УДК 621.396
doi:10.31799/1684-8853-2021-3-29-38

Научные статьи
Articles

Сегментирование множества данных с учетом информации воздействующих факторов

И. С. Лебедев[a], *доктор техн. наук, профессор, orcid.org/0000-0001-6753-2181, isl_box@mail.ru*
[a]*Санкт-Петербургский федеральный исследовательский центр РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*

Введение: *применение методов машинного обучения предполагает сбор и обработку в автономном режиме значений, поступающих от регистрирующих элементов. Большинство моделей обучается на исторических данных, а затем применяется в задачах прогнозирования, классификации, поиска влияющих факторов и воздействий, анализа состояния. В течение времени могут меняться диапазоны регистрируемых значений, что сказывается на качестве результатов классификационных алгоритмов и приводит к тому, что модели должны обучаться постоянно либо перенастраиваться с учетом поступающих значений параметров.* ***Цель:*** *разработка методики, повышающей показатели качества алгоритмов машинного обучения в условиях динамически изменяющихся и нестационарных сред, где распределение данных может изменяться с течением времени.* ***Методы:*** *разбиение (сегментирование) множества данных на основе информации о факторах, влияющих на диапазоны значений целевых переменных.* ***Результаты:*** *предложена методика сегментирования множества данных, основанная на учете факторов, которые влияют на изменение диапазонов значений целевых переменных. Выявление воздействий дает возможность сформировать выборки исходя из текущих и предполагаемых ситуаций. На примере датасета PowerSupply реализовано разбиение множества данных на подмножества, учитывающее влияние факторов на диапазоны значений. Приведена формализация внешних воздействий на основе продукционных правил. Показана обработка факторов с помощью функции принадлежности (индикаторной функции). С ее использованием произведено разбиение выборки данных на конечное число непересекающихся измеримых подмножеств. Приведены экспериментальные значения функции потерь MSE нейросети для предлагаемой методики на выбранном датасете. Показаны результаты качественных показателей классификации (Accuracy, AUC, F-мера) для различных классификаторов.* ***Практическая значимость:*** *результаты могут быть использованы при разработке классификационных моделей методов машинного обучения. Предложенная методика позволяет повысить показатели качества классификации в изменяющихся условиях функционирования.*

Ключевые слова *— машинное обучение, сегментирование множества данных, воздействующие факторы, изменяющиеся условия.*

Для цитирования: Лебедев И. С. Сегментирование множества данных с учетом информации воздействующих факторов. *Информационно-управляющие системы*, 2021, № 3, с. 29–38. doi:10.31799/1684-8853-2021-3-29-38
For citation: Lebedev I. S. Dataset segmentation considering the information about impact factors. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 3, pp. 29–38 (In Russian). doi:10.31799/1684-8853-2021-3-29-38

Введение

Развитие технологий вызывает лавинообразный рост информации. В связи с этим необходимо разрабатывать эффективные методы анализа и обработки постоянно увеличивающихся объемов данных в различных информационных системах.

Традиционное применение методов машинного обучения предполагает сбор и обработку в автономном режиме значений, поступающих от регистрирующих элементов. Большинство моделей обучается на исторических данных, а затем применяется в задачах прогнозирования, классификации, поиска влияющих факторов и воздействий, анализа состояния.

Во многих информационных системах наблюдения производятся одновременно множеством регистрирующих элементов, информация представляется временными рядами. В разных ситуациях могут меняться диапазоны регистрируемых параметров; это влияет на качество результатов классификационных алгоритмов и приводит к тому, что модели должны обучаться постоянно либо перенастраиваться с учетом непрерывно поступающих значений.

В динамически меняющихся и нестационарных средах распределение данных может становиться другим с течением времени, что приводит к «дрейфу концепций» [1, 2], когда возникают изменения условного распределения выходных данных от значений входных признаков, в то время как распределение входных данных может оставаться неизменным.

Рост объемов разнородной информации о поведении информационных, физических процессов, протекающих в технологических системах, и требования повышения качества анализа состояния элементов и узлов обуславливают необходимость адаптировать методы машинного обучения к возникающим изменениям диапазонов значений целевых переменных.

Обзор существующих методов

Большая часть исследований, направленных на повышение качественных показателей идентификации состояния информационных систем в условиях динамически протекающих процессов, фокусируется на проблемах классификации и адаптации дрейфа концепций. Основные виды этого явления представлены и описаны в ряде работ [3–5].

