Предсказание качества для процедуры выбора признаков*

Аминов Т.В.

aminov.tv@phystech.edu

[1]Московский Физико-технический институт

В случае избыточного признакового пространства предсказатеьная модель машинного обучения является неустойчивой. Для повышения устойчивости модели применяются методы выбора признаков. Задача выбора признаков состоит в поиске решения среди экпоненциального числа возможных решений, поэтому она является NP-задачей. Для решения применяются эвристические, субоптимальные алгоритмы. В данной статье предлагается свести дискретную задачу выбора признаков к задаче непрерывной оптимизации. Строится процедуры предсказания качества модели на тестовой выборке для подмножества признаков. Решение задачи выбора признаков восстанавливается из решения непрерывной задачи. Проводится вычислительный эксперимент, чтобы сравнить результаты предложенного метода с существующими алгоритмами.

Ключевые слова: *Выбор признаков, многомерные пространства, отображение булева куба.*

1 Введение

При решении многих задач машинного обучения, таких как классификация и регрессия, исходное пространство признаков оказывается избыточным. Не все признаки коррелируют с целевой переменной, нерелевантные признаки делают модель неустойчивой. Данная работа посвящена проблеме выбора оптимального подмножества признаков. Эта задача является NP-задачей, так как точное решение ищется среди всевозможных $2^n - 1$ вариантов, где n - количество признаков. Поэтому предполагается отыскание приближенного решения.

Задача выбора признаков возникает при распознавании лиц [1]. В статье [2] данная проблема в пространствах малой размерности (порядка 1000). За счет огромного

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Задачу поставил: Роман Р. В. Консультант: Роман Р. В.

прироста вычислительной мощности за последние 20 лет задача выбора признаков может решаться точно за разумное время.В статье [3] пердставлена идея построения оптимальной архитектуры нейронной сети . Существующие методы, независимо от того, основаны ли они на обучении с подкреплением [4] или на эволюционных алгоритмах (EA) [5], выполняют поиск архитектуры в дискретном пространстве, который крайне неэффективен.

Авторы предлагают простой и эффективный метод автоматического выбора признаков непрерывной оптимизации. Весь алгоритм разбивается на три этапа: (1) енкодер — отображает пространтво подмножества признаков в линейное пространство; (2) предиктор — принимает непрерывное представление подмножества признаков в качестве входных данных и прогнозирует ошибку на тестовой выборке; (3) декодер — отображает непрерывное представление обратно в набор признаков. Существенным отличием нашего метода от уже сущесвтующих состоит в том что размерность булева куба признаков очень велика, а значит для прироста скорости нам придется отображать его в пространство гораздо меньшей размерности.

В качестве показателя эффективности предложенного алгоритма предполагается сравнение с уже существующими : CSO (PSO) [6], CFS [7], QPFS [8], на реальных и синтетических данных.

2 Постановка задачи

Пусть $\mathbf{X} = [\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n] \in \mathbb{R}^{m \times n}$ — заданная матрица, где $\boldsymbol{\chi}_j \in \mathbb{R}^m$ — j-ый признак. Пусть $\mathbf{y} \in \mathbb{R}^m$ — значеня функии целевой (закона природы), которые будут пердказываться. Множество $\mathcal{A} \subseteq \{1, \ldots, n\}$ — индикатор $\{0, 1\}$. Существует соответствие между множеством $\mathcal{A}$ и двоичным вектором $\mathbf{a} \in^n$:

$$\mathcal{A} = \{j : a_j = 1\}.$$

Функция $f(\mathbf{x}, \mathbf{w}, \mathcal{A})$ предсказывает $\mathbf{y}$ по заданному объекту $\mathbf{x}$ — сторочка матрицы $\mathbf{X}$, используя только признаки из множества $\mathcal{A}$. Данные, природа которых зависит от конкретной задачи (заданная матрица объектов-признаков $\mathbf{X}$), разбиваются на две части: обучающую $(\mathbf{X}_{\text{tr}}, \mathbf{y}_{\text{tr}})$ и тестовую $(\mathbf{X}_{\text{te}}, \mathbf{y}_{\text{te}})$. Чтобы измерить качество аппроксимации f, вводится функция функция ошибки $s(\mathbf{w}, \mathbf{X}, \mathbf{y}, \mathcal{A})$. Параметры $\mathbf{w}^*$, находятся

