УДК 004.4: 519.7

Галиуллин Р.Р., Кулдавлетов А.И.
Уфимский университет науки и технологий, Уфа

Научный руководитель:
Валеев С.С.
Уфимский университет науки и технологий, Уфа

РАЗРЕЖЕННОСТЬ ДАННЫХ И СЖАТИЕ РАЗМЕРНОСТИ

Аннотация. Статья посвящена анализу разреженности данных как важного аспекта в современных задачах машинного обучения. Рассматриваются причины разреженности, ее влияние на качество моделей и методы эффективной обработки, включая сокращение размерности с использованием PCA. Подчеркивается двойственный характер разреженности как проблемы и источника структурной информации.

Ключевые слова: разреженность данных, PCA, сокращение размерности, машинное обучение, регуляризация, высокоразмерные данные, интерпретируемость моделей.

Большие данные редко бывают полными и равномерно распределенными: зачастую они содержат значительное количество нулевых, отсутствующих или малозначимых значений. Такая неоднородность структуры приводит к явлению, которое называют разреженность данных, т.е. значительная часть информации фактически не несет полезной нагрузки или просто отсутствует.

Разреженность данных является одной из ключевых проблем современной аналитики и машинного обучения. Она напрямую влияет на качество и скорость работы алгоритмов, а также на объем ресурсов, необходимых для хранения и обработки информации. Однако сама по себе разреженность не всегда негативное явление: в ряде случаев она может быть использована в качестве источника структурных признаков, отражающих скрытую закономерность в данных. Поэтому важно не только уметь определять степень разреженности, но и грамотно работать с такими структурами.

Актуальность темы обусловлена тем, что современные интеллектуальные системы – от рекомендательных алгоритмов до нейросетей – часто работают с разреженными данными. Понимание природы разреженности, ее последствий и методов обработки необходимо для разработки устойчивых, производительных и интерпретируемых моделей. Важно учитывать не только форму, но и причины ее возникновения.

Разреженность – это свойство данных, при котором большинство элементов являются нулями, пропущенными или неинформативными значениями. Такая структура характерна для анализа больших данных и

проявляется по-разному: от отсутствующих значений до неравномерного распределения признаков. Осознание причин позволяет лучше понять влияние разреженности на обучение моделей и качество анализа.

Характерный пример – матрицы взаимодействия в рекомендательных системах, где пользователи взаимодействуют лишь с небольшой частью объектов, оставляя большинство ячеек пустыми. Аналогично, в задачах обработки текста модель «мешка слов» создает векторы, заполненные нулями, так как каждый документ содержит лишь малую долю слов из общего словаря. То же наблюдается в высокоразмерных данных, где большинство признаков неактивны для конкретного объекта. Однако разреженность – это не только вызов, но и источник ценной информации. Перейдем к ее положительным аспектам.

Причинами разреженности часто становятся как природа самих данных, так и особенности способов их представления. Например, в пользовательском поведении отсутствует физическая возможность охватить все варианты выбора, а в текстах – естественное ограничение на объем и тематику. Кроме того, разреженность может возникать искусственно, в результате использования обобщенных моделей и признакового кодирования.

Разреженность данных может восприниматься как нежелательное явление, однако ее влияние на анализ и обучение моделей носит двойственный характер. С одной стороны, высокая доля нулевых или пропущенных значений может существенно затруднить извлечение закономерностей. Алгоритмы машинного обучения, сталкиваясь с неполными или нерепрезентативными признаками, подвержены переобучению, особенно в условиях ограниченного объема выборки. Кроме того, обработка разреженных структур требует увеличенных вычислительных ресурсов при использовании стандартных представлений, так как даже пустые элементы необходимо учитывать в памяти и при выполнении операций. Одним из таких методов является сокращение размерности, для чего часто используют метод главных компонент.

С другой стороны, разреженность может быть интерпретирована как форма упорядоченности. В ряде задач она помогает акцентировать внимание на наиболее значимых признаках, отбрасывая избыточные данные. При кооректной реализации, например, с использованием специализированных форматов хранения или методов отбора признаков, разреженные данные позволяют существенно сократить объем памяти и ускорить вычисления. Более того, такие структуры оказываются особенно полезными при построении интерпретируемых моделей, где важны именно редкие, но информативные признаки.

Эффективная работа с разреженными данными требует применения специализированных методов, направленных на уменьшение избыточности и повышение информативности признаков. Одним из базовых подходов является удаление нерелевантных или слабо

вариативных признаков, которые не вносят значимого вклада в обучение модели. Это позволяет сократить размерность пространства и снизить риск переобучения. В случае наличия пропущенных значений возможны различные стратегии их обработки: от простых методов заполнения средними значениями до использования моделей, прогнозирующих недостающие данные на основе имеющихся.