Методы обнаружения и обработки дрейфа концепций делятся на контролируемые, требующие заранее заданной модели или значений параметров, и неконтролируемые подходы.

Трансформация свойств анализируемого явления, связанного с изменением диапазонов значений переменных его кортежей, приводит к тому, что модель становится неактуальной. В связи с этим возникает ряд фундаментальных аспектов применения методов машинного обучения. Текущий этап исследований сфокусирован на задачах обнаружения и реакции на дрейф, идентификации ложных «выбросов» данных, устойчивости к ошибкам первого и второго рода, быстрого обнаружения аномальных событий на небольшом количестве наблюдений.

Используется несколько направлений к решению обозначенных проблемных вопросов.

Первое связано с ансамблями классифицирующих алгоритмов, обученных на подмножествах данных [6–9]. Суть методов состоит в объединении прогнозов моделей. Дрейф концепций определяется анализом статистического расхождения результатов, выдаваемых классификаторами. Если установленная доля ответов находится выше порога, то рассматривается гипотеза о смещении значений целевых переменных. Однако эти методы не являются универсальными, имеют сложности, связанные с формированием модели производящих оценку достоверности классификаторов.

Второе направление базируется на контроле распределения вероятностей. Такие методы направлены на обнаружение возможных изменений диапазонов обрабатываемых данных. Они требуют большого количества ресурсов и в определенных ситуациях характеризуются высокой частотой ложных тревог [10–12].

Третье направление — разработка моделей проявления дрейфа концепта. Такие модели не являются универсальными [10, 13–15], требуют больших вычислительных затрат и адаптации классифицирующих алгоритмов. Они основаны на предварительных знаниях о свойствах концептов, присутствующих в данных. В случае большого количества анализируемых целевых переменных формируется множество классифицирующих моделей и разрабатываются сложные решающие правила.

В большинстве случаев применяемые на сегодня методы являются узкоспециализированными и требуют существенных затрат на реализацию [14, 16–18].

В реальной среде данные всегда имеют несовершенную форму, являются примерами несбалансированных выборок. В то же время существует ряд факторов, влияющих на значения показателей. Они могут быть известны заранее, действовать с определенной периодичностью и изменять регистрируемые результаты в пределах некоторого заранее оцениваемого диапазона.

Использование информации о влияющих на диапазоны значений факторах дает возможность сформировать выборки, позволяющие повысить качество алгоритмов классификации, вследствие чего предлагается методика разбиения (сегментации) множества данных на основе выбранных факторов. В результате ее применения получается несколько подмножеств, каждое из которых определено исходя из влияния выбранного фактора. В дальнейшем для повышения качественных показателей на каждую выборку может назначаться свой классифицирующий алгоритм либо его параметры, влияющие на результат классификации, которые могут меняться с учетом влияния фактора.

Описание предлагаемой методики

Одним из основополагающих факторов, определяющих результат алгоритмов машинного обучения, является формирование обучающего подмножества. Наличие качественных выборок данных во многих случаях гораздо важнее качества алгоритмов [19–21]. Ошибки формирования множеств примеров, на которых обучаются и тестируются классификаторы, предопределяют эффективность модели. Однако при этом необходимо учитывать, что могут проявляться изменения в распределении данных во времени в различных формах.

Формализованное описание постановки задачи можно представить следующим образом.

$\mathbf{X}$ — множество описаний объектов, $x \in \mathbf{X}$ — это d-мерный кортеж признаков в предопределенном векторном пространстве $\mathbf{X} = \mathbf{R}^d$.

Множество классов, поставленных в соответствие описаниям x, отмеченных метками $\{c_1, c_2, ..., c_l\} \in C$, разбивается на бинарное подмножество состояний, объединенных классами $\{C_1, C_2\} \in C$.

Имеется множество факторов V, влияющих на значения признаков. Множество $\mathbf{X}$ характеризуется множеством значений признаков f: $\mathbf{X} \xrightarrow{V} D_f$.

С учетом действующего в текущий момент времени фактора $v \in V$ определяется признаковое описание объекта наблюдения $\mathbf{X}_i \in \mathbf{X}$ в виде $\mathbf{X}_i = (f_1(v, x), ..., f_n(v, x))$ при заданных $f_1, ..., f_n$.

