**Суровцев Степан Сергеевич,** магистрант, 2 курс, инженерная академия, кафедра механики и процессов управления, Российский университет дружбы народов им. Патриса Лумумбы, г. Москва.

**Суровцева Майя Евгеньевна**, АО "Лаборатория Касперского", г. Москва.

# СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ КЛАССИФИКАЦИИ ПРИ РАБОТЕ С РАЗНОРОДНЫМИ НАБОРАМИ ДАННЫХ

**Аннотация**

В статье рассматриваются особенности применения классических и современных алгоритмов классификации к разнородным наборам данных, включающим признаки разного типа, различную размерность и неоднородное распределение классов. На основе анализа публикаций систематизированы сильные и слабые стороны логистической регрессии, деревьев решений, ансамблевых методов, метода опорных векторов, алгоритма k ближайших соседей и нейронных сетей в условиях гетерогенных табличных и текстовых данных. Рассматриваются типичные проблемы предобработки (кодирование категориальных признаков, обработка пропусков, дисбаланс классов) и их влияние на выбор алгоритма и качество модели. Сформулированы рекомендации по выбору методов классификации для различных типов разнородных наборов данных.

**Ключевые слова:** алгоритмы, классификация, разнородные данные, табличные данные, деревья решений, ансамблевые методы, метод опорных векторов, k ближайших соседей, нейронные сети, дисбаланс классов.

**Annotation**

The article discusses the specifics of applying classical and modern classification algorithms to heterogeneous datasets that include features of different types, varying dimensionality, and imbalanced class distributions. Based on a review of publications, the strengths and weaknesses of logistic regression, decision trees, ensemble methods, support vector machines, the k-nearest neighbors algorithm, and neural networks are systematized under conditions of heterogeneous tabular and text data. Typical preprocessing challenges (categorical feature encoding, missing value handling, class imbalance) and their impact on algorithm selection and model performance are discussed. Recommendations for selecting classification methods for various types of heterogeneous datasets are provided.

**Keywords:** algorithms, classification, heterogeneous data, tabular data, decision trees, ensemble methods, support vector machine (SVM), k-nearest neighbors (KNN), neural networks, class imbalance.

Промышленные датасеты редко бывают аккуратными. Наиболее типовая картина: числовые сенсорные показания в одних столбцах, категориальные метки – в других, где-то бинарные флаги, а рядом – текстовые логи и эмбеддинги, вытащенные из сигналов или изображений. Такой набор данных и называют разнородными, или гетерогенными данными. И именно он ломает предположения, заложенные в основу большинства базовых алгоритмов классификации. Существует масса ориентиров по классификации, но большинство из них честны лишь наполовину: авторы берут однородные наборы – сплошные числа или сплошные категории – и выдают выводы, которые неплохо смотрятся в таблицах, но разваливаются при контакте с реальным рабочим датасетом. Потому что гетерогенная таблица – это другой случай. KNN начинает страдать от смешанных метрик расстояния, метод опорных векторов (SVM) требует аккуратной нормализации разнотипных признаков, а нейронные сети на табличных данных с дисбалансом классов без

дополнительной доработки часто проигрывают банальному градиентному бустингу.

В данном исследовании раскроем ответ на главный вопрос: при каких конкретных условиях деревья решений, ансамблевые методы и остальные участники забега реально выигрывают – а не просто выглядят убедительно на синтетических примерах.

Гетерогенный датасет – это не просто «разные колонки». Это случай, когда в одной таблице одновременно находятся вещественные числа с нормальным распределением, порядковые категории, бинарные флаги и сырой текст, а у части строк половина значений просто отсутствует. Распределения признаков могут существенно различаться: в одном столбце наблюдается плавная кривая, тогда как в соседнем присутствует тяжёлый хвост с выбросами, выходящими за пределы трёх стандартных отклонений. Чаще всего данный эффект еще не сопровождается дисбалансом классов, а он почти всегда есть. Именно поэтому предобработка в таких задачах – не рутина, а архитектурное решение. Выбор между one-hot и target encoding для категорий напрямую влияет на то, как дерево решений разобьёт пространство признаков. Стратегия заполнения пропусков – удалить строки, заполнить медианой или восстановить через модель – меняет распределение обучающей выборки и, соответственно, результаты сравнения алгоритмов. SMOTE для борьбы с дисбалансом даёт одну картину метрик, метод балансировки датасета при дисбалансе классов – совсем другую.

