УДК 004.056 **DOI** 10.15622/ia.24.2.1

И.С. Лебедев

АДАПТИВНОЕ ПОСТРОЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ НА ОСНОВЕ АНАЛИЗА ФУНКЦИОНАЛА КАЧЕСТВА ОБРАБОТКИ СЕГМЕНТОВ ПОСЛЕДОВАТЕЛЬНОСТИ

Лебедев И.С. **Адаптивное построение регрессионных моделей на основе анализа функционала качества обработки сегментов последовательности.**

Аннотация. Рассмотрена задача адаптивного построения модели, направленной на повышение показателей качества обработки информационных последовательностей. В методах обработки данных, которые нашли применение во многих прикладных областях, применяемый анализ объектов наблюдения является вычислительно ресурсоемким и в случае изменения свойств данных, требует большого количества итераций. В статье предложена методика выбора сегментов информационной последовательности, полученных разными способами, отличающаяся использованием функционала качества регрессионных моделей обработки подпоследовательностей. Поступающие на вход последовательности объектов наблюдения подвергаются разделению различными предварительно заданными алгоритмами сегментации. На каждом полученном сегменте обучаются заранее выбранные регрессионные модели и, в зависимости от полученных значений вычисленного функционала качества, происходит назначение лучших по качественным показателям моделей на сегменты. Это позволяет формировать агрегационную модель обработки данных. На основе эксперимента на модельных данных и выборках проведена оценка предлагаемой методики. Получены значения показателя качества MSE и MAE для разных алгоритмов обработки и при различном количестве сегментов. Предлагаемая методика дает возможность повысить показатели MSE и MAE за счет сегментации и назначения регрессионных моделей, которые имеют наилучшие показатели на отдельных сегментах. Предложенное решение направлено на дальнейшее усовершенствование ансамблевых методов. Его применение позволяет повысить оперативность настройки базовых алгоритмов в случае трансформации свойств данных и улучшить интерпретируемость результатов. Методика может применяться при разработке моделей и методов обработки информационных последовательностей.

Ключевые слова: машинное обучение, адаптивные модели, повышение качества обработки, регрессионные модели.

1. Введение. В условиях лавинообразного роста информационных потоков, вызванного внедрением средств сбора и накопления данных во всевозможные процессы, создание эффективных моделей при решении задач регрессии, предсказания поведения системы является проблемным вопросом. Анализ информационных последовательностей, полученных в ходе функционирования разнородных объектов наблюдения, является сложной задачей, поэтому в целях автоматизации действий, направленных на повышение качественных показателей обработки данных применяются саморегулирующиеся вычисления. Их относят к методам, позволяющим автоматически реагировать на динамически изменяющиеся свойства входных данных. Саморегулирующиеся

вычисления могут значительно повысить качественные показатели обработки входных регрессионных последовательностей [1].

Одним из часто используемых методов машинного обучения, применяемых для предсказания поведения системы, является линейная регрессия. В случае наличия в данных случайного статистического шума и близких к линейным зависимостей она позволяет получить достаточно качественное приближение, но при резких изменениях наблюдаемых значений такая модель обработки становится не адекватной. Решение обозначенного вопроса может быть выполнено на основе кусочно-линейной аппроксимации, предполагающей применение сегментированной регрессии. Происходит разделение последовательности данных на отдельные сегменты, а затем на каждом из них строится линейная регрессия [2]. В результате последовательность аппроксимируется набором прямых, ограниченных сегментами.

В случае «линейности» данных сегментированная линейная регрессия может обеспечить хорошую аппроксимацию. Однако для эффективной реализации необходимо решать ряд вопросов, связанных с выбором сегментов. Заданная длина последовательности, количество сегментов, выбросы и шумы, свойства данных могут существенно оказывать влияние на результат. Возникает большое количество возможных вариантов разбиения [3].

При обработке регрессионных зависимостей очень часто возникают ситуации, когда взаимосвязь между ответом и объясняющими переменными резко меняется в определенных точках [4]. Такие точки проявляются во многих областях. Например, в социологии различные срезы общественного мнения, в медицине риски заболеваний могут существенно отличаться от возраста людей, в энергетике генерация электроэнергии – зависеть от климатических, температурных условий, в анализе состояния информационных систем и сетей при определенных внешних воздействиях может проявляться резкий рост информационных, служебных сообщений. Для обнаружения точек, где взаимосвязь между переменными резко меняется, применяются методы на основе сплайнов и методы кусочной регрессии. В первом случае задача состоит в том, чтобы найти наиболее подходящую модель для прогнозировании или для поиска экстремальных значений [5 – 7], а во втором – кусочная регрессия направлена на выявление закономерностей между ответными и объясняющими переменными на интервалах ограниченных точками изменений.

Обнаружение точек изменения является одной из важных проблем для регрессионного анализа. Ее решение зависит от обработки шумов, выбросов, аномальных значений выборки данных. Возникает задача

анализа свойств данных в целях формирования оптимальных выборок для обучения моделей и их последующего использования. Кроме того, существует большое разнообразие методов и способов разбиения последовательностей, которые обуславливают необходимость применения анализа результатов вычислений для оценки и выбора способов, позволяющих достичь оптимальных качественных показателей.

2. Обзор существующих методов. Задачи прогнозирования, регрессии решаются с помощью различных методов машинного обучения. Одними из наиболее популярных моделей обработки являются нейронные сети. На сегодняшний день для решения задач используются архитектуры CNN, RNN, GAN, DNN, которые совершенствуются [8]. Постоянно происходит их развитие, разрабатываются новые нейросетевые модели, например KAN. Они обладают свойствами адаптивности, гибкости настройки на решение предметно ориентированных задач, имеют большие возможности по расширению и масштабируемости, позволяют обеспечить высокую скорость обработки данных. Однако при их использовании возникает ряд проблемных вопросов вычислительной сложности, ограничения памяти при обучении на больших наборах данных, связанных с "исчезающими" и "взрывными" градиентами, стабильностью обучения, необходимостью построения архитектуры обработки и хранения данных при решении масштабных задач информационного анализа. Несмотря на все преимущества, позволяющие вычислять скрытые закономерности, по-прежнему для обучения нейросети требуется большое количество данных, что не всегда возможно обеспечить.

В общем виде нейросеть можно рассматривать как суперпозицию функций, где роль функций выполняют нейроны и их совокупности [9 – 11]. В рамках такого представления появляется возможность в качестве альтернативы рассмотреть замену суперпозиции функций, основанной на нейронах, функциями – моделями обработки данных, например реализующими алгоритмы наивного Байеса, деревьев решений, опорных векторов или других нейросетевых структур, последовательность использования которых будет приводить к такому же решению задачи, как и с помощью нейросети, имеющей "сложное" построение.

Определяя правила последовательного применения менее сложных и менее ресурсоемких моделей по отдельности, выбора наиболее подходящих алгоритмов в зависимости от свойств данных и текущих процессов обработки возможно получить общую модель, обладающую лучшей интерпретируемостью результатов, меньшей ресурсоемкостью, и достигающую соизмеримых с нейросетью результатов. Однако выбор

последовательности обработки таких моделей является нетривиальной задачей, подразумевающей оптимизацию протекающих в них процессов.

Каждая из моделей оптимизируется под определенные свойства последовательностей данных. Причем процессы оптимизации методов осуществляются по двум основным направлениям. С одной стороны выполняется «повышение качества» обрабатываемых данных, а с другой – построение эффективной модели обработки [12].

В целях повышения качества обработки данных используются методы формирования пространства признаков. Среди них применительно к рассматриваемой проблематике можно выделить подходы, выполняющие разделение данных, на основе кластеризации, поиска точек разладки временных рядов, обнаружения «дрейфа концепта» при трансформации свойств данных. Проблемные вопросы нахождения точек, где изменяются свойства предикторов и целевой переменной, являются важными для решения многих прикладных задач, использующих сегментированную регрессию. В этих целях применяется большое количество методов, таких как байесовский анализ, метод максимального правдоподобия, квантильная регрессия, непараметрические методы, аналитические способы. Однако сложность аппроксимирующей функции, проходящей через такие точки, часто приводит к большим вычислительным и ресурсным затратам при ее поиске.

Процессы разделения выборки в методах машинного обучения многими исследователями рассматриваются как вспомогательные. В большинстве случаев им уделяют внимание в рамках решения специализированных задач. Они зависят от вида и свойств обрабатываемой информации. Применительно к обработке информационных последовательностей можно выделить ряд основных направлений. Например, в работе [13] используются классические методы кластеризации. Несмотря на их относительную простоту, они позволяют повысить качественные показатели обработки только данных с определенными свойствами. В условиях высокой размерности, когда может существовать более одной точки изменения возникает проблема их обнаружения и локализации. Эта задача может решаться методами сегментации данных, которые используют динамическое программирование [14 – 16], бинарную сегментацию, байесовские методы [17]. В работах [17 – 19] предложены байесовские подходы для регрессий по точкам изменения. Однако описанные в них решения имеют высокую вычислительную сложность, требуют большое количество итераций при моделировании цепей Маркова методом Монте-Карло. Работа [20] использует статистические пороги для сегментации последовательности. При таком решении для повышения точности

обработки необходимо учитывать выбросы и шумы объектов наблюдения. В [21] представлен метод сегментированной регрессии (SEG) с использованием линейного перехода для оценки точек изменения регрессионной модели. В [22] предложены методы глубокого обучения. Качественные показатели обработки данных в этом случае зависят от ресурсоемкости и анализа данных. В работе [23] для поиска сегментов применяется расширенный тест Дики-Фуллера (ADF) на стационарность данных. В [24] рассматривается метод, использующий шаблоны для определения сегментов временного ряда. В [25] предложен подход, базирующийся на поиске максимального правдоподобия, применяющий метод динамического программирования. В [26, 27] был реализован метод одновременной многомасштабной оценки точки изменения (SMUCE), определяющий не только сами точки, но и доверительные интервалы. В [28] для разбиения выборки предложен метод бинарной сегментации.

Характеристики рассмотренных методов, предпосылки применения и возникающие при этом сложности представлены в таблице 1.

Разделение выборки на сегменты позволяет определить внутреннюю структуру данных для дальнейшего анализа и обработки, исследовать вероятные связи между объектами наблюдений. В случае рассмотрения информационных последовательностей возникает необходимость применения алгоритмов разделения данных, обнаруживающих новые закономерности.

Второе направление связано с поиском наиболее эффективной модели обработки данных. В простейших задачах применяются базовые алгоритмы, например линейная, логистическая регрессия, метод машинных векторов, нейросетевые подходы. Достигаемые ими значения показателей качества обработки зависят от свойств обрабатываемых выборок. Наличие выбросов, линейность данных, независимость переменных оказывают существенное влияние на показатели качества результата обработки.