Временная метка	Значения признаков			
	Признак 1 f_1	Признак 2 f_2	...	Признак n f_n
t_0	x_{10}	x_{20}	...	x_{n0}
t_1	x_{11}	x_{21}	...	x_{n1}
t_2	x_{12}	x_{22}	...	x_{n2}
...	...	...	...	...
t_m	x_{1m}	x_{2m}	...	x_{nm}

Влияющий фактор	Метка класса (принадлежность событию)	Временная метка	Значения признаков			
			Признак 1 f_1	Признак 2 f_2	...	Признак n f_n
v_1	c_0	t_0	x_{10}	x_{20}	...	x_{n0}
	c_1	t_1	x_{11}	x_{21}	...	x_{n1}
	c_0	t_2	x_{12}	x_{22}	...	x_{n2}
	c_0	t_3	x_{10}	x_{20}	...	x_{n0}
	c_0	t_4	x_{11}	x_{21}	...	x_{n1}
v_2	c_1	t_5	x_{12}	x_{22}	...	x_{n2}
	c_1	t_6	x_{10}	x_{20}	...	x_{n0}
	c_0	t_7	x_{11}	x_{21}	...	x_{n1}
	c_1	t_8	x_{12}	x_{22}	...	x_{n2}
...	...	...	...	...	...	...
v_k	c_1	t_m	x_{1m}	x_{2m}	...	x_{nm}

■ ***Рис. 1.*** Преобразование датасета
■ ***Fig. 1.*** Dataset transformation

Обучающее множество приобретает вид размеченной выборки $\{v_j, \{(x_i, c_i)\}_{i=1}^{N}\}_{j=1}^{M}$, где i — количество кортежей в момент, когда оказывал влияние рассматриваемый j-й фактор v_j.

Необходимо построить классифицирующий алгоритм, учитывающий влияние фактора v, $\alpha : \mathbf{X} \xrightarrow{V} C$. Алгоритм определяет по входному признаковому описанию соответствие классу.

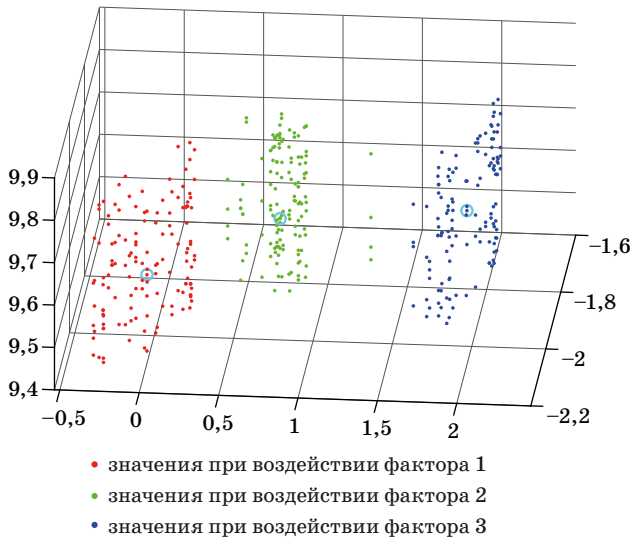
Обучение модели усложняется не только большой размерностью признакового пространства, но и наличием влияющих на значения признаковых переменных факторов.

Основным ограничением методов машинного обучения является то, что алгоритмы классификации не всегда могут быть эффективны в условиях постоянно функционирующей под влиянием различных внешних и внутренних воздействий системы. Система находится в динамике, осуществляются постоянные переходы из одного состояния в другое. Внешние и внутренние факторы изменяют значения характеристик. Например, подключение нового устройства сетевого сегмента может в течение определенного момента времени вызывать изменения анализируемых значений параметров, таких как рост количества служебных сообщений, конфигурационных команд, увеличение времени задержки и т. д.

Накопленная статистическая информация о функционировании системы представляется в виде матрицы, где временной отметке соответствуют значения признаков, полученных от регистрирующих устройств и элементов. Различные факторы влияют на свойства объектов. Изменения значений признаков являются результатом воздействий факторов. Часть из них, например сезонность, являются известными в текущий момент времени.

Внесение дополнительных полей в разметку обучающего подмножества дает возможность учитывать информацию о влияющих факторах. На рис. 1 представлен вид записей исходного датасета и датасета после разметки и добавления информации о влияющих факторах. В результате такого представления записи связываются с отдельными факторами.

Появляется возможность разбить пространство на подпространства, где происходит классификация отдельных признаковых сегментов, и рассматривать множество полученных значений для каждого фактора. На рис. 2 показан пример



■ ***Рис. 2.*** Пример значений из датасета при воздействии разных факторов
■ ***Fig. 2.*** Dataset values example exposed to various factors

разбиения множества данных на подмножества с учетом наблюдаемых факторов.

