ВЕСТНИК НАУКИ
И ТВОРЧЕСТВА

СРАВНЕНИЕ МЕТОДОВ МАСШТАБИРОВАНИЯ ПРИЗНАКОВ: STANDARDSCALER, MINMAX, ROBUST

Клюковкин Георгий Константинович,
ведущий инженер-программист,
RingCentral, г. Санкт-Петербург

E-mail: kliukovkin@gmail.com

Аннотация. В данной статье проведён сравнительный анализ трёх широко применяемых методов масштабирования признаков – StandardScaler, MinMaxScaler и RobustScaler. В работе описана теоретическая основа каждого метода, проанализированы их преимущества и ограничения, представлены визуальные и табличные сравнения, а также даны практические рекомендации по выбору метода в зависимости от структуры данных и типа модели. На основании полученных результатов выделены сценарии применения каждого подхода и предложены направления для дальнейших исследований, включая использование нелинейных преобразований. Статья может быть полезна специалистам в области анализа данных, машинного обучения и прикладной статистики.

Ключевые слова: машинное обучение, масштабирование признаков, StandardScaler, MinMaxScaler, RobustScaler, выбросы в данных, нормализация, предобработка данных, градиентный спуск, устойчивость к выбросам, линейная регрессия, нейронные сети, методы обработки данных.

Актуальность исследования

В современных задачах машинного обучения одной из ключевых предпосылок успешной работы алгоритмов является корректная предварительная обработка данных, в том числе масштабирование признаков. Это особенно важно для моделей, чувствительных к масштабу переменных, таких как методы ближайших соседей, опорные векторы и алгоритмы кластеризации. При отсутствии масштабирования одни признаки могут доминировать над другими, искажаются расстояния, ухудшается сходимость градиентных методов и, как следствие, снижается общая точность модели.

В связи с этим возрастает интерес к изучению и сравнительному анализу различных подходов к масштабированию. Наиболее широко используемыми методами являются StandardScaler, MinMaxScaler и RobustScaler, каждый из которых имеет свои особенности и области применения. StandardScaler

предполагает нормальное распределение и чувствителен к выбросам, MinMaxScaler приводит данные к фиксированному диапазону, а RobustScaler учитывает медиану и межквартильный размах, что делает его устойчивым к выбросам. Несмотря на широкое применение этих методов, выбор наилучшего из них в конкретной ситуации до сих пор остаётся нерешённой прикладной задачей. Это обусловливает актуальность сравнительного исследования эффективности масштабирования признаков с использованием различных методов при решении задач классификации и регрессии.

Цель исследования

Цель настоящего исследования заключается в систематическом сравнении методов масштабирования признаков (StandardScaler, MinMaxScaler и RobustScaler) с точки зрения их влияния на производительность алгоритмов машинного обучения.

Материалы и методы исследования

В ходе исследования были использованы данные из открытых источников, в том числе искусственно сгенерированные датасеты и реальные выборки, демонстрирующие различия в распределениях признаков и наличии выбросов. Масштабирование осуществлялось с помощью стандартных инструментов библиотеки Scikit-learn на языке Python. Методы StandardScaler, MinMaxScaler и RobustScaler применялись к данным с различными характеристиками, включая нормальное распределение, диапазонные ограничения и выбросы.

Результаты исследования

В современной науке о машинном обучении масштабирование признаков представляет собой фундаментальный шаг предварительной обработки данных, направленный на приведение параметров модели к единому масштабу. Без нормализации или стандартизации диапазоны исходных переменных могут существенно различаться, что приводит к доминированию одних признаков над другими и искажению мер расстояния, используемых во многих алгоритмах. Это особенно критично для методов, основанных на градиентном спуске (например, линейная регрессия, нейронные сети), где неравномерность масштабов замедляет сходимость, а также для алгоритмов, измеряющих сходство через расстояния, например, KNN и SVM.

Масштабирование признаков устраняет смещение при обучении, когда параметры с большими значениями получают избыточное влияние, а также ускоряет обучение и делает регуляризацию более эффективной. Без применения такого преобразования встречаются следующие затруднения: неправильный вес признаков, замедленное или нестабильное обучение моделей, особенно при наличии выбросов и разношкальных переменных, что влияет на результаты модели и усложняет её интерпретацию [5].

На рисунке 1 представлено изменение плотности распределения двух признаков (x1 и x2) в результате применения различных методов масштабирования.