В целях преодоления обозначенных проблемных вопросов для повышения качества обработки регрессионных последовательностей в [29] используется многомодельный подход, направленный на формирование ансамбля моделей и алгоритмов, сочетающий несколько методов машинного обучения. Могут использоваться системы принятия решений, основанные на правилах, априорных знаниях о данных, системы взвешенного, выборочного голосования, которые оценивают модели только на основе их прошлых результатов прогнозирования [30], каскады простых алгоритмов и глубоких нейронных сетей [31].

Таблица 1. Характеристики рассмотренных методов разделения последовательности

Методы разделения	**Ссылка на источник**	**Характеристики**	
		Предпосылки применения	**Проблемы применения**
Классические методы кластеризации	[13]	Имеют относительную простоту реализации.	Обладают сильной зависимостью от свойств данных.
Методы динамического программирования	[14 – 16], [25]	Применяются при не дифференцируемости целевой функции, дискретном изменении переменных.	Имеют большую ресурсоемкость.
Методы бинарной сегментации	[17], [28]	Простота реализации, высокая скорость обработки.	Большая вероятность ошибочной сегментации.
Байесовские методы	[18, 19]	Могут быть реализованы на выборках небольшого объема, используются для данных высокой размерности.	Характеризуются субъективностью предварительного описания, что может приводить к слабым качественным показателям.
Метод на основе статистических порогов для сегментации последовательности	[20]	Обладают высокой чувствительностью, имеют апробированный математический аппарат.	Качество результатов зависит от анализа и интерпретации данных, (необходимо определять и интерпретировать выбросы и шумы объектов наблюдения).
Метод сегментированной регрессии	[21]	Характеризуется простотой вычислительных алгоритмов.	При увеличении разрядности данных падает точность определения точек изменения, чувствителен к изменчивости данных (сдвигу дисперсии).
Методы глубокого обучения	[22]	Имеют возможность достижения высоких качественных показателей при адекватной модели.	Обладают зависимостью от исходных свойств последовательности, сложностью построения модели, ресурсоемкостью, требуют большое количество итераций.
Метод сегментирования с применением расширенного теста Дики-Фуллера (ADF)	[23]	Адаптирован для рядов, обладающих стационарностью.	В случае нестационарного ряда и необходимости проверки нескольких точек разладки существенно возрастает сложность.
Метод шаблонов для определения сегментов временного ряда	[24]	Использует множество метрик сходства и различия.	Зависит от качества шаблонов и выбранных критериев сходства и различия.
Метод одновременной многомасштабной оценки точки изменения (SMUCE)	[26, 27]	Позволяет оценить число скачков, их расположение, а также доверительные интервалы для точек разладки.	Требует решения ресурсоемкой оптимизационной задачи, в которой на интервалах может существовать достаточно большое число локальных решений.

Все эти методы в той или иной степени улучшают отдельные качественные показатели, однако основными их недостатками является сложность обучения, ресурсоемкость и увеличение времени работы алгоритма. Кроме того, очень часто возникают ситуации, когда неправильно подобранные модели и способы агрегации их результатов ухудшают общий прогноз. А в случае трансформации свойств данных без организации постоянных процессов обучения модель с течением времени может потерять адекватность [32, 33].

В связи с этим возникает задача оценки входной выборки, разделения ее на сегменты с использованием информации о свойствах данных и формирование моделей обработки, показывающих лучшие качественные показатели для вычисленных свойств обрабатываемой последовательности.

3. Предлагаемый подход. Адаптивное формирование модели происходит в несколько этапов. В начале осуществляется выбор базовых алгоритмов обработки. На этом шаге, в большинстве случаев, уделяется особое внимание реализации процессов обучения, которые должны обеспечить разнообразие моделей. Это решается различными «манипуляциями» над данными, где происходит формирование признаковых пространств, выборок, и использованием принципиально разных моделей обработки. На следующем шаге выполняется анализ базовых моделей, в результате которого, обычно, возникает необходимость исключения тех из них, которые не позволяют достичь высокую точность прогнозирования. После этого выбирается способ обработки полученных результатов.

Большое количество рутинных повторяющихся операций обуславливает необходимость применения анализа результатов вычислений на отдельных этапах формирования модели обработки. При наличии достаточных вычислительных ресурсов становится возможным проводить автоматическое разделение последовательностей данных различными способами применяя разные алгоритмы, например разбиения последовательности или кластеризации. А затем, на отдельных последовательностях сегментов данных вычислять качественные показатели разных заранее выбранных моделей и назначать на каждый отдельный сегмент "лучшую" модель. На рисунке 1 приведены процессы формирования модели и их взаимосвязи.

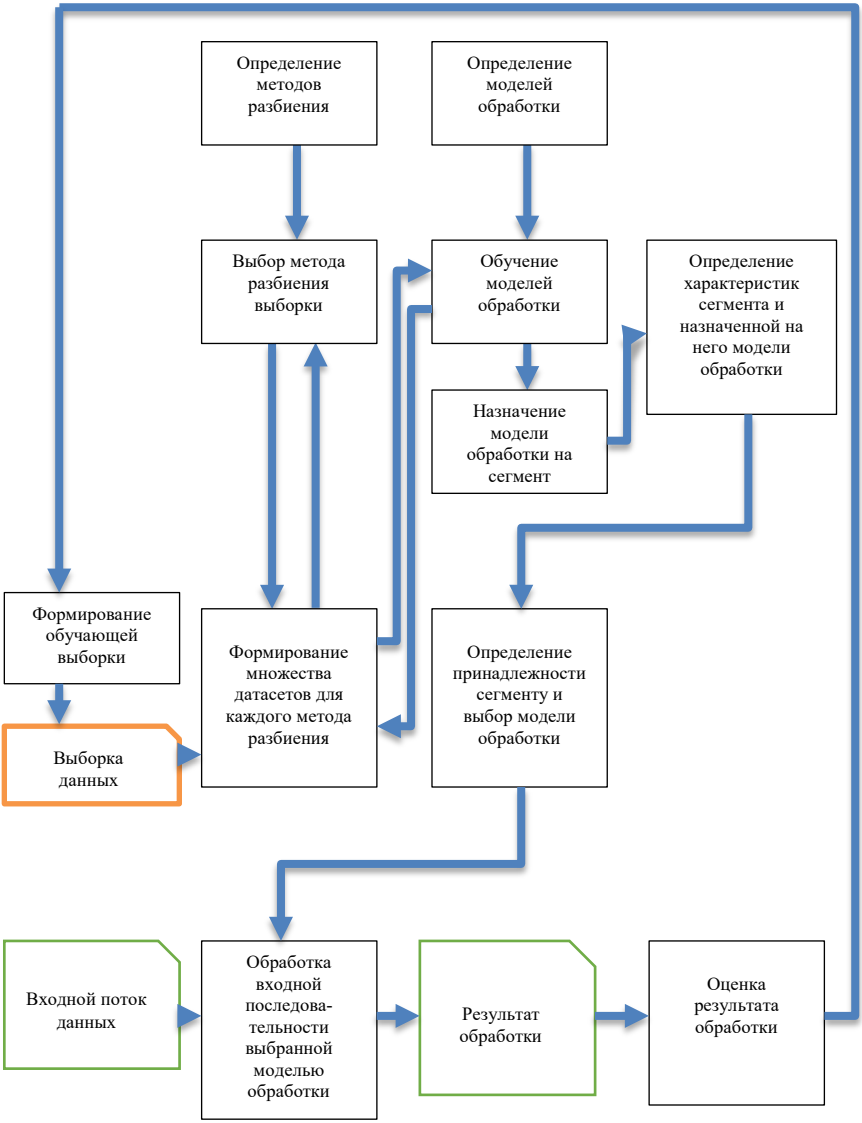

Рис. 1. Процессы формирования модели

Таким образом, в целях улучшения качества обработки в статье предлагается методика, направленная на оценку показателей качества

моделей обработки на отдельных сегментах, полученных разными способами разделения информационной последовательности.

Предполагается, что последовательность данных может быть разделена разными способами. В результате меняется количество объектов наблюдения, состав и свойства сегментов, связанные, например, с распределением данных. А это приводит к тому, что при различных способах разбиения последовательности на каждом сегменте лучшие качественные показатели будут достигать разные модели. В связи с возможной трансформацией свойств данных анализ разделяемых последовательностей необходимо осуществлять постоянно, чтобы оперативно реагировать на возникающие изменения и настраивать модель обработки.

4. Формальная постановка задачи. Имеется выборка последовательности объектов наблюдения X, определены модели обработки $\{a_1,...,a_N\} \in A$ и методы сегментации данных $\{\mu_l,...,\mu_L\} \in \Omega$. Информационная последовательность X разбивается на отдельные сегменты. В результате получается множество способов разбиения X^{μ}, сегменты которых обрабатываются моделями А.

Необходимо найти метод μ* разделения последовательности на сегменты $X^{\mu^*} = \{X_{1^{\mu^*}}^{\mu^*},...,X_{m^{\mu^*}}^{\mu^*}\}$ и назначить на каждый сегмент модель обработки $a_i \in A$, имеющую лучшее значение функционала качества $Q(a_i(x), X_{j^{\mu^*}}^{\mu^*}) \to \max_{a_i \in A, \mu^* \in \Omega}$.

Таким образом, основной акцент делается на назначение лучших по качественным показателям моделей обработки сегментов, что в определенных случаях позволяет обращать меньше внимания на реализацию вычислительно сложных процедур сегментации данных регрессии, учитывающих свойства объектов наблюдения в последовательностях данных. Возникает задача разработки методики выбора сегментов регрессионной последовательности, полученных разными способами и алгоритмами, отличающейся использованием функционала качества моделей обработки на подпоследовательности, что дает возможность формировать агрегационную модель, осуществляющую назначение лучших по качественным показателям моделей на сегменты.

В отличие от классических подходов регрессионная последовательность данных сначала разделяется выбранными методами сегментации, а затем на сегментах обучаются модели и определяются их характеристики. Базовыми алгоритмами обработки для анализа последовательностей данных могут быть линейная, логистическая регрессии, метод опорных векторов, нейросетевые и другие модели.