Однако, главная проблема здесь не техническая. Если предобработка не согласована с предположениями конкретного алгоритма, сравнение методов теряет смысл. KNN, SVM и градиентный бустинг живут в принципиально разных мирах относительно масштабирования и типов входных данных. Сравнивать их «в лоб» на сырых гетерогенных данных не предоставит

никакого эффективного результата, ни по скорости обработки, ни по итоговому набору данных.

**Логистическая регрессия: интерпретируемость в обмен на гибкость**

Логистическая регрессия сохраняет позиции базового инструмента классификации – и на это есть веские причины. Интерпретируемые коэффициенты, предсказуемое поведение при L1/L2-регуляризации, низкие вычислительные затраты. Для бинарной классификации на однородных числовых данных это надёжный и воспроизводимый эталон, но граница применимости наступает быстро. При работе с реально гетерогенными данными – категориальными признаками высокой кардинальности, нелинейными зависимостями, значительной долей пропусков – линейное допущение о разделимости перестаёт выполняться. Для заменых сложных или неточных объектов, данных или зависимостей нелинейных структур необходимо явное конструирование полиномиальных признаков и взаимодействий, что ведёт к росту размерности и снижению обобщающей способности модели.

**Деревья решений: естественная работа со смешанными типами признаков**

Алгоритмы Classification and Regression Trees (CART) и C4.5 обрабатывают разнородные данные без предварительной нормализации и специальной подготовки признаков. Модель самостоятельно строит иерархию пороговых разбиений для числовых и номинальных переменных, формируя интерпретируемые правила классификации. На ряде гетерогенных признаков одиночное дерево превосходит логистическую регрессию и наивный байесовский классификатор – в первую очередь за счёт отсутствия жёстких предположений о распределении входных данных. Однако высокая дисперсия структуры остаётся системной проблемой. Незначительные изменения в обучающей выборке приводят к существенным перестройкам дерева,

особенно при большом числе признаков и зашумлённых данных. На высокоразмерных гетерогенных наборах склонность к переобучению выражена наиболее сильно.

**Ансамблевые методы: практический стандарт для табличных задач**

Случайный лес решает проблему дисперсии одиночного дерева через агрегацию независимых моделей. Градиентный бустинг (например, XGBoost, LightGBM или CatBoost) идёт дальше, последовательно минимизируя остаточные ошибки предыдущих итераций. На табличных данных с гетерогенной структурой признаков именно бустинг демонстрирует стабильно высокое качество. Такое заключение подтверждается эмпирически. Сравнительные исследования на сотнях датасетов из открытых хранилищ данных показывают: градиентный бустинг по деревьям систематически превосходит линейные модели и нейронные сети в задачах классификации на классических табличных данных. CatBoost демонстрирует особую эффективность при высокой кардинальности категориальных признаков за счёт встроенного метода кодирования категориальных признаков ordered target encoding. При этом качество ансамблей напрямую зависит от настройки гиперпараметров. Без систематического поиска по пространству скорости обучения (learning_rate), максимальной глубины дерева (max_depth), доли выборки для каждого дерева (subsample) и смежных параметров высок риск получить модель с существенным разрывом между метриками на обучающей и тестовой выборках.

**Метод опорных векторов: высокая точность при строгих условиях предобработки**

SVM с RBF-ядром (Гауссово ядро) способен строить сложные нелинейные разделяющие поверхности и показывает высокие результаты на компактных, качественно предобработанных наборах данных. Однако, для гетерогенных данных метод предъявляет более жёсткие требования:

масштабирование каждого признака, продуманное кодирование категориальных переменных, тщательный подбор параметров C и γ.

На больших датасетах с высокой размерностью для SVM появляются вычислительные ограничения: время обучения растёт квадратично с числом объектов, что делает метод практически неприменимым для выборок свыше нескольких сотен тысяч записей. В этом контексте ансамблевые методы выигрывают как по качеству классификации, так и по масштабируемости.

**Метод k ближайших соседей: структурные ограничения в гетерогенном пространстве**

Метод KNN включают в сравнительные исследования в качестве базовой непараметрической модели и на малоразмерных задачах с корректно подобранной метрикой расстояния он действительно конкурентоспособен, но относительно работы с гетерогенными даннымы возникает принципиальная структурная проблема.

Совместное использование числовых признаков разного масштаба, номинальных переменных и текстовых представлений приводит к вырождению евклидова расстояния: различия между ближайшими и дальними соседями статистически нивелируются. Это классическое проявление «проклятия размерности», усиленное типовой неоднородностью признакового пространства. Дополнительным ограничением служит необходимость хранения полной обучающей выборки и линейный рост затрат на поиск соседей при процессе использования обученной модели для принятия решений на основе новых данных.