из условия минимизацией:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} s(\mathbf{w}, \mathbf{X}_{\text{tr}}, \mathbf{y}_{\text{tr}}, \mathcal{A}).$$

Необходимо найти множество $\mathcal{A}$ на котором функция ошибки на тестовой выборке будет минимальна. Создается набор данных$\{(\mathbf{a}_i, s_i)\}_{i=1}^N$. Выбираются N векторов $\mathbf{a}$ из двоичного куба ^{n}. Затем оценивается ошибка s_i для каждого набора $\mathbf{a}_i$. Идея в том чтобы отобразить некоторые элементы ^{n} в непрерывное линейное пространство меньшей размерности $h < n$. Таким образом каждый двоичный вектор $\mathbf{a}$ будет обладать непрерывным представлением $\mathbf{u} \in \mathbb{R}^h$. Это представление создается моделью $e(\mathbf{a}, \mathbf{w}_e)$. После этого другая модель $p(\mathbf{u}, \mathbf{w}_p)$ пытается предсказать ошибку на тестовой выборке s. Тогда модель принимает вид $s = p(\mathbf{u}, \mathbf{w}_p) = p(e(\mathbf{a}, \mathbf{w}_e), \mathbf{w}_p)$. Возможно использование любой функции потерь для оценки параметров $\mathbf{w}_e, \mathbf{w}_p$. Например это может быть квадратичная функция ошибки

$$L_{error}(\mathbf{w}_e, \mathbf{w}_p, \mathbf{a}, s) = \|p(e(\mathbf{a}, \mathbf{w}_e), \mathbf{w}_p) - s\|^2 \to \min_{\mathbf{w}_e, \mathbf{w}_p}.$$

Так же требуется построить модель обратного отображения вектора $\mathbf{a}$ по его непрерывному представлению $\mathbf{u}$. Это небходимо для восстановления набора признаков, который дает лучшее описание зависимости $\mathbf{y}$ от $\mathbf{x}$. Эта модель будет $\sigma(\mathbf{u}, \mathbf{w}_\sigma)$. Вводится функция ошибки отображения - кросс-итропия между начальным вектором $\mathbf{a}$ и тем что получаетя на выходе модели декодер σ

$$L_{rec}(\mathbf{w}_e, \mathbf{w}_\sigma, \mathbf{a}) = \sum_{i=1}^{n} a_i \log(\sigma_i) + (1 - a_i) \log(1 - \sigma_i) \to \min_{\mathbf{w}_e, \mathbf{w}_\sigma},$$

где $\sigma_i = \sigma(e(\mathbf{a}, \mathbf{w}_e), \mathbf{w}_\sigma)_i$. Общая функция потерь является линейной комбинацией ошибки предсказания и ошибки отображения

$$L = L_{error} + \alpha L_{rec} \to \min$$

Совокупностью параметров $\mathbf{w}_e, \mathbf{w}_p, \mathbf{w}_\sigma$ задается наиболее подходящее подмножество признаков. Максимизируется предсказательное качество модели