Выбор методов сегментации, например алгоритмов кластеризации или поиска точек изменений, определяется ресурсоемкостью и вычислительной сложностью. Предпосылкой предлагаемого метода с одной стороны является предположение о неоднородности данных, а с другой возникающая неоднозначность, например, в условиях высокой размерности, когда может существовать более одной точки в области локализации. Объекты наблюдения регрессионной последовательности могут образовывать разные области, где меняются тренды и свойства данных. В результате получаются сегменты, где может изменяться направление тренда, или в одних случаях будет линейная зависимость, в других нелинейная. Это приводит к тому, разные подпоследовательности внутри X лучше аппроксимируются различными функциями.

Рассматривая способы разбиения методами $\mu_l \in \Omega$ последовательностей данных $\{X^{\mu_l}_{1_{\mu_l}},...,X^{\mu_l}_{m_{\mu_l}}\} \in X^{\mu_l}$, получая различные комбинации сегментов данных, необходимо выбрать метод разбиения μ^* и определить агрегируемую функцию a^* обработки данных, состоящую из моделей, при которых на сегментах достигаются лучшие показатели качества

$$a^*, \mu^* = \arg\max_{\mu_l \in \Omega, a_i \in A} Q(a_i(x, X^{\mu_l})). \quad (1)$$

Определяя метод разбиения выборки и назначая на каждый сегмент алгоритм, имеющий по результатам обучения самые высокие показатели на сегменте по сравнению с другими выбранными, получается агрегированная модель обработки последовательностей

$$a^*(x) = \left\{ \begin{matrix} a_{i_{\mu_l}}(x), x \in X^{\mu_l}_{1_{\mu_l}} \\ a_{M_{\mu_l}}(x), x \in X^{\mu_l}_{M_{\mu_l}} \end{matrix} \right\}. \quad (2)$$

Таким образом, решение направлено на выбор лучшего из заранее определенных методов сегментации и количества сегментов. Это дает возможность сформировать агрегированную модель обработки, где на каждый сегмент назначается свой алгоритм, показывающий в процессе обучения лучший результат на выделенном сегменте.

5. Реализация метода. Применение предлагаемой методики предполагает ряд операций, необходимых для настройки моделей и выполнения ими предопределенных задач. В качестве подготовительного

этапа необходимо определить методы разбиения $\{\mu_l,...,\mu_L\} \in \Omega$ последовательности объектов наблюдения и выбрать модели обработки данных $\{a_1,...,a_N\} \in A$. Для верификации методов разбиения и настройки моделей обработки заранее создается начальная тестовая выборка данных, повторяющая свойства генеральной совокупности. В листинге 1 представлен псевдокод модифицированного алгоритма [33, 34], который реализует формирование модели.

Входные данные:

X – начальная тестовая выборка,

μ_l – методы разбиения выборки X в количестве L,

a_i – модели обработки данных, в количестве N,

$Q(a(x), X)$ – функционал качества,

M – максимальное количество сегментов.

Выходные данные:

μ^* – метод разбиения выборки, где достигается максимальный показатель качества

m^{μ^*} – количество сегментов, полученных выбранным μ^* методом разбиения выборки

$X^{\mu^*}_{1^{\mu^*}},...,X^{\mu^*}_{m^{\mu^*}}$ – сегменты выборки X, обработанной выбранным методом разбиения

$a^{\mu^*}(x, X^{\mu^*})$ – модель обработки выборки X, сегментированной методом μ^*

Начало

1. Цикл перебора методов разбиения выборки: $for\ l = 1,...,L$

2. Цикл, увеличивающий количество сегментов: $for\ m = 1,...,M$

3. Обработка методом разбиения μ_l выборки X, формирование сегментов $\{X^{\mu_l}_{1_{\mu_l}},...,X^{\mu_l}_{j_{\mu_l}},...,X^{\mu_l}_{m_{\mu_l}}\} \in X^{\mu_l}$ для метода разбиения μ_l и количества сегментов m

4. Цикл перебора количества сегментов: $for\ j = 1,...,m$

5. Цикл перебора моделей обработки a_i: $for\ i = 1,...,N$

6. Обучение модели a_i на сегменте $X^{\mu_l}_{j_{\mu_l}}$

7. Конец цикла п. 5

8. Определение лучшей из моделей $\{a_1,...,a_N\} \in A$ на сегменте $X^{\mu_l}_{j_{\mu_l}}$ по значению показателя качества модели $a^{j_{\mu_l}} = \arg\max_{a_i \in A} Q(a_i(x, X^{\mu_l}_{j_{\mu_l}}))$

9. Конец цикла п. 4

10. Определение выборки $\{X^{\mu_l}_m = \{X^{\mu_l}_{1_{\mu_l}},...,X^{\mu_l}_{m_{\mu_l}}\}$ для выбранного метода и выбранного количества сегментов

11. Формирование агрегационной модели из моделей, $\{a^{1_{\mu_l}},...,a^{m_{\mu_l}}\} \in A$, показывающих лучшие результаты по значению показателя качества, на m сегментов на выборке $X_m^{\mu l}$ после обработки методом μ_l

$$a_m^{\mu_l}(x, X_m^{\mu_l}) = \begin{Bmatrix} a^{1_{\mu_l}}(x, X_{1_{\mu_l}}^{\mu_l}), x \in X_{1_{\mu_l}}^{\mu_l} \\ ... \\ a^{m_{\mu_l}}(x, X_{m_{\mu_l}}^{\mu_l} l), x \in X_{m_{\mu_l}}^{\mu_l} \end{Bmatrix}$$

12. Конец цикла п. 2

13. Определение количества сегментов разбиения выборки X методом μ_l, при котором был достигнут лучший показатель качества $m^{\mu_l} = \arg\max_{m\in\{1,...,M\}} Q(a_m^{\mu_l}(x, X_m^{\mu_l}))$

Определение сегментов выборки $X_{m^{\mu_l}}^{\mu l} = \{X_{1^{\mu_l}}^{\mu l},...,X_{m^{\mu_l}}^{\mu l}\}$

Определение модели, которая достигает лучшего показателя качества при разбиении методом μ_l

$$a_{m^{\mu_l}}^{\mu_l}(x, X_{m^{\mu_l}}^{\mu_l}) = \arg\max_{a_m^{\mu_l}\in A} Q(a_m^{\mu_l}(x, X_m^{\mu_l}))$$

14. Конец цикла п. 1

15. Окончательный выбор метода разбиения выборки, где достигается максимальный показатель качества

$$\mu^* = \arg\max_{\mu_l\in\mu} Q(a_{m^{\mu l}}^{\mu l}(x, X_{m^{\mu l}}^{\mu l}))$$

Определение количества сегментов, полученных выбранным методом разбиения выборки

$$m^{\mu^*} = \arg\max_{m\in\{1,...,M\}} Q(a_{m^{\mu l}}^{\mu l}(x, X_{m^{\mu l}}^{\mu l}))$$

Определение сегментов выборки, обработанной выбранным методом разбиения

$$X^{\mu^*} = \{X_{1^{\mu^*}}^{\mu^*},...,X_{m^{\mu^*}}^{\mu^*}\}$$

Окончательное формирование модели обработки

$$a^{\mu^*}(x, X^{\mu^*}) = \begin{Bmatrix} a_{1^{\mu^*}}^{\mu^*}(x, X_{1^{\mu^*}}^{\mu^*}), x \in X_{1^{\mu^*}}^{\mu^*} \\ ... \\ a_{m^{\mu^*}}^{\mu^*}(x, X_{m^{\mu^*}}^{\mu^*}), x \in X_{m^{\mu^*}}^{\mu^*} \end{Bmatrix}$$

Конец

Листинг 1. Псевдокод алгоритма формирования модели обработки

В представленном алгоритме максимальное число рассматриваемых сегментов из одной последовательности равно M^2. Условно накладные расходы на усредненное время обучения и обработки для модели a_i на сегменте можно оценить величиной p, тогда общая сложность алгоритма составляет O(pLNM2). Количество методов разбиения выборки L и количество моделей обработки N обычно относительно небольшие величины. На рост сложности влияет число рассматриваемых сегментов M, что является существенным ограничением предлагаемого метода.

Однако в целях оптимизации и ускорения алгоритмов обработки временных рядов и информационных последовательностей возможно применение ряда подходов, направленных на анализ и объединения «сходных» по свойствам данных сегментов, отбрасывания способов разбиения выборки при качественных показателях обучения моделей меньше заданного порога, применение распараллеливания процессов обучения для моделей.

Реализация модели предполагает выполнение ряда действий по формированию информационной последовательности и определению методов обработки и сегментации. Каждому методу разбиения задаются параметры, например количество разбиений, функция расстояния для кластеризации, функция определения изменения тренда для методов поиска точки разладки. В целях упрощения можно считать, что методы с разными параметрами являются различными методами разбиения. Затем каждому методу подается на вход выборка данных X и формируются сегменты $X^{\mu l}_{1_{\mu l}},...,X^{\mu l}_{M_{\mu l}}$. В соответствие методу разбиения $\mu_l \in \Omega$ ставится $X^{\mu l}_{1_{\mu l}},...,X^{\mu l}_{M_{\mu l}} \in X^{\mu l}$ множество сегментов выборки.

На объектах наблюдения x каждого из множества сегментов $X^{\mu l}_{1_{\mu l}},...,X^{\mu l}_{M_{\mu l}}$, образованных методом разбиения $\mu_l \in \Omega$, осуществляется обучение всех заранее определенных моделей $\{a_1,...,a_N\} \in A$ обработки последовательности.

Далее происходит регуляция системы, где для каждого метода разбиения $\mu_l \in \Omega$ для всех обученных моделей $\{a_1,...,a_N\} \in A$ на полученных с его применением сегментах $X^{\mu l}_{1_{\mu l}},...,X^{\mu l}_{M_{\mu l}}$ сравниваются значения функционала качества $Q(a_i(x, X^{\mu_i}_{j_{\mu_i}}))$. Каждому методу разбиения $\mu_l \in \Omega$ на каждом сегменте $X^{\mu l}_{j_{\mu l}}$ ставятся в соответствие

модели, имеющие лучшие показатели качества. Применяя выражение (1) становится возможным вычислить выбрать метод разбиения $\mu_l \in \Omega$ при котором лучшие модели показывают лучшие результаты на сегментах. Итоговая модель определяется выражением (2). Формируются кортежи $< a_{j_{\mu l}}, X^{\mu l}_{j_{\mu l}} >$, где на сегмент назначается модель.

