Предсказание качества для процедуры выбора признаков*

Аминов Т.В.

aminov.tv@phystech.edu

[1]Московский Физико-технический институт

При решении задач машинного обучения, в случае избыточного пространства признаков, предсказатеьная модель является неустойчивой. Для повышения устойчивости модели применяется выбор признаков. Задача выбора признаков является NP-задачей. Для решения применяются эвристические, субоптимальные алгоритмы. В данной статье предлагается свести дискретную задачу выбора признаков к задаче непрерывной оптимизации. Строится модель предсказания качества на тестовой выборке подмножества признаков. Решение задачи выбора признаков восстанавливается из решения непрерывной задачи. Проводится вычислительный эксперимент, чтобы сравнить результаты предложенного метода с существующими алгоритмами.

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

1 Введение

При решении многих задач машинного обучения, таких как классификация, существует большое количество признаков. 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 всех $2^n - 1$ вариантов. Поэтому предполлагается отыскание приближенного решения.

Задача выбора признаков возникает при распознавании лиц [1]. В этой статье [2] данная проблема в пространсвтах 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 эта задача может решаться точно за разумное время.В данной статье мы ставим задачу применить идею построения оптимальной архитектуры нейронной сети [3] для построения оптимального подмножества признаков.

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. В этой статье авторы предлагают простой и эффективный метод автоматического проектирования нейронной архитектуры на основе непрерывной оптимизации. Весь алгоритм разбивается на три этапа: (1) енкодер - отображает структуру нейронной сети в линейное пространство; (2) предиктор - принимает непрерывное представление сети в качестве входных данных и прогнозирует ошибку на тестовой выборке; (3) декодер - отображает непрерывное представление сети обратно в ее архитектуру.

Существенным отличием нашего метода 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.

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.