$$\mathbf{u}^* = \arg\max_{\mathbf{u}} p(\mathbf{u}, \mathbf{w}_p).$$

$$\mathbf{a}^* = \sigma(\mathbf{u}^*, \mathbf{w}_\sigma).$$

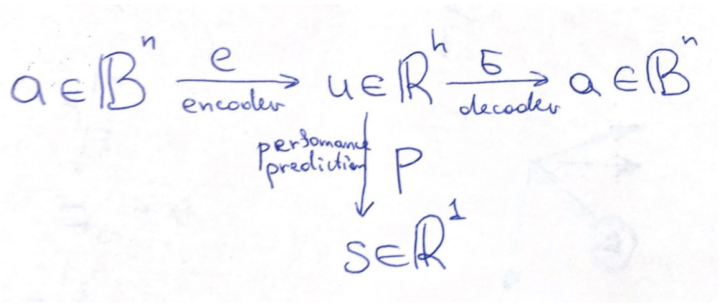

Рис. 1 Общая схема модели

2.1 QPFS

QPFS — один из многих алгоритмов отбора признаков. Этот алгоритм сводит задачу отбора признаков к задаче квадратичной оптимизации.

$$(1-\alpha)\cdot \underbrace{\mathbf{z}^\mathsf{T}\mathbf{Q}\mathbf{z}}_{\mathrm{Sim}(\mathbf{X})} - \alpha \cdot \underbrace{\mathbf{b}^\mathsf{T}\mathbf{z}}_{\mathrm{Rel}(\mathbf{X},\boldsymbol{\nu})} \to \min_{\substack{\mathbf{z}\geqslant \mathbf{0}_n \\ \mathbf{1}_n^\mathsf{T}\mathbf{z}=1}}. \tag{1}$$

$$\|\boldsymbol{\nu} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \to \min_{\boldsymbol{\theta}\in\mathbb{R}^n}.$$

где $\boldsymbol{\nu}$ вектор целевой функции для задачи линейной регрессии Авторы оригинальной статьи [8] предлагают выбирать α and $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\nu})$ для (1):

$$\alpha = \frac{\overline{\mathbf{Q}}}{\overline{\mathbf{Q}} + \overline{\mathbf{b}}}, \quad \overline{\mathbf{Q}} = \mathrm{mean}(\mathbf{Q}), \ \ \overline{\mathbf{b}} = \mathrm{mean}(\mathbf{b}).$$

Параметры QPFS задаются следующим образом:

$$\mathbf{Q} = \left[|\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)|\right]_{i,j=1}^{n}, \quad \mathbf{b} = \left[|\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\nu})|\right]_{i=1}^{n}. \tag{2}$$

Здесь $\mathrm{corr}(\cdot,\cdot)$, абсолютная величина выборочного коэффициента корреляции Пирсона:

$$\mathrm{corr}(\boldsymbol{\chi}, \boldsymbol{\nu}) = \frac{\sum_{i=1}^{m}(\boldsymbol{\chi}_i - \overline{\boldsymbol{\chi}})(\boldsymbol{\nu}_i - \overline{\boldsymbol{\nu}})}{\sqrt{\sum_{i=1}^{m}(\boldsymbol{\chi}_i - \overline{\boldsymbol{\chi}})^2 \sum_{i=1}^{m}(\boldsymbol{\nu}_i - \overline{\boldsymbol{\nu}})^2}}.$$

Литература

[1] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. pages 3025–3032, 06 2013.

[2] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *Ann. Statist.*, 44(2):813–852, 04 2016.

[3] Shengcao Cao, Xiaofang Wang, and Kris M. Kitani. Learnable embedding space for efficient neural architecture compression. In *International Conference on Learning Representations*, 2019.

[4] Leslie Pack Kaelbling, Michael Littman, and Andrew P Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 04 1996.

[5] Q. Song, J. Ni, and G. Wang. A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):1–14, Jan 2013.

[6] Zhaleh Manbari, Fardin Akhlaghian Tab, and Chiman Salavati. Hybrid fast unsupervised feature selection for high-dimensional data. *Expert Systems with Applications*, 124, 06 2019.

[7] Shenkai Gu, Ran Cheng, and Yaochu Jin. Feature selection for high-dimensional classification using a competitive swarm optimizer. *Soft Computing*, 22(3):811–822, Feb 2018.

[8] Irene Rodriguez-Lujan, Ramon Huerta, Charles Elkan, and Carlos Santa Cruz. Quadratic programming feature selection. *J. Mach. Learn. Res.*, 11:1491–1516, August 2010.