В дальнейшем при обработке данных сегменту ставится в соответствие информация о его свойствах. В зависимости от решаемой задачи в качестве свойств могут использоваться статистические характеристики среднего значения данных сегмента, направление тренда, скорость изменения значений и т.д. Совершенствование предложенной модели возможно с использованием методов постоянного обучения, по аналогии с нейросетевыми решениями [35]. Для этого может быть применена многоуровневая обработка [32, 34, 36], где модели разделяются на уровни. При ее реализации в начале выполняется настройка моделей на обучающей выборке, решающих задачи обработки информации. Производится выбор оптимального метода разбиения $\mu_l \in \Omega$ и количества сегментов. Выполняется первоначальное обучение. Вычисляются свойства сегментов. Поступающий на вход информационный поток подвергается обработке. Производится анализ и вычисление свойств объектов наблюдения. Они соотносятся со свойствами сегментов, и для их обработки выбирается заранее определенная модель, имеющая лучшие значения функционала качества на схожем по свойствам сегменте обучающей выборки. Предсказанные моделями результаты сравниваются с реальными значениями объектов наблюдения, полученными от регистрирующих систем и устройств. В случае увеличения ошибок выше заранее определенного порога принимается решение о формировании выборки данных. Над выборкой проводятся манипуляции, указанные в алгоритме (листинг 1), происходит настройка и обучение модели. После этого анализируемый поток данных поступает на вход обученной модели.

Таким образом, возможно применение методов самообучения для рассматриваемого решения. Однако эти действия необходимо выполнять в параллельном режиме, чтобы избежать существенного роста вычислительной сложности.

6. Эксперимент. Цель проведения эксперимента состояла в оценке повышения качественных показателей обработки информационных последовательностей при применении рассматриваемой методики. Так как предлагаемая методика основана на сегментировании выборки данных, то были рассмотрены два основных подхода к разделению. В первом случае применялся контролируемый подход, когда

осуществлялось нахождение границ сегмента на основе вычисления смены тренда информационной последовательности. Во втором случае был применен подход, разделяющий последовательность на равные по количеству наблюдений сегменты. Оценка предлагаемого метода осуществлялась для задачи прогнозирования одномерной и многомерной последовательности данных.

В предлагаемом методе в качестве базовых моделей в зависимости от задачи обработки можно использовать модели любой сложности. В эксперименте при выборе моделей обработки приоритет отдавался имеющим высокую скорость обучения. Сравнивались результаты линейной регрессии (LR), квадратичной регрессии (QR), машины опорных векторов (SVM), регрессии гауссова процесса (GR), деревьев решений (DT). Сравнение качественных показателей осуществлялось для различных наборов данных. В одном случае сегментирование выборки осуществлялось простым разделением на равное количество объектов наблюдения. В другом – подвергалась обработке алгоритмом оптимального обнаружения точек изменения с линейными вычислительными затратами, реализованного в среде Matlab [37]. Информационная последовательность объектов наблюдения разделялась в первом случае на m равных частей, а во втором – на m сегментов с использованием информации о точках изменений. Затем каждый сегмент аппроксимировался моделями регрессии. Поступающий на вход модели объект наблюдения входной последовательности анализировался на принадлежность сегменту. Вычислялись его значения $\widehat{y_i}$ полученные моделью, назначенной на сегмент, и сравнивалось с истинным значением y_i. Показатели качества обработки данных определялись метриками MSE и MAE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \tag{4}$$

где $\widehat{y_i}$ – предсказанное значение, y_i – реальное значение, n – количество объектов наблюдения.

6.1. Эксперимент на модельных данных. Для иллюстрации предлагаемой методики был проведен эксперимент на модельных данных. Рассматриваемая информационная последовательность подвергалась сегментации. Модельные данные эксперимента для одномерной регрессии [38] приведены на рисунке 2.

В верхней части рисунка приведено разбиение на 4 сегмента равных по количеству объектов наблюдения, в нижней части – алгоритмом оптимального обнаружения точек изменения с линейными вычислительными затратами с 4 сегментами.

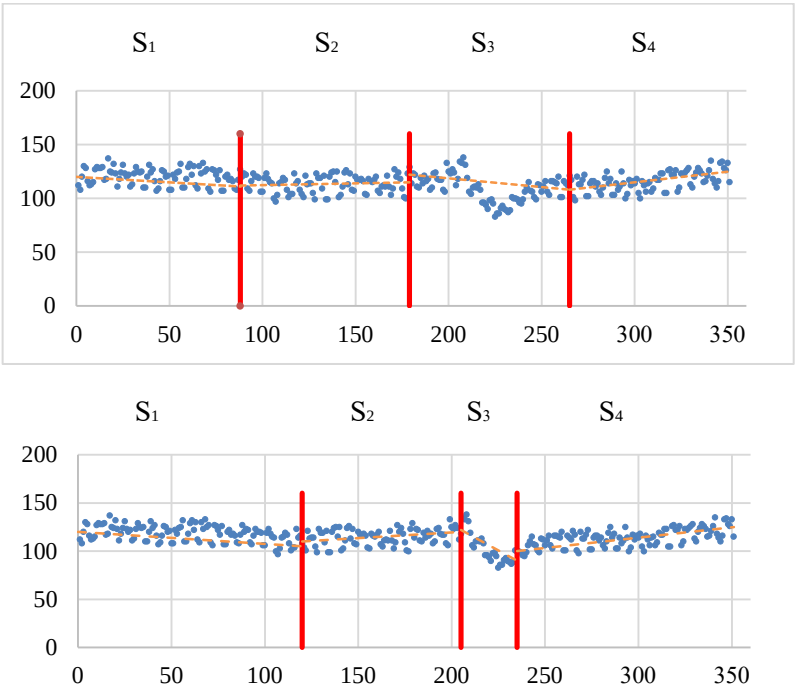

Рис. 2. Деление последовательности на 4 сегмента (на равные части – верхняя часть, алгоритмом оптимального обнаружения точек изменения с линейными вычислительными затратами – нижняя часть)

На рисунке 3 представлены гистограммы функций потерь. На них приведены значения MSE и MAE на каждом из 4 сегментов рисунка 2, полученным разными способами и аппроксимированными линейными и квадратичными функциями.

Для анализируемых данных вычисленные значения выражений (3) и (4) показывают преимущество использования методов разделения последовательности. Значения функций потерь MSE и MAE оказались для данных после сегментирования ниже, чем для всей выборки целиком. Разбиение алгоритмом оптимального обнаружения точек изменения с линейными вычислительными затратами для рассматриваемого случая

оказалось предпочтительнее, чем при разделении на равных по количеству объектов наблюдения сегменты.

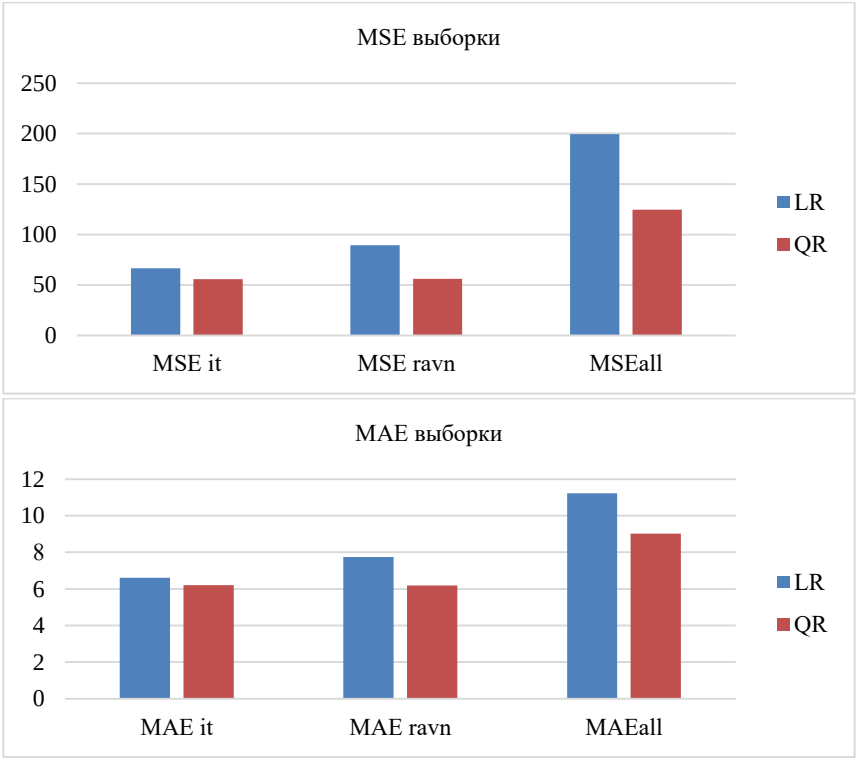

Рис. 3. Функция потерь MSE и MAE при апроксмации LR и QR на всей выборке и при делении на 4 сегмента равным количеством объектов наблюдения (ravn) и алгоритмом оптимального обнаружения точек изменения с линейными вычислительными затратами (it)

Полученные сегменты могут быть неоднородны. На одних сегментах могут лучше показывать себя одни алгоритмы, на других – другие. Разделив последовательность, например, на 4 сегмента можно увидеть, что на различных сегментах значения функции потерь будут отличаться. На рисунке 4 приведены значения MSE и MAE аппроксимирующих функций на каждом из 4 сегментов, полученных алгоритмом оптимального обнаружения точек изменения с линейными вычислительными затратами и простым разделением на равные отрезки.

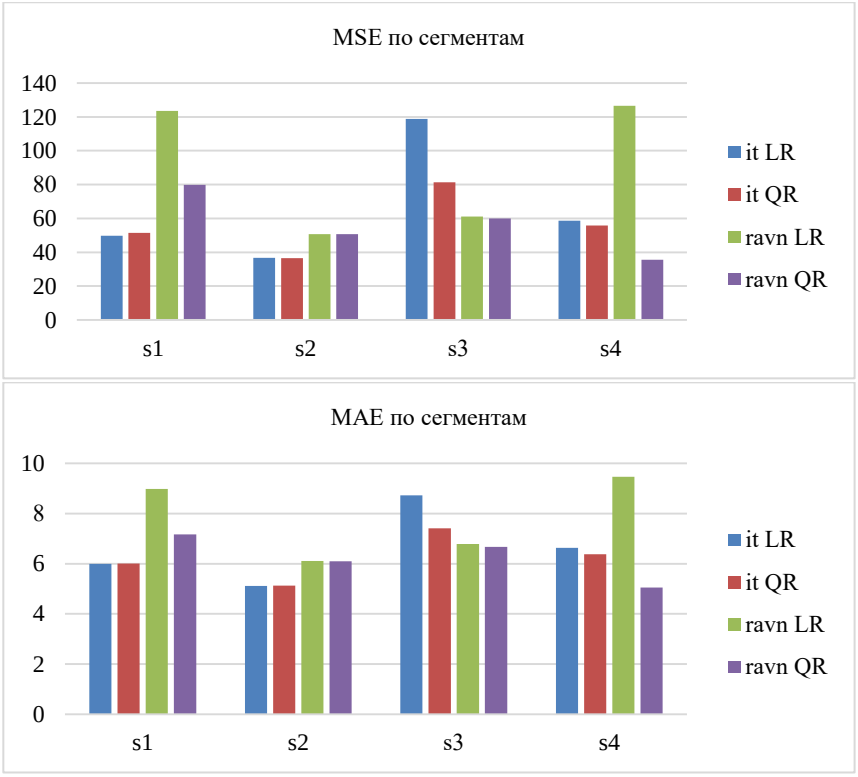

Рис. 4. Значения MSE и MAE алгоритмов LR, QR для сегментов, полученных равным количеством объектов наблюдения (ravn) и алгоритмом оптимального обнаружения точек изменения с линейными вычислительными затратами (it)

Анализ гистограмм рисунка 4 показывает, что при разбиении алгоритмом оптимального обнаружения точек изменения с линейными вычислительными затратами на сегментах s1, s2 меньшие значения функции потерь показывает линейная регрессия, а на сегментах s3, s4 – квадратичная регрессия. При простом разделении линейная регрессия показывает лучшие результаты на s2, а на остальных сегментах – квадратичная регрессия. При определении значений функционала качества для данных внутри сегментов возможно решение задачи по назначению модели обработки на сегмент, где он показывает лучшие значения.