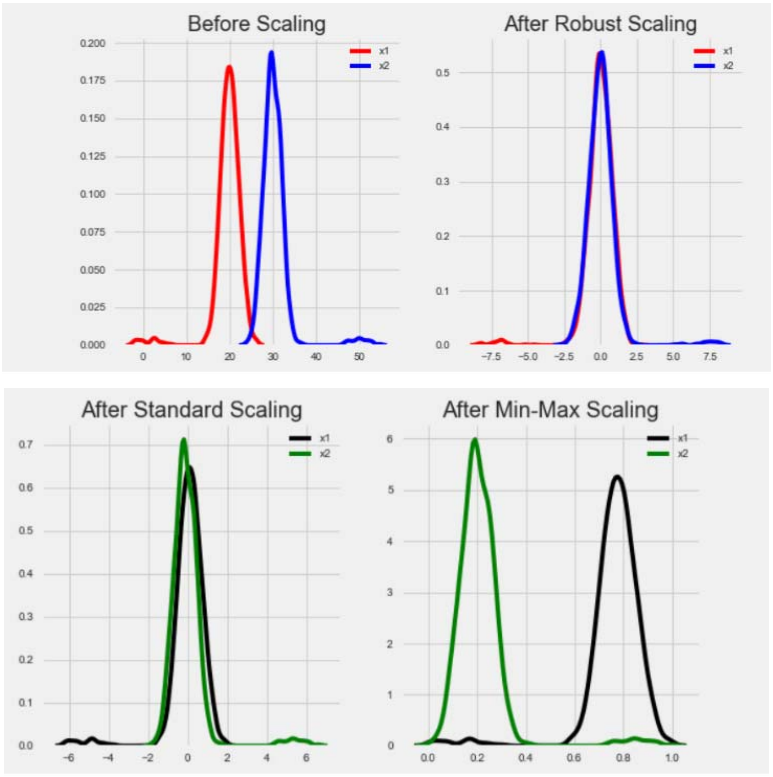

Рис. 1 Визуальное сравнение распределений признаков до и после применения методов масштабирования

Графики наглядно иллюстрируют, как разные подходы по-разному влияют на структуру данных, и подчеркивают необходимость осознанного выбора метода в зависимости от особенностей распределения и задач анализа.

Понимание необходимости масштабирования признаков и осознание возможных проблем при его отсутствии формируют прочную теоретическую базу для дальнейшего анализа. Однако знание общей концепции масштабирования недостаточно для эффективного применения на практике. Каждый конкретный метод обладает своими алгоритмическими особенностями, преимуществами и ограничениями, что обуславливает необходимость их детального рассмотрения.

1) Метод StandardScaler реализует Z-преобразование (стандартизацию), при котором каждому признаку задаются нулевое среднее и единичная дисперсия с помощью преобразования вида:

$$x' = \frac{x - \mu}{\sigma}$$

, где μ – среднее значение, σ – стандартное отклонение признака. Хотя принудительно метод не делает распределение Гауссовым, он центрирует и нормализует масштаб данных, сохраняя форму формы распределения ядра выборки [1].

Применение стандартизации требует оценки μ и σ на тренировочных данных, после чего одна и та же линейная трансформация применяется и к тестовой выборке. Такой подход предотвращает утечку информации из теста в тренировочный процесс.

Стандартизация приводит к ускорению сходимости градиентных методов благодаря выравниванию шагов обновления для различных переменных, снижению риска численных переполнений и стабилизации параметров моделей. Это делает метод эффективным для регрессии, классификации и нейронных сетей, где целенаправленное обучение параметров требует однородного масштаба данных.

Однако StandardScaler чувствителен к выбросам, так как μ и σ могут значительно меняться при наличии экстремальных значений. Это обуславливает искажение масштабирования: при наличии выбросов основная часть данных может оказаться чрезмерно сжата в относительно узкий диапазон значений.

На рисунке 2 видно, как после стандартизации основная масса данных сжимается вокруг центра, при этом хвосты выбросов остаются, но начинают искажать оценку масштаба (распределение в пределах $\pm 1\sigma$ содержит примерно 68 % значений).

