

Предсказание качества для процедуры выбора признаков*

Аминов Т.В.

aminov.tv@phystech.edu

¹Московский Физико-технический институт

При решении задач машинного обучения пространство признаков как правило избыточно. Для уменьшения ошибки на тестовой выборке применяется отбор признаков. Отыскание точного решения является NP-задачей. Как правило для решения применяются эвристические или субоптимальные алгоритмы. В данной статье предлагается свести данную дискретную задачу к задаче непрерывной оптимизации. На тестовой выборке строится модель предсказания качества выборки для выбранного подмножества признаков. Решение задачи выбора признаков восстанавливается из решения непрерывной задачи. Проводится вычислительный эксперимент, чтобы сравнить результаты нового метода с уже 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. Тем не менее, не все признаки имеют отношение к решению проблемы, и иногда, в том числе нерелевантные признаки могут ухудшить эффективность обучения. Данная работа посвящена проблеме выбора подмножества признаков, дающих наилучшее приближение. Эта проблема является слишком тяжелой для нахождения точного решения, поэтому предполагается отыскание решения, близкого к точному.

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

Эта проблема возникает при распознавание лиц с помощью машинного обучения [1], Так же предполагается использование результатов статьи [2], в ней исследуется наша проблема для не очень большого количества признаков (порядка 1000). За счет огромного прироста вычислительной мощности за последние 20 лет эта задача может решаться точно за разумное время.

В нашей статье мы ставим задачу применить идею построения оптимальной архитектуры нейронной сети [2], для построения оптимального подмножества признаков. Существующие методы, независимо от того, основаны ли они на обучении с подкреплением (11) или на эволюционных алгоритмах (ЕА) [3], 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.

В качестве показателя эффективности предложенного алгоритма предполагается сравнение с уже существующими : CSO (PSO) [4], CFS [5], QPFS [6], 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] 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.
- [4] Zhaleh Manbari, Fardin Akhlaghian Tab, and Chiman Salavati. Hybrid fast unsupervised feature selection for high-dimensional data. *Expert Systems with Applications*, 124, 06 2019.
- [5] 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.
- [6] Irene Rodriguez-Lujan, Ramon Huerta, Charles Elkan, and Carlos Santa Cruz. Quadratic programming feature selection. *J. Mach. Learn. Res.*, 11:1491–1516, August 2010.