Далее в эксперименте производилась оценка влияния количества сегментов m на значения MSE и MAE. Число сегментов увеличивалось

(от 1 (вся выборка) до 35) и оценивались значения функций потерь MSE и MAE для аппроксимирующих функций и способов разбиения.

На рисунке 5 приведена зависимость усредненных значений MSE и MAE линейной и квадратичной регрессии от количества сегментов, полученных при делении последовательности на равные части и при использовании алгоритма оптимального обнаружения точек изменения с линейными вычислительными затратами.

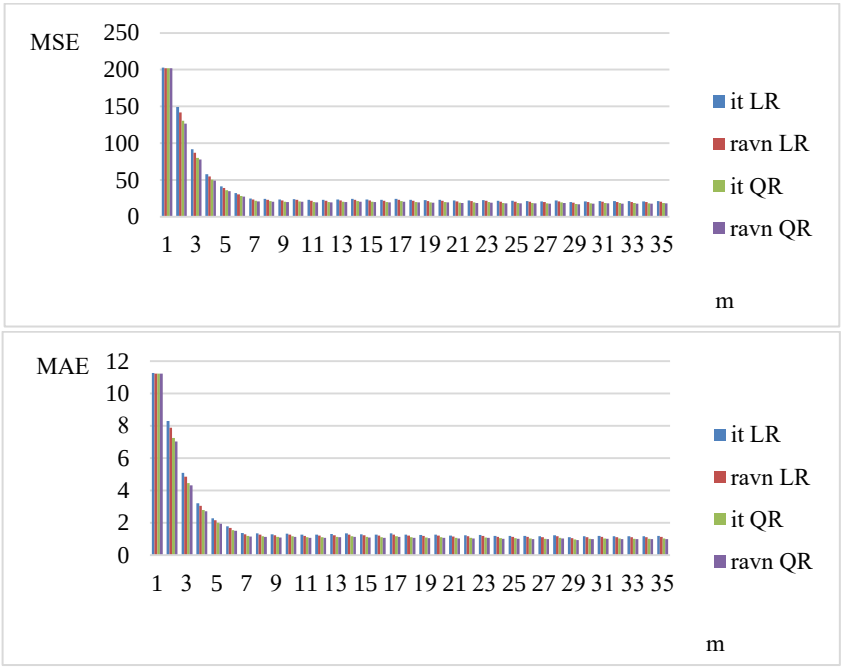

Рис. 5. Зависимость значений MSE и MAE от количества сегментов m для линейной регрессии при разделении на равные части по количеству объектов (ravn) и алгоритмом оптимального обнаружения точек изменения (it) с линейными вычислительными затратами

Рисунок 5 показывает, что на достигаемые значения MSE и MAE, оказывает влияние количество сегментов. Значения функции потерь для рассматриваемых моделей уменьшаются при увеличении количества сегментов. Кроме того, на графиках видно, что до определенного момента, чем больше сегментов создается, тем меньше становятся MSE и MAE для выбранных аппроксимирующих функций и способов сегментации. Затем показатели функций потерь моделей выходят на

плато и дальнейшее увеличение количества сегментов к существенному улучшению функционала качества не приводят.

В то же время, необходимо отметить, что при неограниченном разрастании количества сегментов может возникнуть ситуация, когда будут появляться «микросегменты» содержащие, недостаточное количество объектов наблюдения для построения адекватной модели. Это требует дополнительного анализа и обработки.

6.2. Применение методики при появлении точек разрыва. При анализе последовательности данных могут возникать ситуации, где появляются точки разрыва. При прохождении такой точки может меняться вид регрессионного уравнения.

Несмотря на применение алгоритмов IJD, SEG [28, 37, 39] и ряда других, использующих средние значения, дисперсии данных, изменение наклона линейной регрессии, их результаты зависят от последовательностей данных и могут приводить к существенным отличиям в определении границ и составе объектов наблюдения в сегментах.

На рисунке 6 приведены модельные данные с разрывом.

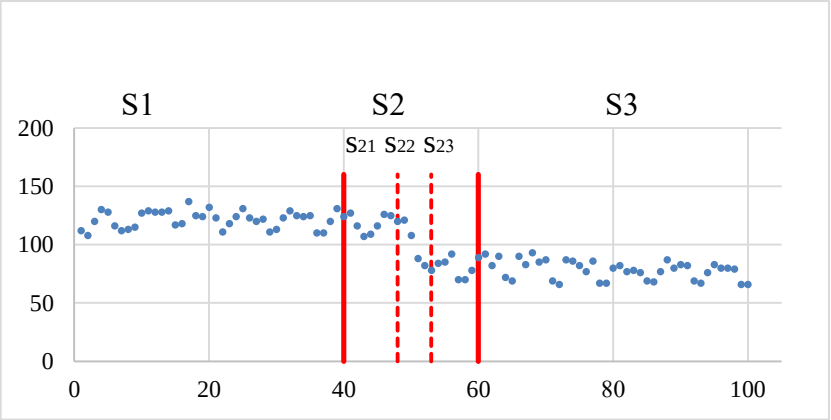

Рис. 6. Локализация окрестностей точек потенциального разрыва дальнейшем делением сегментов

Для обработки подобных последовательностей необходима либо сложная функция аппроксимации, либо определение точки потенциального разрыва как границы сегмента. Повышение сложности функции аппроксимации приводит к увеличению ресурсоемкости, вычислительной сложности, одновременно с этим может возникнуть эффект переобучения модели, а это не гарантирует повышение качества результата обработки.

Определение точки разрыва, как границы сегмента имеет свои сложности, возникающие из-за нечетких окрестностей в неоднозначных случаях. Например, при переходе системы из одного режима работы в другой наблюдаются пограничные ситуации, когда полученные значения не позволяют однозначно трактовать состояние. В этом случае на основе предлагаемой методики возможно зафиксировать модель, получить сегменты, определить показатели качества и в случае, если они оказываются хуже заданного уровня, то произвести дальнейшее деление.

На рисунке 7 показано, что для рассматриваемых данных дальнейшее деление сегмента S2 дает возможность уменьшить значения функции потерь MAE и MSE для линейной и квадратичной регрессии как в случае применения анализа точек изменения, так и в случае простого деления сегмента.

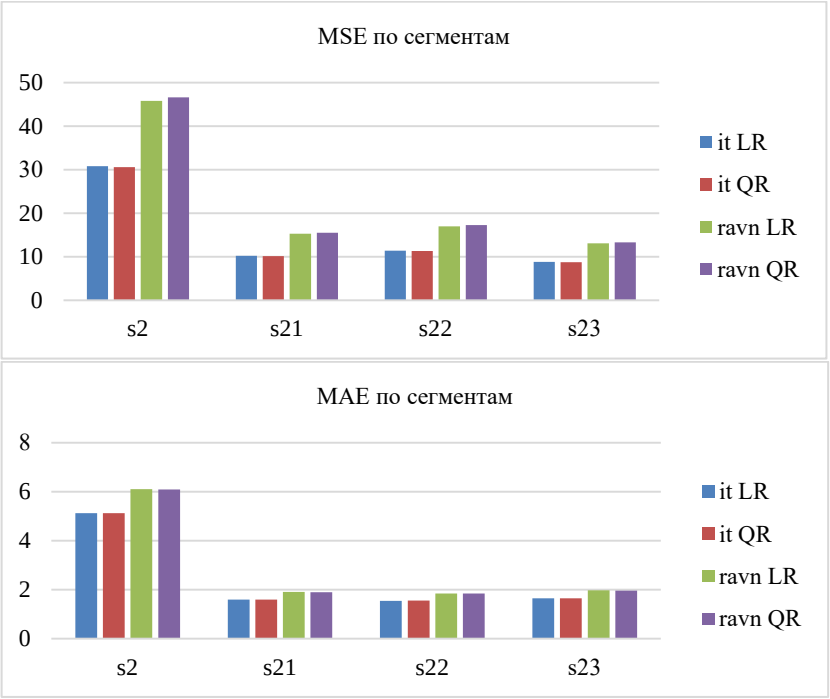

Рис. 7. Значения функции потерь при делении сегмента

Таким образом, для заданной модели можно уменьшать размер сегмента, анализируя выбранный показатель качества. Если значение оказываются хуже заданного уровня, то производить дальнейшее

деление. Затем, определяя значения функционала качества, осуществлять выбор моделей. При этом необходимо также учитывать их ресурсоемкость и сложность процессов обучения.

6.3. Эксперимент на множественной регрессии. В условиях высокой размерности, когда может существовать более одной точки изменения, возникает задача их обнаружения и локализации.

В большинстве практических случаев анализируемые данные являются многомерными, что накладывает определенные сложности их предварительной обработки, например при обнаружении выбросов и шумов.

Возникают определенные проблемы, связанные с оптимальным определением точки изменения в регрессионной модели.

Для оценки достигаемых качественных показателей при сегментировании многомерной последовательности данных была применена множественная регрессия. В качестве исследуемого набора данных была использована выборка [38], которая обрабатывалась моделями линейной регрессии LR, регрессия гауссова процесса GR, машина опорных векторов SVM, деревья решений DT.