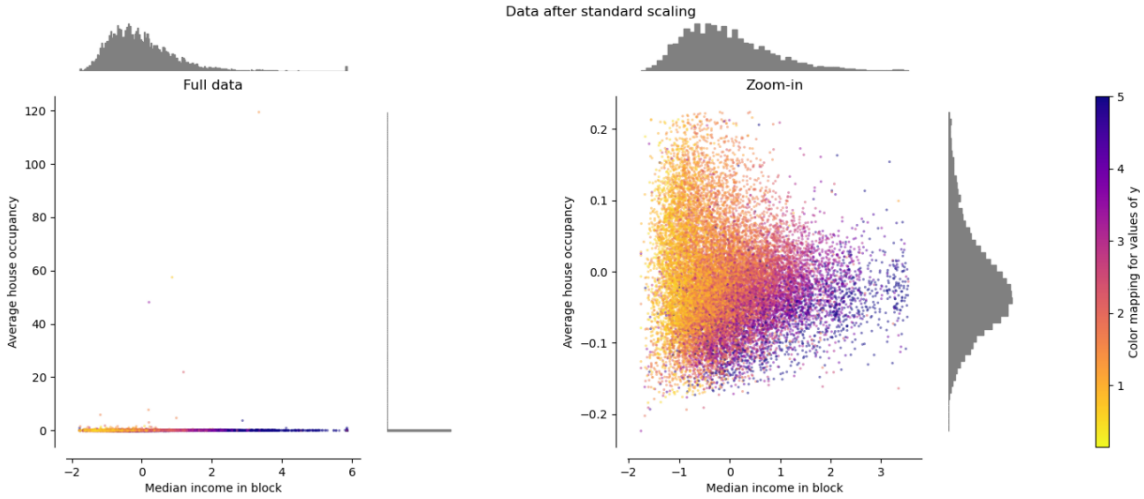

Рис. 2 Влияние стандартизации признаков (StandardScaler) на распределение данных и визуализацию кластерной структуры

Таким образом, StandardScaler рекомендуется применять в следующих случаях:

– Данные имеют близкое к нормальному распределение (или хотя бы умеренную асимметрию);

– Алгоритмы, чувствительные к масштабу (градиентный спуск, SVM, KNN, etc.);

– Отсутствие значительного числа выбросов или готовность их предварительно удалить/обработать.

2) Метод MinMaxScaler – это классический способ линейного масштабирования, переводящий каждый признак на заранее заданный диапазон (по умолчанию [0, 1]). Он рассчитывается по формуле:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)};$$

, где *min(x)* и *max(x)* – это соответственно минимальное и максимальное значение признака на обучающей выборке [2].

MinMaxScaler особо полезен в ситуациях, когда важен ограниченный диапазон входных данных – например, нейронные сети часто требуют значения признаков в диапазоне [0, 1] или [-1, +1] для нормальной работы функции активации. Также он прост в понимании и реализации и улучшает численную стабильность алгоритмов [4].

Однако метод сильно чувствителен к выбросам: крайние значения определяют диапазон, и в результате большинство наблюдений могут «сжаться» в узкий сегмент шкалы. Из-за этого внутренняя структура, особенно кластеризация, может быть искажена. Например, на примере с признаками «средний доход» и «заполняемость домов» большинство точек сосредоточено около 0, а выбросы дистанцируются на 1, что подтверждает практика.

На рисунке 3 видно, что после применения MinMaxScaler box-графики признаков уже лежат в [0, 1], однако при наличии выбросов наблюдения с аналогичной природой концентрируются в самом низу шкалы, что снижает вариативность.