**Нейронные сети: эффективны при достаточном объёме данных и сложных взаимодействиях**

Полносвязные сети и специализированные архитектуры для табличных данных (например, TabNet, FT-Transformer) демонстрируют высокую эффективность в задачах с изображениями, текстом и временными рядами.

Для классических гетерогенных табличных данных картина менее однозначна. Результаты сравнительных тестов за 2021–2023 фиксируют устойчивую закономерность: градиентный бустинг превосходит нейронные сети на большинстве табличных наборов, особенно при объёме выборки до 50–100 тысяч объектов. Нейронные сети требуют продуманного выбора архитектуры, стратегии эмбеддинга категориальных признаков и нормализации числовых переменных и, как правило, существенно большего объёма данных для устойчивой сходимости.

Область, где нейросетевой подход оправдан: наборы данных с большим объёмом, сложными нелинейными взаимодействиями между признаками и достаточными ресурсами для полноценного подбора архитектуры. В остальных сценариях ансамблевые методы остаются более надёжным и воспроизводимым выбором.

**Стандартная схема эксперимента – и её слабые места**

Сравнительные исследования алгоритмов классификации опираются на широкий круг открытых датасетов: медицинские данные, задачи биоинформатики, обнаружения сетевых вторжений, тематических текстов. Репозиторий UCI, коллекция KDD Cup 99, специализированные диагностические наборы – всё это давно стало стандартным полигоном для проверки моделей.

Стандартная схема эксперимента выглядит так: берётся несколько датасетов разного размера и структуры, проводится унифицированная предобработка, обучается набор алгоритмов (логистическая регрессия, деревья решений, случайный лес, бустинг, SVM, kNN, иногда нейронные сети) и результаты оцениваются по accuracy, F1, ROC-AUC и времени обучения с обязательной кросс-валидацией. Однако, в подобных исследованиях кроется системная проблема. «Унифицированная предобработка» на практике означает очень разные вещи у разных авторов. Один использует one-hot для категорий, другой – target encoding, третий вообще может удалить

категориальные признаки. При этом выбор стратегии кодирования напрямую влияет на то, какой алгоритм окажется сильнее. Сравнение становится валидным только тогда, когда предобработка согласована с предположениями каждого конкретного метода.

**Что говорят данные: устойчивые закономерности**

Из накопленных исследований вырисовывается несколько чётких паттернов – и они достаточно стабильны, чтобы на них опираться.

Одиночные деревья решений, наивный байес и kNN редко выходят в лидеры на сложных гетерогенных наборах. Их ценность в другом: быстрый старт, читаемая логика, минимальные требования к предобработке. Для первичного анализа или задач с жёсткими требованиями к интерпретируемости – вполне рабочий выбор.

Ансамблевые методы на деревьях занимают верхние строчки рейтингов на разнородных табличных данных последовательно и воспроизводимо. Причём в ряде работ показано, что гетерогенные ансамбли – те, которые объединяют несколько принципиально разных базовых алгоритмов, – дополнительно выигрывают у однородных. Разнообразие базовых моделей здесь работает как реальное преимущество, а не методологическая экзотика.

Метод опорных векторов держится уверенно при умеренной размерности и качественно подобранных гиперпараметрах, но при росте числа признаков и усилении гетерогенности «стоимость» подбора параметров и обучения начинает расти непропорционально быстро. В промышленных сценариях с большими объёмами это уже не просто неудобство – это реальное ограничение применимости.

Нейронные сети на классических табличных данных не показывают стабильного превосходства над бустингом, особенно при выборках до 100 тысяч объектов. Исключением являются комбинированные данные с

текстовыми или мультимодальными признаками: специализированные архитектуры типа FT-Transformer получают своё преимущество.

Корректная импутация пропусков, согласованное кодирование категорий и грамотная работа с дисбалансом классов способны перевернуть итоговый рейтинг алгоритмов. Метод, который в литературе выглядит явным фаворитом, на конкретном датасете с небрежной предобработкой может потерять 10–15 процентных пунктов по F1 – и оказаться в хвосте таблицы.

**Заключение**

Гетерогенные данные не прощают универсальных решений. Разные типы признаков, шум, пропуски, дисбаланс классов – каждый из этих факторов способен перевернуть рейтинг алгоритмов на конкретном датасете. Никакого «лучшего метода на все случаи» не существует, и литература это подтверждает однозначно, но, если нужна отправная точка – ансамблевые методы на деревьях решений остаются наиболее надёжным выбором для разнородных табличных данных. Логистическая регрессия, SVM и нейронные сети полезны как дополнительные модели для валидации и учёта специфики предметной области.