На рисунке 8 представлены графики изменений MSE(m) и MAE(m) от количества сегментов m для базовых алгоритмов LR, GR, SVM, DT. На графиках прослеживается, что для большинства относительно простых моделей, например LR, SVM, GR к повышению показателей качества обработки может приводить уменьшение сегмента без учета свойств содержащихся в нем объектов наблюдения.

Однако предлагаемая методика, использующая уменьшение размера сегмента, позволяет получить эффект не для всех моделей. Например, для рассматриваемой выборки данных уменьшение размера сегмента оказывает более слабое влияние на алгоритм DT по сравнению с другими. Значение показателей MSE и MAE для него понижается незначительно.

Несмотря на сравнительную простоту реализации предложенного метода имеется ряд ограничений, который необходимо учитывать при формировании моделей обработки. Информационные последовательности могут иметь различные свойства в разных сегментах, что оказывает влияние на выбор наиболее эффективных моделей их обработки.

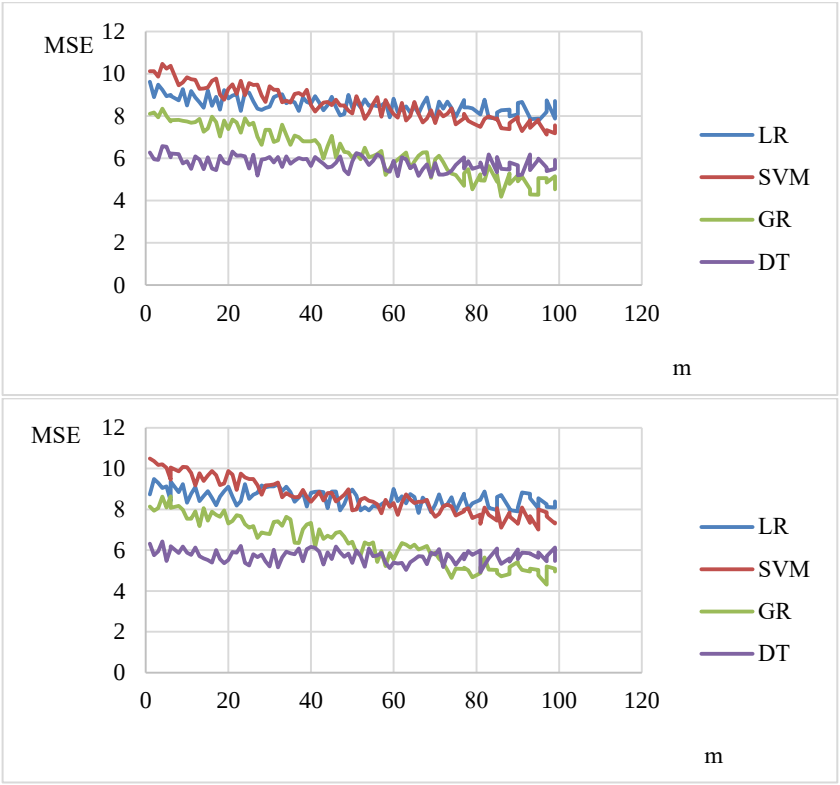

Рис. 8. Зависимость значений MSE различных алгоритмов от количества сегментов m для множественной регрессии при сегментировании делением на равные отрезки (вверху) и алгоритмом оптимального обнаружения точек изменения с линейными вычислительными затратами (внизу)

6.4. Эксперимент на данных датасетов. Далее эксперимент был посвящен анализу ряда различных наборов данных множественной регрессии. В качестве экспериментальных данных рассматривались выборки [38 – 40]. Объем первого набора около 30000 записей, 3 предиктора и 1 целевая переменная, второго – 32000 записей, 123 предиктора и 1 целевая переменная, третьего – 32000 записей, 8 предикторов и 1 целевая переменная.

Данные представлялись временными рядами.

В качестве исследуемых определены модели обработки данных и составленные из них ансамбли LR, SVM, GR, DT, LR+SVM, SVM+GR, не требующие больших вычислительных затрат. Производилась разбиение

выборки на 15 и 25 сегментов алгоритмом оптимального обнаружения точек изменения с линейными вычислительными затратами и разделением на равные по количеству объектов наблюдения, определялся функционал качества каждой модели. В качестве анализируемых показателей были выбраны MSE и MAE.

Результаты качественных показателей, полученных при предсказании для сегментов и всей выборки целиком, представлены в таблицах 2-4.

Таблица 2. Обработка выборки [38]

	Модель	Вся выборка		15 сегментов		25 сегментов	
		MSE	MAE	MSE	MAE	MSE	MAE
Алгоритм обнаружения точек изменения	LR	137,08	9,25	126,33	8,88	**116,43**	**8,52**
	SVM	218,13	11,67	201,03	11,20	**185,27**	**10,75**
	GR	80,51	7,09	74,20	6,81	**68,38**	**6,53**
	DT	46,35	5,38	42,71	5,16	**39,37**	**4,96**
	LR+SVM	103,29	8,03	95,19	7,71	**87,73**	**7,40**
	SVM+GR	64,02	6,32	59,00	6,07	**54,37**	**5,83**
Разделение на равные сегменты	LR	165,87	10,17	152,86	9,77	**140,88**	**9,38**
	SVM	263,94	12,83	243,25	12,32	**224,18**	**11,83**
	GR	97,42	7,80	89,78	7,49	**82,74**	**7,19**
	DT	56,08	5,92	51,68	5,68	**47,63**	**5,45**
	LR+SVM	124,98	8,83	115,18	8,48	**106,15**	**8,14**
	SVM+GR	77,46	6,95	71,39	6,67	**65,79**	**6,41**

Таблица 3. Обработка выборки [41]

	Модель	Вся выборка		15 сегментов		25 сегментов	
		MSE	MAE	MSE	MAE	MSE	MAE
Алгоритм обнаружения точек изменения	LR	0,0025	0,0300	0,0024	0,0291	**0,0022**	**0,0282**
	SVM	0,0049	0,0553	0,0046	0,0536	**0,0043**	**0,0520**
	GR	0,0016	0,0316	0,0015	0,0307	**0,0014**	**0,0297**
	DT	0,0009	0,0237	**0,0008**	0,0230	**0,0008**	**0,0223**
	LR+SVM	0,0036	0,0474	0,0034	0,0460	**0,0032**	**0,0446**
	SVM+GR	0,0016	0,0316	0,0015	0,0307	**0,0014**	**0,0297**
Разделение на равные сегменты	LR	0,0036	0,0360	0,0034	0,0349	**0,0032**	**0,0339**
	SVM	0,0071	0,0664	0,0066	0,0644	**0,0062**	**0,0624**
	GR	0,0023	0,0379	0,0022	0,0368	**0,0020**	**0,0357**
	DT	0,0013	0,0284	0,0012	0,0276	**0,0011**	**0,0268**
	LR+SVM	0,0052	0,0569	0,0049	0,0552	**0,0046**	**0,0535**
	SVM+GR	0,0023	0,0379	0,0022	0,0368	**0,0020**	**0,0357**

Таблица 4. Обработка выборки [42]

	Модель	Вся выборка		15 сегментов		25 сегментов	
		MSE	MAE	MSE	MAE	MSE	MAE
Алгоритм обнаружения точек изменения	LR	6,9169	0,7100	6,5081	0,6887	**6,1235**	**0,6680**
	SVM	9,7344	0,8400	9,1591	0,8148	**8,6178**	**0,7904**
	GR	4,7089	0,5700	4,4306	0,5529	**4,1688**	**0,5363**
	DT	3,7249	0,4500	3,5048	0,4365	**3,2976**	**0,4234**
	LR+SVM	7,9524	0,6300	7,4824	0,6111	**7,0402**	**0,5928**
	SVM+GR	5,1076	0,5500	4,8057	0,5335	**4,5217**	**0,5175**
Разделение на равные сегменты	LR	11,6896	0,9230	10,9987	0,8953	**10,3487**	**0,8685**
	SVM	16,4511	1,0920	15,4789	1,0592	**14,5641**	**1,0275**
	GR	7,9580	0,7410	7,4877	0,7188	**7,0452**	**0,6972**
	DT	6,2951	0,5850	5,9230	0,5675	**5,5730**	**0,5504**
	LR+SVM	13,4396	0,8190	12,6453	0,7944	**11,8979**	**0,7706**
	SVM+GR	8,6318	0,7150	8,1217	0,6936	**7,6417**	**0,6727**

Результаты в таблицах 2-4 показывают, что уменьшение «размера» сегмента последовательности, в основном, уменьшают функции потерь MSE и MAE от 2 до 5%. Предложенный метод, во-первых, облегчает сложность процессов анализа. Во-вторых, он может приводить к тому, что при обработке появляется возможность построения более простой аппроксимирующей функции. В-третьих, применение сегментации дает возможность локализовывать группы объектов наблюдения и нивелировать влияние различных эффектов, связанных с шумовыми данными и выбросами. Однако для применения предлагаемого решения, необходимо анализировать, чтобы данные внутри сегмента повторяли свойства генеральной совокупности, в противном случае могут возникнуть проблемы репрезентативности обрабатываемой выборки и модель не будет адекватна.

Полученные результаты показывают, что оценка потенциальных качественных показателей при обучении моделей на сегментах дает возможность определить для каждого сегмента алгоритм обработки, обладающий лучшими показателями функционала качества.

Сегментация выборок данных во многих случаях создает предпосылки для повышения качественных показателей обработки. Получаемые различными методами сегменты характеризуются разными свойствами. В связи с чем при их обработке результаты алгоритмов могут существенно отличаться. Формирование сегмента информационной последовательности данных обычно приводит к возможности построению более простой аппроксимирующей функции, уменьшению

вычислительной сложности. Однако может вызывать проблемы, связанные с необходимостью более детального анализа данных на предмет выбросов и шумов, которые на практике могут быть сложны в интерпретации.

Уменьшение размера сегмента посредством увеличения их количества целесообразно до определенного предела. После его достижения существенного прироста качественных показателей моделей обработки данных не происходит. Назначение алгоритмов с лучшими качественными показателями на сегменты позволяет повысить качественные показатели обработки выборки.

Представленная методика направлена на совершенствование моделей обработки данных. При ее реализации возможно параллельное функционирование алгоритмов.

7. Заключение. Основное преимущество предлагаемой методики построения регрессионных моделей на основе анализа функционала качества обработки сегментов последовательности – отсутствие необходимости использовать сложные методы обнаружения точек изменения.

Причем, когда размерность данных относительно небольшая точки изменения свойств интуитивно понятны, но с ростом размерности данных интерпретировать такие точки становится затруднительно. Кроме того, каждый применяемый метод будет давать свой результат разделения данных, не совпадающий с другими методами, что может сказываться на результате обработки данных.