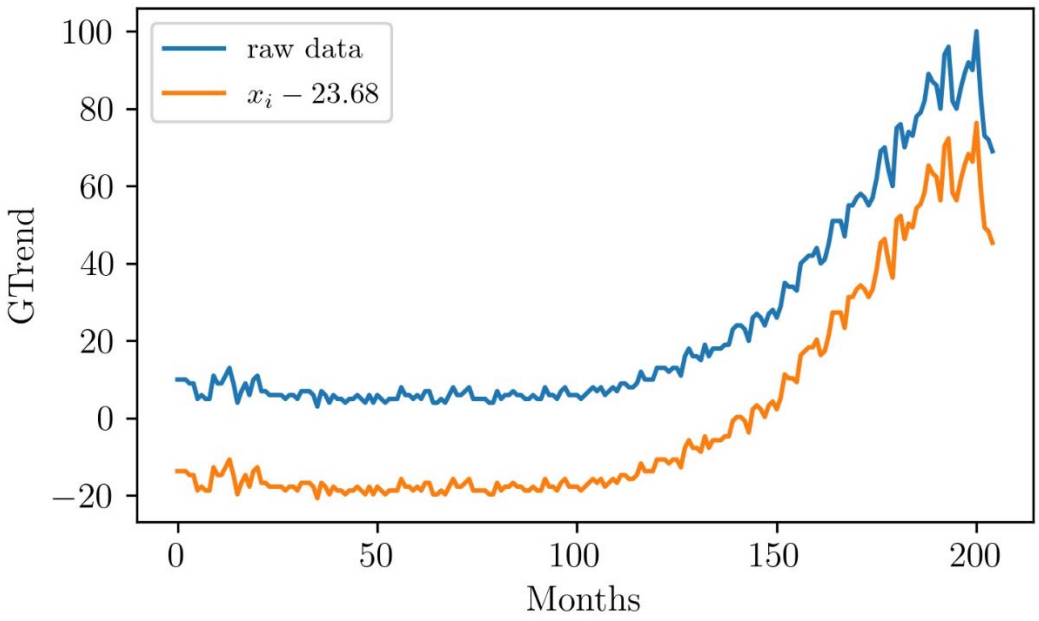

Рис. 3 Влияние масштабирования методом MinMax на временной ряд: сравнение необработанных и нормализованных данных

MinMaxScaler желательно применять:

– Когда признаки не нормальны, и требуется ограниченный диапазон ввода (например, сигнальные данные, изображения);

– При отсутствии или незначительном количестве выбросов;

– В нейронных сетях и алгоритмах, формально не требующих нормального распределения, но требующих стандартизованных входов (KNN, SVM, NN) .

3) Метод RobustScaler – это статистически устойчивый подход к масштабированию признаков, основанный на использовании медианы и межквартильного размаха (IQR) вместо среднего и стандартного отклонения. После удаления медианы и деления на IQR каждое значение преобразуется по формуле:

$$x' = \frac{x - \text{медиана}}{\text{IQR}} = \frac{x - Q_2}{Q_3 - Q_1}$$

, где Q_1, Q_2, Q_3 – соответственно 25-й, 50-й и 75-й процентили признака.

Такой подход обладает высокой устойчивостью к выбросам: медиана и IQR остаются практически неизменными при появлении экстремальных значений, в отличие от среднего и стандартного отклонения, сильно искажённых этими выбросами. Именно это делает метод предпочтительным при наличии шумных данных и экстремальных точек [3].

На рисунке 4 видно, как RobustScaler сохраняет компактность основной группы значений, игнорируя влияние выброса, в отличие от StandardScaler, смещающего масштабирование.

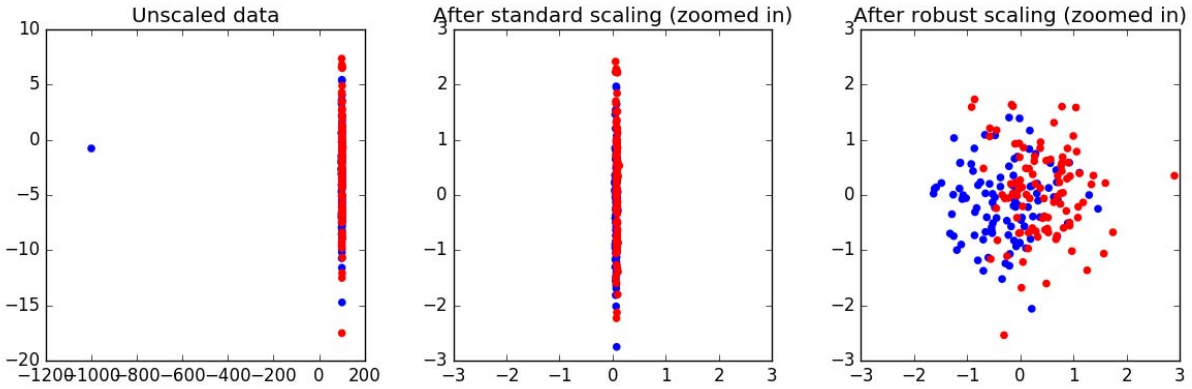

Рис. 4 Влияние различных методов масштабирования на данные с выбросами: сравнение необработанных, стандартизованных и устойчиво масштабированных данных

RobustScaler рекомендуется использовать в следующих случаях:

– Наличие значительного количества выбросов, которые нежелательно удалять.

– Требуется сохранить «устойчивые» статистические свойства данных.

– Алгоритмы чувствительны к масштабу, а обычная стандартизация и нормализация приводят к потере информативности из-за выбросов.

Таблица 1 обобщает ключевые особенности трёх наиболее широко используемых методов масштабирования данных – StandardScaler, MinMaxScaler и RobustScaler. В ней отражены применяемые математические формулы, основные преимущества и потенциальные ограничения каждого подхода.

Таблица 1

Сравнительная характеристика методов масштабирования признаков: StandardScaler, MinMaxScaler и RobustScaler

Таблица 2 систематизирует подходы к выбору метода масштабирования признаков (StandardScaler, MinMaxScaler и RobustScaler) в зависимости от структуры данных и требований к моделям машинного обучения.