Важным инструментом является регуляризация, особенно в задачах с большим количеством признаков. Методы L1-регуляризации, например, не только уменьшают переобучение, но и способствуют автоматическому отбору наиболее значимых переменных, обнуляя коэффициенты малозначимых. Кроме того, при работе с разреженными структурами необходимо учитывать особенности хранения данных. Применение специализированных форматов, таких как разреженные матрицы, позволяет существенно снизить объем используемой памяти и ускорить выполнение линейной алгебры и других операций. Эти подходы особенно актуальны при построении масштабируемых моделей, работающих с большими и высокоразмерными наборами данных.

Метод главных компонент (PCA) является одним из наиболее распространенных подходов к снижению размерности и сжатию информации в задачах анализа данных. Его основная цель заключается в поиске новых ортогональных признаков – главных компонент – которые максимизируют дисперсию исходных данных. Эти компоненты представляют собой линейные комбинации исходных признаков и позволяют сосредоточить основную информацию в меньшем числе измерений. Такой подход особенно полезен при наличии разреженности, поскольку позволяет отсеять признаки, не вносящие существенного вклада в структуру данных.

В контексте разреженных матриц PCA способствует выявлению скрытых зависимостей между признаками, что позволяет заменить исходное высокоразмерное пространство более компактным и информативным представлением. Это особенно актуально для задач машинного обучения, где избыточная размерность может не только замедлять обучение, но и ухудшать обобщающую способность модели. Применение PCA снижает вычислительные затраты и способствует повышению устойчивости алгоритмов к шуму.

Сокращение объема данных за счет уменьшения размерности признакового пространства играет важную роль в построении эффективных моделей машинного обучения. Высокая размерность приводит к проклятию размерности: расстояния между объектами теряют информативность, выборка становится разреженной, ухудшается обобщающая способность и возрастает риск переобучения. Сокращение числа признаков до информативного минимума повышает устойчивость модели к шуму, улучшает интерпретируемость.

Дополнительно уменьшение числа признаков ускоряет обучение: снижается вычислительная сложность, объем памяти и время настройки параметров. Это особенно важно при использовании ресурсоемких методов, таких как нейронные сети, где обучение на сжатом наборе может быть существенно быстрее. Таким образом, сокращение объема данных повышает качество модели и делает обучение более практичным и масштабируемым.

Разреженность данных, характерная для многих современных наборов, существенно влияет на эффективность аналитических моделей. При правильной обработке она становится источником структурной информации, позволяющей сжимать данные без потери информативности. Методы, такие как PCA и регуляризация, повышают скорость и устойчивость обучения. Эффективная работа с разреженными структурами – ключевой элемент построения масштабируемых и надежных интеллектуальных систем.

Список использованных источников:

1. Паттерсон Д. Глубокое обучение с точки зрения практика / Д. Паттерсон, А. Гибсон. М.: ДМК Пресс, 2018. 418 с. ISBN 978-5-97060-481-6 // Лань: электронно-библиотечная система. URL: https://e.lanbook.com/book/116122 (дата обращения: 13.04.2025). URL: для авториз. пользователей.

2. Разреженный файл // Википедия. [2023]. Дата обновления: 02.12.2023. URL: https://ru.wikipedia.org/?curid=471221&oldid=134609381 (дата обращения: 13.04.2025).

3. Шалев-Шварц Ш. Идеи машинного обучения: учебное пособие / Ш. Шалев-Шварц, Ш. Бен-Давид; перевод с англ. А.А. Слинкина. М.: ДМК Пресс, 2019. 436 с. ISBN 978-5-97060-673-5 // Лань: электронно-библиотечная система. URL: https://e.lanbook.com/book/131686 (дата обращения: 13.04.2025). URL: для авториз. пользователей.

Galiullin R.R., Kuldavletov A.I.
Ufa University of Science and Technology, Ufa

Scientific supervisor:
Valeev S.S.
Ufa University of Science and Technology, Ufa

DATA SPARSITY AND DIMENSIONALITY REDUCTION

Abstract. The article is devoted to the analysis of sparsity of data as an important aspect in modern machine learning tasks. The reasons for sparsity, its impact on the quality of models, and effective processing methods, including

dimensionality reduction using PCA, are considered. The dual nature of sparsity as a problem and a source of structural information is emphasized.

Keywords: sparsity of data, PCA, dimension reduction, machine learning, regularization, high-dimensional data, interpretability of models.