В информационных последовательностях могут присутствовать выбросы, однако определение объектов наблюдения, подлежащих исключению, является сложной задачей. Определить ситуацию, связанную с выбросом или закономерным появлением объекта наблюдения не всегда представляется возможны, что влияет на результат обнаружения границ сегментов.

В предложенной методике предлагается осуществлять деление выборки данных на сегменты, а далее "подбирать" на них аппроксимирующие модели. Критерием выбора модели является заранее определенный показатель качества, который сравнивается с достигаемыми значениями других моделей, что позволяет выбрать на отрезке лучшую модель.

Предлагаемая методика является относительно простой, позволяет быстро разделять последовательности для различных видов данных регрессии. Эксперименты показывают, что, сегментируя выборку данных и применяя различные модели аппроксимации, можно достигать определенного уровня значений показателя качества.

В то же время предложенная методика построения регрессионных моделей обработки данных обладает определенным ограничениями. Во-первых, рассматриваемая выборка должна обладать свойствами генеральной совокупности. В случае несоблюдения этого свойства будет возникать проблема адекватности предлагаемой модели обработки данных. Во-вторых, в процессе разделения данных на сегменты могут возникать ситуации, когда в обучающей выборке оказывается слишком мало объектов наблюдения, что не позволяет правильно оценить свойства и обучить модели на этих сегментах. В-третьих, при использовании предлагаемого решения желательно проводить анализ данных, чтобы избегать проблем переобучения.

Тем не менее предложенная методика позволяет формировать сегменты информационной последовательности на основе использования функционала качества моделей обработки информационных последовательностей. А это, в свою очередь, дает возможность реализовать агрегационную модель, где выполняется назначение лучших по качественным показателям моделей на сегменты.

Литература

1. Chen H.Y., Chen C. Evaluation of Calibration Equations by Using Regression Analysis: An Example of Chemical Analysis // Sensors. 2022. vol. 22. no. 2. DOI: 10.3390/s22020447.
2. Schober P., Vetter T.R. Segmented Regression in an Interrupted Time Series Study Design // Anesthesia and Analgesia. 2021. vol. 132. no. 3. pp. 696–697.
3. Bozpolat E. Investigation of the self-regulated learning strategies of students from the faculty of education using ordinal logistic regression analysis // Educational Sciences: Theory & Practice. 2016. no. 16(1). pp. 301–318.
4. Jarantow S.W., Pisors E.D., Chiu M.L. Introduction to the use of Linear and Nonlinear Regression Analysis in Quantitative Biological Assays // Current Protocols. 2023. no. 3. DOI: 10.1002/cpz1.801.
5. Britzger D. The Linear Template Fit // The European Physical Journal C. 2022. vol. 82(8). DOI: 10.1140/epjc/s10052-022-10581-w.
6. Perperoglou A., Sauerbrei W., Abrahamowicz M., Schmid M. A review of spline function procedures in R // BMC Medical Research Methodology. 2019. vol. 19. pp. 1–16.
7. Ren J., Tapert S., Fan C.C., Thompson W.K. A semi-parametric Bayesian model for semi-continuous longitudinal data // Statistics in Medicine. 2022. vol. 41. no. 13. pp. 2354–2374.
8. Taye M.M. Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions // Computation. 2023. vol. 11. no. 3. DOI: 10.3390/computation11030052.
9. Колмогоров А.Н. О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного // Доклады АН СССР. 1957. Т. 114. № 5. С. 953–956.
10. Girosi F., Poggio T. Representation Properties of Networks: Kolmogorov's Theorem is Irrelevant. Neural Computation. 1989. vol. 4. no. 1. pp. 465–469.
11. Parhi R., Nowak R.D. Banach Space Representer Theorems for Neural Networks and Ridge Splines // Journal of Machine Learning Research. 2021. vol. 22(1). pp. 1960–1999.

12. Marques H.O., Swersky L., Sander J., Campello R.J., Zimek A. On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection, and ensembles // Data Mining and Knowledge Discovery. 2023. vol. 37. no. 4. pp. 1473–1517.
13. Li Y., Guo X., Lin W., Zhong M., Li Q., Liu Z., Zhong W., Zhu Z. Learning dynamic user interest sequence in knowledge graphs for click-through rate prediction // IEEE Transactions on Knowledge and Data Engineering. 2023. vol. 35. no. 1. pp. 647–657.
14. Rinaldo A., Wang D., Wen Q., Willett R., Yu Y. Localizing changes in highdimensional regression models // The 24th International Conference on Artificial Intelligence and Statistics. 2021. pp. 2089–2097.
15. Aue A., Rice G., Sönmez O. Detecting and dating structural breaks in functional data without dimension reduction // Journal of the Royal Statistical Society. Series B, Statistical Methodology. 2018. vol. 80. no. 3. pp. 509–529.
16. Datta A., Zou H., Banerjee S. Bayesian high-dimensional regression for change point analysis // Statistics and its Interface. 2019. vol. 12. no. 2. pp. 253–264. DOI: 10.4310/SII.2019.v12.n2.a6.
17. Melnyk I., Banerjee A. A spectral algorithm for inference in hidden semi-Markov models // Journal of Machine Learning Research. 2017. vol. 18. no. 35. pp. 1–39.
18. Haynes K., Fearnhead P., Eckley I.A. A computationally efficient nonparametric approach for changepoint detection // Statistics and Computing. 2017. vol. 27. pp. 1293–1305. DOI: 10.1007/s11222-016-9687-5.
19. Muggeo V. Estimating regression models with unknown break-points // Statistics in Medicine. 2003. vol. 22(19). pp. 3055–3071.
20. Lu K.P., Chang S.T. A fuzzy classification approach to piecewise regression models // Applied Soft Computing Journal. 2018. vol. 69. pp. 671–688.
21. Bardwell L., Fearnhead P. Bayesian detection of abnormal segments in multiple time series // Bayesian Analysis. 2017. vol. 12. no. 1. pp. 193–218.
22. Huang J., Chen P., Lu L., Deng Y., Zou Q. WCDForest: a weighted cascade deep forest model toward the classification tasks // Applied Intelligence, 2023. vol. 53. no. 23. pp. 29169–29182. DOI: 10.1007/s10489-023-04794-z.
23. Tong W., Wang Y., Liu D. An Adaptive Clustering Algorithm Based on Local-Density Peaks for Imbalanced Data Without Parameters // IEEE Transactions on Knowledge and Data Engineering. 2023. vol. 35. no. 4. pp. 3419–3432.
24. Lu K.P., Chang S.T. Fuzzy maximum likelihood change-point algorithms for identifying the time of shifts in process data // Neural Computing and Applications. 2019. vol. 31. pp. 2431–2446.
25. Nevendra M., Singh P. Software defect prediction using deep learning // Acta Polytechnica Hungarica. 2021. vol. 18. no. 10. pp. 173–189.
26. Tallman E., West M. Bayesian predictive decision synthesis // Journal of the Royal Statistical Society. Series B: Statistical Methodology. 2024. vol. 86. no. 2. pp. 340–363.
27. Korkas K., Fryzlewicz P. Multiple change-point detection for non-stationary time series using wild binary segmentation. Statistica Sinica. 2017. vol. 27. pp. 287–311. DOI: 10.5705/ss.202015.0262.
28. Silva R.P., Zarpelão B.B., Cano A., Junior S.B. Time Series Segmentation Based on Stationarity Analysis to Improve New Samples Prediction // Sensors. 2021. vol. 21(21). DOI: 10.3390/s21217333.
29. Barzegar V., Laflamme S., Hu C., Dodson J. Multi-Time Resolution Ensemble LSTMs for Enhanced Feature Extraction in High-Rate Time Series // Sensors. 2021. vol. 21(6). DOI: 10.3390/s21061954.
30. Si S., Zhao J., Cai Z., Dui H. Recent advances in system reliability optimization driven by importance measures // Frontiers of Engineering Management. 2020. vol. 7. no. 3. pp. 335–358.

31. Xu S., Song Y., Hao X. A Comparative Study of Shallow Machine Learning Models and Deep Learning Models for Landslide Susceptibility Assessment Based on Imbalanced Data // Forests. 2022. vol. 13. no. 11. DOI: 10.3390/f13111908.
32. Лебедев И.С. Адаптивное применение моделей машинного обучения на отдельных сегментах выборки в задачах регрессии и классификации // Информационно-управляющие системы. 2022. № 3(118). С. 20–30.
33. Тихонов Д.Д., Лебедев И.С. Метод формирования сегментов информационной последовательности с использованием функционала качества моделей обработки // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24. № 3. С. 474–482.
34. Lebedev I.S., Sukhoparov M.E. Adaptive Learning and Integrated Use of Information Flow Forecasting Methods // Emerging Science Journal. 2023. vol. 7. no. 3. pp. 704–723.
35. Osipov V., Nikiforov V., Zhukova N., Miloserdov D. Urban traffic flows forecasting by recurrent neural networks with spiral structures of layers // Neural Computing and Applications. 2020. vol. 32. no. 18. pp. 14885–14897.
36. Lebedev I.S., Sukhoparov M.E. Improving the Quality Indicators of Multilevel Data Sampling Processing Models Based on Unsupervised Clustering // Emerging Science Journal. 2024. vol. 8. no. 1. pp. 355–371.
37. Jin H., Yin G., Yuan B., Jiang F. Bayesian hierarchical model for change point detection in multivariate sequences // Technometrics. 2022. vol. 64. no. 2. pp. 177–186.
38. Power Supply dataset. URL: http://www.cse.fau.edu/~xqzhu/stream.html (дата обращения: 16.05.2024).
39. Lu K.-P., Chang S.-T. An Advanced Segmentation Approach to Piecewise Regression Models // Mathematics. 2023. vol. 11(24). DOI: 10.3390/math11244959.
40. Energy generation dataset. URL: https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather/data?select=energy_dataset.csv (дата обращения: 16.05.2024).
41. Pima Indians Diabetes Database URL: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (дата обращения: 16.05.2024).
42. E-Commerce Data URL: https://www.kaggle.com/datasets/carrie1/ecommerce-data (дата обращения: 16.05.2024).

Лебедев Илья Сергеевич — д-р техн. наук, профессор, главный научный сотрудник, лаборатория интеллектуальных систем, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: методы машинного обучения, представление и обработка слабоструктурированных данных, применение методов искусственного интеллекта в системах информационной безопасности. Число научных публикаций — 200. isl_box@mail.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)508-3311.