Следующий рубеж – алгоритмы, нативно работающие со смешанными представлениями: таблица + текст + граф в одной модели. И автоматизация подбора предобработки под конкретную структуру данных. Подобное решение уже нельзя отнести к академической задаче – это скорее нужно в крупном проекте.

**Литература**

1. A Comparative Analysis of Classification Algorithms on Diverse Datasets // Engineering, Technology & Applied Science Research. 2018. Vol. 8, № 2. P. 2790-2795.

2. Петровский М.И., Глазкова В.В. Алгоритмы машинного обучения для задачи анализа и рубрикации электронных документов // Вестник Новосибирского государственного университета. 2006. Т. 4, № 4. С. 57-72.
3. Ayidagn K.A., Gite S. Analysis of Feature Selection Algorithms and a Comparative Study on Heterogeneous Classifier for High Dimensional Data Survey // International Journal of Engineering Trends and Technology. 2017. Vol. 53, № 2. P. 59-63.
4. Основы классификации данных с использованием алгоритмов машинного обучения // Science and Education. 2024. № 3. С. 15-25.
5. Comparative Analysis of Data Mining Classification Algorithm Performance for Searching Prospective Student Interests // International Journal of Information Systems and Technology. 2022. Vol. 6, № 1. P. 45-54.
6. Сравнительный анализ алгоритмов классификации и рубрикации текстовых документов. Донецк: ДонНТУ, 2018. 45 с.
7. Comparative Study of Machine Learning Algorithm in Data Classification. Final Year Project Report. UTAR, 2025. 78 p.
8. Сравнительный анализ производительности алгоритмов мультиклассовой классификации // Труды конференции MLSD. 2024. С. 1-8.
9. A Comparison of Classification Methods across Different Data Complexity Scenarios and Datasets // Expert Systems with Applications. 2021. Vol. 167.
10. A Comparative Study of Machine Learning Algorithms for Tabular Data Classification // International Journal of Engineering Development and Research. 2025. Vol. 13, № 4. P. 210-218.
11. Classification Algorithm for Heterogeneous Network Data Streams Based on Big Data Active Learning // Computational Intelligence and Neuroscience. 2022. №2. P.1-10.
12. The Heterogeneous Ensembles of Standard Classification Algorithms (HESCA): the Whole Is Greater than the Sum of Its Parts // arXiv preprint arXiv:1710.09220. 2017.

**Literature**


1. A comparative analysis of classification algorithms on diverse datasets // Engineering, Technology &
   Applied Science Research. 2018. Vol. 8, No. 2. P. 2790–2795.
2. Petrovsky M.I., Glazkova V.V. Machine learning algorithms for the task of a nalysis and categorization of electronic documents //
   Novosibirsk State University Bulletin. 2006. Vol. 4, No. 4. P. 57–72.
3. Ayidagn K.A., Gite S. Analysis of feature selection algorithms and a compar ative study on heterogeneous classifier for high-dimensional data survey //
   International Journal of Engineering Trends and Technology. 2017. Vol. 53, No. 2. P. 59–63.
4. Fundamentals of data classification using machine learning algorithms //
   Science and Education. 2024. No. 3. P. 15–25.
5. Comparative analysis of data mining classification algorithm performance fo r searching prospective student interests //
   International Journal of Information Systems and Technology. 2022. Vol. 6, No. 1. P. 45–54.
6. Comparative analysis of text document classification and categorization algo rithms. Donetsk: DonNTU, 2018. 45 p.
7. Comparative study of machine learning algorithm in data classification //
   Final Year Project Report. UTAR, 2025. 78 p.
8. Comparative analysis of multiclass classification algorithm performance //
   Proceedings of the MLSD Conference. 2024. P. 1–8.
9. A comparison of classification methods across different data complexity sce narios and datasets // Expert Systems with Applications. 2021. Vol. 167.
10. A comparative study of machine learning algorithms for tabular data classifi cation //
    International Journal of Engineering Development and Research. 2025. Vol. 13, No. 4. P. 210–218.

11. Classification algorithm for heterogeneous network data streams based on bi g data active learning //
Computational Intelligence and Neuroscience. 2022. No. 2. P. 1–10.
12. The heterogeneous ensembles of standard classification algorithms (HESCA ): The whole is greater than the sum of its parts //
arXiv preprint arXiv:1710.09220. 2017.