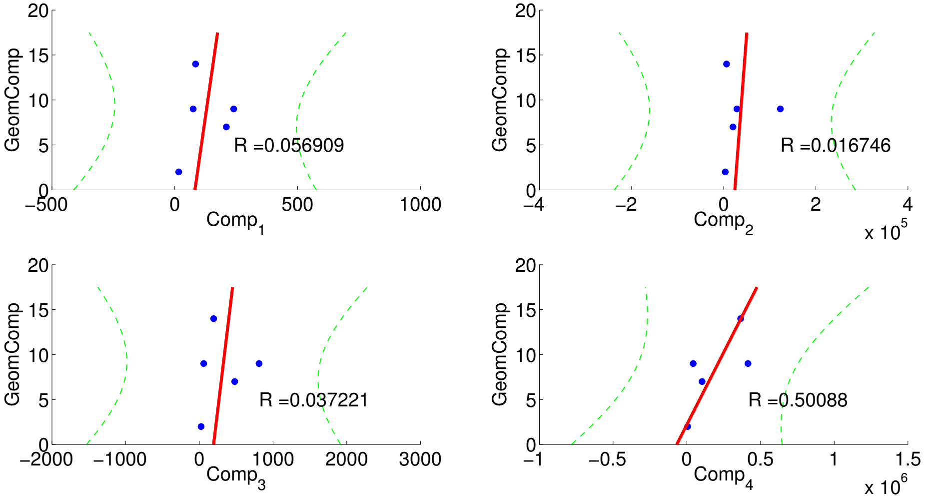
1. Временные ряды акселлерометра. Количество признаков — 600, классов — 4.
2. Синтетически сгенерированная выборка. Количество признаков — 2, классов — 2.
3. Распознавание сортов вин. Количество признаков — 13, классов — 3.
4. Распознавание ирисов. Количество признаков — 4, классов — 3.
5. Распознавание патологий кожного покрова груди. Количество признаков — 9, классов — 6.

Вычислительный эксперимент

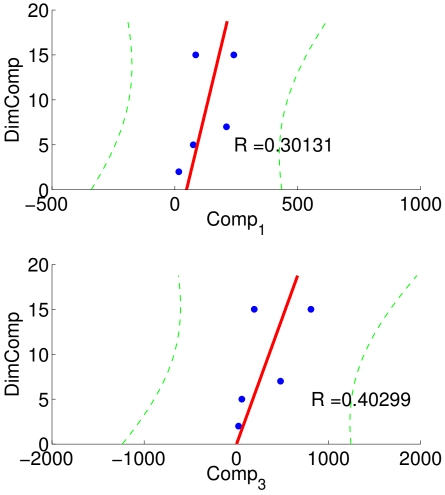
DataSet	$Comp_1$	$Comp_2$	$Comp_3$	$Comp_4$
№1	210	2.0243e+04	4800	1.0360e+05
№2	75	2.8950e+04	60	4.3208e+04
№3	85	6.2355e+03	195	3.6551e+04
№4	16	3.6485e+03	24	5.6853e+03
№5	240	1.2314e+05	810	5.1547e+05

DataSet	GeomComp	DimComp
№1	7	7
№2	9	5
№3	14	20
№4	2	2
№5	6	15

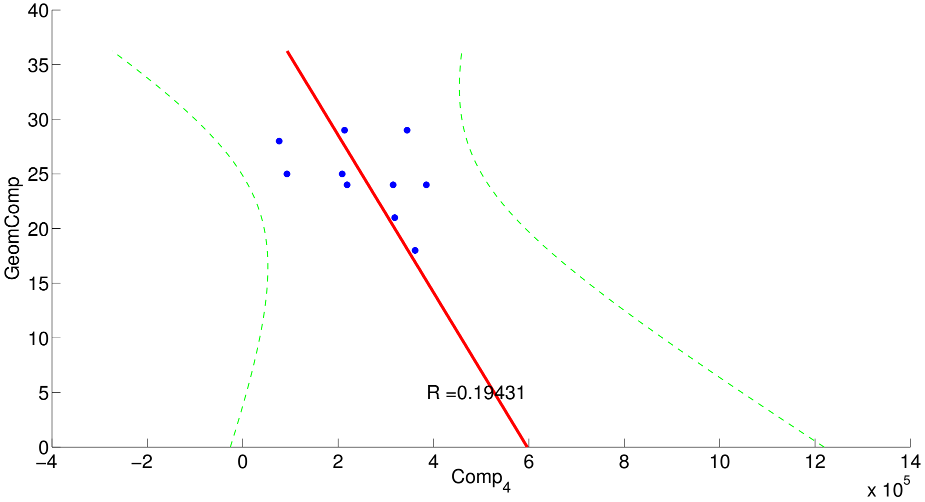
Прогнозирование оптимальной структурной сложности модели по геометрической сложности



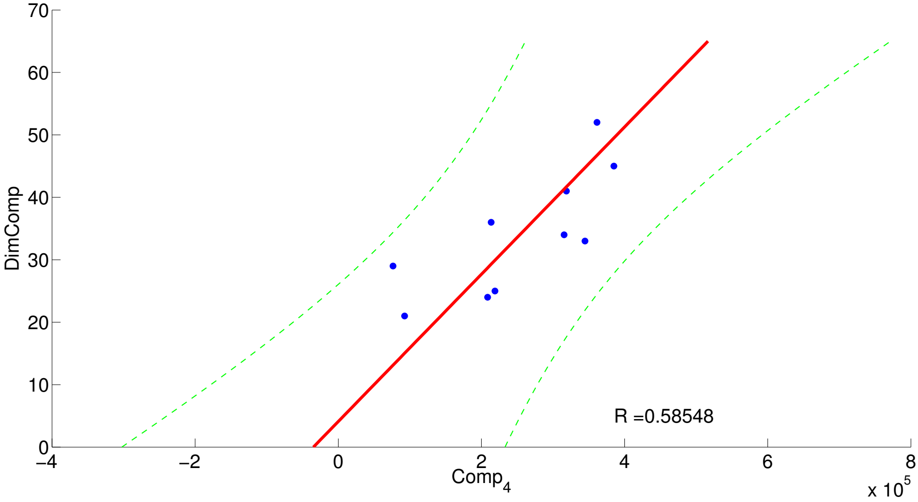
Прогнозирование оптимальной структурной сложности модели по размерной сложности



Прогнозирование оптимальной структурной сложности модели по геометрической сложности



Прогнозирование оптимальной структурной сложности модели по размерной сложности



Заключение

- Реализован и исследован алгоритм прогнозирования структурной сложности нейронной сети по сложности выборки.
- Предложены четыре критерия структурной сложности универсальной модели нейросети.
- Предложены критерии геометрической и размерной сложности выборки.
- Проведена серия численных экспериментов на модельных данных. В результате получено, что пара Comp_4 и DimComp являются хорошо корреллирующими между собой. Определение структурной сложности по быстровычислимой размерной сложности позволяет значительно сократить перебор гиперпараметров нейросетей.