Поддержка исследований. Исследование выполнено за счет гранта Российского научного фонда № 25-21-00269, https://rscf.ru/project/25-21-00269/.

DOI 10.15622/ia.24.2.1

I. LEBEDEV

ADAPTIVE REGRESSION MODEL CONSTRUCTION BASED ON THE FUNCTIONAL QUALITY ANALYSIS OF THE SEQUENCE SEGMENT PROCESSING

Lebedev I. **Adaptive Regression Model Construction Based on the Functional Quality Analysis of the Sequence Segment Processing.**

Abstract. The article considers the problem of constructing an adaptive model aimed at improving the quality indicators of processing information sequences. In data processing techniques that have found application in many application areas, the applied analysis of observation objects is computationally resource-intensive and requires many iterations in case of changes in data properties. The article proposes a technique for selecting segments of an information sequence obtained in different ways, which differs in the use of the quality functional of regression models for processing subsequences. The sequences of observation objects received at the input of the model are divided by various specified segmentation algorithms. Pre-selected regression models are trained on each obtained segment and, depending on the obtained values of the calculated quality functional, the best models in terms of quality indicators are assigned to the segments. This allows us to form an aggregation model for data processing. Based on the experiment on model data and samples, the proposed technique is assessed. The values of the quality indicator MSE and MAE are obtained for different processing algorithms and with a different number of segments. The proposed method makes it possible to increase the MSE and MAE indicators by segmentation and assignment of regression models that have the best indicators on individual segments. The proposed solution is aimed at further improvement of ensemble methods. Its application allows to increase the efficiency of setting up basic algorithms in case of data property transformation and to improve the interpretability of results. The method can be used in developing models and methods for processing information sequences.

Keywords: machine learning, adaptive models, improving the quality of processing, regression models.

References

1. Chen H.Y., Chen C. Evaluation of Calibration Equations by Using Regression Analysis: An Example of Chemical Analysis. Sensors. 2022. vol. 22. no. 2. DOI: 10.3390/s22020447.
2. Schober P., Vetter T.R. Segmented Regression in an Interrupted Time Series Study Design. Anesthesia and Analgesia. 2021. vol. 132. no. 3. pp. 696–697.
3. Bozpolat E. Investigation of the self-regulated learning strategies of students from the faculty of education using ordinal logistic regression analysis. Educational Sciences: Theory & Practice. 2016. no. 16(1). pp. 301–318.
4. Jarantow S.W., Pisors E.D., Chiu M.L. Introduction to the use of Linear and Nonlinear Regression Analysis in Quantitative Biological Assays. Current Protocols. 2023. no. 3. DOI: 10.1002/cpz1.801.
5. Britzger D. The Linear Template Fit. The European Physical Journal C. 2022. vol. 82(8). DOI: 10.1140/epjc/s10052-022-10581-w.
6. Perperoglou A., Sauerbrei W., Abrahamowicz M., Schmid M. A review of spline function procedures in R. BMC Medical Research Methodology. 2019. vol. 19. pp. 1–16.
7. Ren J., Tapert S., Fan C.C., Thompson W.K. A semi-parametric Bayesian model for semi-continuous longitudinal data. Statistics in Medicine. 2022. vol. 41. no. 13. pp. 2354–2374.

8. Taye M.M. Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. Computation. 2023. vol. 11. no. 3. DOI: 10.3390/computation11030052.
9. Kolmogorov A.N. [On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition]. Reports of the USSR Academy of Sciences – Doklady AN SSSR. 1957. vol. 114. no. 5. pp. 953–956.
10. Girosi F., Poggio T. Representation Properties of Networks: Kolmogorov's Theorem is Irrelevant. Neural Computation. 1989. vol. 4. no. 1. pp. 465–469.
11. Parhi R., Nowak R.D. Banach Space Representer Theorems for Neural Networks and Ridge Splines. Journal of Machine Learning Research. 2021. vol. 22(1). pp. 1960–1999.
12. Marques H.O., Swersky L., Sander J., Campello R.J., Zimek A. On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection, and ensembles. Data Mining and Knowledge Discovery. 2023. vol. 37. no. 4. pp. 1473–1517.
13. Li Y., Guo X., Lin W., Zhong M., Li Q., Liu Z., Zhong W., Zhu Z. Learning dynamic user interest sequence in knowledge graphs for click-through rate prediction. IEEE Transactions on Knowledge and Data Engineering. 2023. vol. 35. no. 1. pp. 647–657.
14. Rinaldo A., Wang D., Wen Q., Willett R., Yu Y. Localizing changes in highdimensional regression models. The 24th International Conference on Artificial Intelligence and Statistics. 2021. pp. 2089–2097.
15. Aue A., Rice G., Sönmez O. Detecting and dating structural breaks in functional data without dimension reduction. Journal of the Royal Statistical Society. Series B, Statistical Methodology. 2018. vol. 80. no. 3. pp. 509–529.
16. Datta A., Zou H., Banerjee S. Bayesian high-dimensional regression for change point analysis. Statistics and its Interface. 2019. vol. 12. no. 2. pp. 253–264. DOI: 10.4310/SII.2019.v12.n2.a6.
17. Melnyk I., Banerjee A. A spectral algorithm for inference in hidden semi-Markov models. Journal of Machine Learning Research. 2017. vol. 18. no. 35. pp. 1–39.
18. Haynes K., Fearnhead P., Eckley I.A. A computationally efficient nonparametric approach for changepoint detection. Statistics and Computing. 2017. vol. 27. pp. 1293–1305. DOI: 10.1007/s11222-016-9687-5.
19. Muggeo V. Estimating regression models with unknown break-points. Statistics in Medicine. 2003. vol. 22(19). pp. 3055–3071.
20. Lu K.P., Chang S.T. A fuzzy classification approach to piecewise regression models. Applied Soft Computing Journal. 2018. vol. 69. pp. 671–688.
21. Bardwell L., Fearnhead P. Bayesian detection of abnormal segments in multiple time series. Bayesian Analysis. 2017. vol. 12. no. 1. pp. 193–218.
22. Huang J., Chen P., Lu L., Deng Y., Zou Q. WCDForest: a weighted cascade deep forest model toward the classification tasks. Applied Intelligence, 2023. vol. 53. no. 23. pp. 29169–29182. DOI: 10.1007/s10489-023-04794-z.
23. Tong W., Wang Y., Liu D. An Adaptive Clustering Algorithm Based on Local-Density Peaks for Imbalanced Data Without Parameters. IEEE Transactions on Knowledge and Data Engineering. 2023. vol. 35. no. 4. pp. 3419–3432.
24. Lu K.P., Chang S.T. Fuzzy maximum likelihood change-point algorithms for identifying the time of shifts in process data. Neural Computing and Applications. 2019. vol. 31. pp. 2431–2446.
25. Nevendra M., Singh P. Software defect prediction using deep learning. Acta Polytechnica Hungarica. 2021. vol. 18. no. 10. pp. 173–189.
26. Tallman E., West M. Bayesian predictive decision synthesis. Journal of the Royal Statistical Society. Series B: Statistical Methodology. 2024. vol. 86. no. 2. pp. 340–363.

27. Korkas K., Fryzlewicz P. Multiple change-point detection for non-stationary time series using wild binary segmentation. Statistica Sinica. 2017. vol. 27. pp. 287–311. DOI: 10.5705/ss.202015.0262.
28. Silva R.P., Zarpelão B.B., Cano A., Junior S.B. Time Series Segmentation Based on Stationarity Analysis to Improve New Samples Prediction. Sensors. 2021. vol. 21(21). DOI: 10.3390/s21217333.
29. Barzegar V., Laflamme S., Hu C., Dodson J. Multi-Time Resolution Ensemble LSTMs for Enhanced Feature Extraction in High-Rate Time Series. Sensors. 2021. vol. 21(6). DOI: 10.3390/s21061954.
30. Si S., Zhao J., Cai Z., Dui H. Recent advances in system reliability optimization driven by importance measures. Frontiers of Engineering Management. 2020. vol. 7. no. 3. pp. 335–358.
31. Xu S., Song Y., Hao X. A Comparative Study of Shallow Machine Learning Models and Deep Learning Models for Landslide Susceptibility Assessment Based on Imbalanced Data. Forests. 2022. vol. 13. no. 11. DOI: 10.3390/f13111908.
32. Lebedev I.S. [Adaptive application of machine learning models on separate segments of a data sample in regression and classification problems]. Informatsionno-upravliaiushchie sistemy – Information and Control Systems. 2022. no. 3(118). pp. 20–30.
33. Tikhonov D.D., Lebedev I.S. [Method for generating information sequence segments using the quality functional of processing models]. Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki – Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2024. vol. 24. no. 3. pp. 474–482.
34. Lebedev I.S., Sukhoparov M.E. Adaptive Learning and Integrated Use of Information Flow Forecasting Methods. Emerging Science Journal. 2023. vol. 7. no. 3. pp. 704–723.
35. Osipov V., Nikiforov V., Zhukova N., Miloserdov D. Urban traffic flows forecasting by recurrent neural networks with spiral structures of layers. Neural Computing and Applications. 2020. vol. 32. no. 18. pp. 14885–14897.
36. Lebedev I.S., Sukhoparov M.E. Improving the Quality Indicators of Multilevel Data Sampling Processing Models Based on Unsupervised Clustering. Emerging Science Journal. 2024. vol. 8. no. 1. pp. 355–371.
37. Jin H., Yin G., Yuan B., Jiang F. Bayesian hierarchical model for change point detection in multivariate sequences. Technometrics. 2022. vol. 64. no. 2. pp. 177–186.
38. Power Supply dataset. Available at: http://www.cse.fau.edu/~xqzhu/stream.html (accessed 16.05.2024).
39. Lu K.-P., Chang S.-T. An Advanced Segmentation Approach to Piecewise Regression Models. Mathematics. 2023. vol. 11(24). DOI: 10.3390/math11244959.
40. Energy generation dataset. Available at: https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather/data?select=energy_dataset.csv (accessed 16.05.2024).
41. Pima Indians Diabetes Database. Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (accessed 16.05.2024).
42. E-Commerce Data. Available at: https://www.kaggle.com/datasets/carrie1/ecommerce-data (accessed 16.05.2024).

Lebedev Ilya — Ph.D., Dr.Sci., Professor, Chief scientific officer, Laboratory of intelligent systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: machine learning methods, representation and processing of weakly structured data, application of artificial intelligence methods in information security systems. The number of publications — 200. isl_box@mail.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)508-3311.

Acknowledgements. This research is supported by RSF (grant № 25-21-00269, https://rscf.ru/project/25-21-00269/).