Метод масштабирования	Формула преобразования	Преимущества	Недостатки
StandardScaler	$(x - \mu) / \sigma$	Центрирует данные, приводит к единичной дисперсии; ускоряет градиентные методы	Чувствителен к выбросам, нарушается масштаб при экстремальных значениях
MinMaxScaler	$(x - min) / (max - min)$	Приводит значения к фиксированному диапазону; подходит для нейронных сетей	Высокая чувствительность к выбросам; искажение плотности распределения
RobustScaler	$(x - медиана) / IQR$	Устойчив к выбросам; сохраняет структуру основной массы данных	Не нормализует до стандартной шкалы, может быть неэффективен при асимметрии

Таблица 2

Практические рекомендации по выбору метода масштабирования

Условие / Алгоритм	StandardScaler	MinMaxScaler	RobustScaler
Данные близки к нормальному распределению	Да	Возможно	Да
Требуется ограниченный диапазон [0, 1]	Нет	Да	Возможно
Наличие выбросов в данных	Нет	Нет	Да
Использование градиентного спуска (SVM, логистическая регрессия)	Да	Да	Да
Нейронные сети	Возможно	Да	Возможно
Алгоритмы, чувствительные к расстоянию (PCA, KNN)	Да	Да	Да
Деревья решений, случайный лес, бустинг	Нет	Нет	Нет

Перспективы дальнейших исследований в области масштабирования признаков заключаются в расширении сравнительного анализа за счёт включения методов, основанных на нелинейных преобразованиях, таких как квантильное преобразование, логарифмическое и степенное масштабирование, а

также нормализация на основе гауссовского отображения. Эти подходы позволяют более гибко адаптировать данные с выраженной асимметрией, нестандартными распределениями и высокой плотностью выбросов. Особый интерес представляет исследование влияния нелинейного масштабирования на обучение глубоких нейронных сетей и ансамблевых моделей, а также на устойчивость моделей в условиях зашумлённых или неполных данных.

Дальнейшие работы могут быть направлены на разработку адаптивных гибридных схем масштабирования, автоматически подбирающих метод в зависимости от структуры признака и модели, а также на создание рекомендаций для конкретных отраслевых задач, например, в биоинформатике, финтехе или промышленной аналитике.

Выводы

Проведённое исследование подтвердило, что выбор метода масштабирования признаков оказывает влияние на эффективность алгоритмов машинного обучения. Метод StandardScaler показал высокую эффективность при работе с нормально распределёнными данными без выраженных выбросов. MinMaxScaler оказался предпочтительным в ситуациях, когда требуется ограниченный диапазон значений, но продемонстрировал высокую чувствительность к экстремальным значениям. RobustScaler проявил наилучшую устойчивость к выбросам, сохраняя информативность основной массы данных, что делает его незаменимым при анализе шумных выборок.

Практические рекомендации, основанные на результатах анализа, позволяют выбирать оптимальный метод масштабирования в зависимости от свойств данных и используемой модели. В перспективе исследование может быть дополнено сравнением с методами на основе нелинейных преобразований, что позволит более гибко обрабатывать данные с выраженной асимметрией и сложной структурой распределения.

Литература:

1. Определение остатка от деления чисел в Python: методы решения [Электронный ресурс]. – Режим доступа: https://sky.pro/wiki/python/opredelenie-ostatka-ot-deleniya-chisel-v-python-metody-resheniya/.

2. Feature Scaling: MinMax, Standard and Robust Scaler – Machine Learning Geek [Электронный ресурс]. – Режим доступа: https://machinelearninggeek.com/feature-scaling-minmax-standard-and-robust-scaler/.

3. Sklearn Feature Scaling with StandardScaler, MinMaxScaler, RobustScaler and MaxAbsScaler – MLK – Machine Learning Knowledge [Электронный ресурс]. – Режим доступа: https://machinelearningknowledge.ai/sklearn-feature-scaling-with-standardscaler-minmaxscaler-robustscaler-and-maxabsscaler.

4. StandardScaler vs. MinMaxScaler vs. RobustScaler: Which one to use for your next ML project? | by Sarp Nalcin | Medium [Электронный ресурс]. – Режим доступа: https://medium.com/%40onersarpnalcin/standardscaler-vs-minmaxscaler-vs-robustscaler-which-one-to-use-for-your-next-ml-project-ae5b44f571b9.

5. The choice of scaling technique matters for classification performance [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2212.12343