Сокращение размерности в задачах классификации текстов: компромисс между скоростью обучения и качеством модели машинного обучения

А.В. Павельев[1], М.Е. Бурлаков[1]

[1]*Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34а, Самара, Россия, 443086*

Аннотация

Задача классификации текстов по темам является актуальной задачей современного анализа данных. При формировании признакового описания, так или иначе связанного с частотой появления слова в тексте, размерность признакового пространства получается высокой (>1000). В данной работе мы рассматриваем целесообразность применения техник сокращения размерности для моделей машинного обучения, способных работать с разреженными матрицами, и проводим численный эксперимент, доказывающий, что во многих случаях с точки зрения вычислительной сложности оптимальной стратегией является усложнение модели, а не сокращение размерности датасета.

Ключевые слова

Анализ данных, машинное обучение, классификация текстов, разреженные матрицы признаков, сокращение размерности

1. Введение

Задача классификации текстов, то есть отнесение текста к одной из заранее заданных тем – это актуальная задача анализа данных как с теоретической, так и с практической точек зрения [1]. В результате применения TF-IDF [2], для коллекции документов мы получаем следующее признаковое описание: каждый текст представляется в виде вектора, где большая часть координат нулевая, а значения нескольких координат являются неотрицательными действительными числами. Весь датасет представляют по сути своей разреженную матрицу размерности N×d. Высокая размерность признакового пространства значит, что модель машинного обучения должна быть достаточно сложна, и что обучение и использование модели потребует больших вычислительных и временных затрат [3].

Одним из часто используемых методов для решения этой проблемы – это сокращение размерности признакового пространства с помощью метода главных компонент (англ. principal component analysis - PCA) или усечённого сингулярного разложения матрицы (англ. truncated singular value decomposition - SVD).

2. Численный эксперимент

В нашей работе мы строили многоклассовый классификатор для текстов средней длины – постов в социальной сети. Использовалась следующая схема: сначала формировалась матрица объекты-признаки на основе TF-IDF, затем либо происходило дополнительное сокращение размерности, либо сразу на ней обучался классификатор XGBoost – реализация градиентного бустинга над ансамблем решающих деревьев [4] с разными гиперпараметрами. В каждом случае оценивались две величины: время обучения модели, включая понижение размерности и общее качество обученной модели (т.е. отношение верно классифицированных текстов к общему количеству в обучающей выборке). Всего использовалась сбалансированная выборка из 60000 текстов, относящихся к 6 классам. Результаты изображены на рисунках 1 и 2.

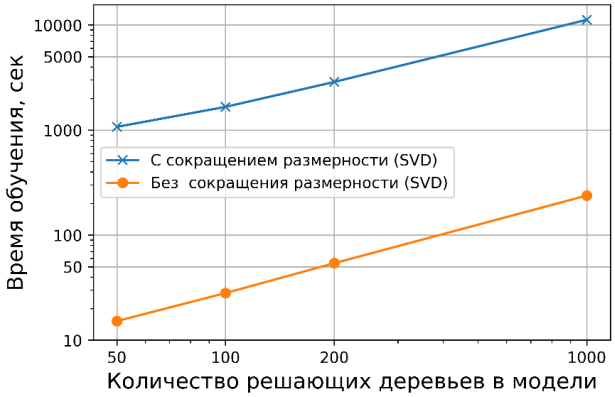

Рисунок 1: Зависимость времени обучения моделей от сложности (количества деревьев)

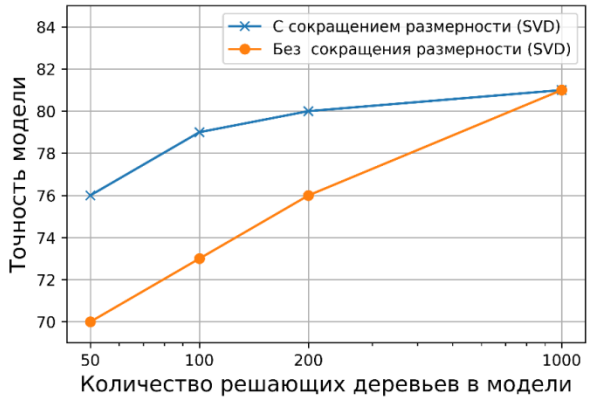

Рисунок 2: Зависимость качества модели от сложности (количества деревьев)

3. Заключение

В данной работе мы рассмотрели понижение размерности в задачах классификации текстов на основе усеченного сингулярного разложения и обсудили его главный недостаток в рамках этих задач: переход от разреженной матрицы объекты-признаки к сокращенной плотной матрице. На примере численных экспериментов по обучению моделей градиентного бустинга решающих деревьев с использованием и без использования SVD на TF-IDF признаках мы показали, что для простых моделей сокращение размерности позволяет получить лучшее качество классификации в обмен на более длительное обучение. Для сложных же моделей качество классификации будет примерно равным в обоих случаях, а выигрыш в скорости обучения и использования моделей без сокращения размерности оказался значительным.

4. Литература

[1] Dalal, M.K. Automatic text classification: a technical review / M.K. Dalal, M.A. Zaveri // International Journal of Computer Applications. – 2011. – Т. 28, № 2. – С. 37-40.

[2] Zheng, A. Feature engineering for machine learning: principles and techniques for data scientists / A. Zheng, A. Casari // O'Reilly Media, Inc., 2018.

[3] Kuo, F.Y. Lifting the curse of dimensionality / F.Y. Kuo, I.H. Sloan // Notices of the AMS. – 2005. – Т. 52, № 11. – С. 1320-1328.

[4] Chen, T. Xgboost: A scalable tree boosting system / T. Chen, C. Guestrin // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – 2016. – P. 785-794.