Задача отнесения состояния наблюдаемого объекта решается с учетом текущего информационного воздействия.

Анализ и учет влияющих на данные факторов позволяют реализовать разбиение множества на подмножества. В дальнейшем, определяя свойства полученных выборок, можно осуществить решение задачи о применении наиболее эффективных алгоритмов обработки.

Применение методики

Рассмотрим классификатор $\varphi(x, \mathbf{W})$. На вход поступает кортеж значений x. Для принятия решения используется весовая матрица $\mathbf{W}$. Возможны два направления разделения множества данных: использование продукционных правил и функций принадлежности.

Применение продукций предполагает, что влияющие на значения данных факторы поддаются эвристике. Она дает возможность определить правила, учитывающие воздействия на значения выборки. Например, периодичность процессов в экономике, энергетике, других областях позволяет сформировать ряд продукционных правил. В общем виде такая модель представляется в предикативном виде:

$$M = <\Phi, V, \mathbf{W}, \mathbf{X}>,$$

где Φ — классифицирующие алгоритмы, использующие весовые матрицы для сравнения поступающего кортежа данных; V — множество влияющих факторов на целевые переменные в выборке данных; $\mathbf{W}$ — множество весовых матриц классификаторов, значения матриц зависят от фактора, влияющего на данные в системе; $\mathbf{X}$ — множество описаний объектов, состоящее из подмножества выборок данных, каждому подмножеству соответствуют свои весовые матрицы классификаторов.

Выбор значений $\mathbf{w}_j \in \mathbf{W}$ может быть осуществлен на основе продукционной модели. Определяется подмножество данных $\mathbf{X}_j$ с учетом влияния фактора v_j. Каждому подмножеству может назначаться классифицирующий алгоритм $\varphi \in \Phi$. Сгруппированным переменным «влияющий фактор — подмножество» v_j, $\mathbf{X}_j$ определяется матрица $\mathbf{w}_i$ с учетом свойств классифицирующих алгоритмов. Продукция, реализующая правило, примет вид

$$(v_j, \mathbf{X}_j) \xrightarrow{\varphi_k} w_j. \quad (1)$$

Выражение (1) позволяет после определения текущего воздействия v_j использовать матрицу $\mathbf{w}_j$. Поступающий на вход новый кортеж x идентифицируется классификатором $\varphi(x, \mathbf{w}_j)$.

В дальнейшем возможно реализовать правила не только для разбиения на сегменты множества данных, но и для выбора соответствующего классификатора в момент действия фактора. Это позволит учитывать изменения в данных и повышать показатели качества классифицирующей модели в целом.

Второе направление базируется на использовании функции принадлежности. Оно может применяться, когда существуют воздействия, поддающиеся аналитическому описанию (например, длина светового дня в зависимости от времени года, широта места в подсистеме подачи электроэнергии на объекты городского хозяйства, часы пиковой нагрузки в информационной системе). Такие факторы v_j могут быть обработаны с помощью функции принадлежности (индикаторной функции). На ее основе производится разбиение выборки данных $\mathbf{X}$ на конечное число не пересекающихся измеримых подмножеств $\mathbf{X}_1$, ..., $\mathbf{X}_n$. В простейшем случае функция принадлежности μ подмножества $\mathbf{X}_j \in \mathbf{X}$, где x — кортеж обучающей выборки, может быть представлена в виде

$$\mu_{\mathbf{X}_j}(x, v_j) = \begin{cases} 1, & x \in \mathbf{X}_j \\ 0, & x \notin \mathbf{X}_j \end{cases}. \quad (2)$$

Выражение (2) позволяет определить принадлежность элемента выборки данных $x \in \mathbf{X}$ подмножеству $\mathbf{X}_j$ в момент действия фактора v_j.

В общем случае выборка состоит из n подмножеств. Принадлежность подмножеству определяется функциями $\mu_{\mathbf{X}_1}(x)$, $\mu_{\mathbf{X}_2}(x)$, ..., $\mu_{\mathbf{X}_n}(x)$.

Получается разбиение $\mathbf{X}_1 \cup \mathbf{X}_2 \cup ... \cup \mathbf{X}_n = \mathbf{X}$ при условии $\mathbf{X}_j \cap \mathbf{X}_i \neq 0 \ \forall \ i \neq j$. Объединение подмножеств совпадает с множеством $\mathbf{X}$, подмножества не пересекаются.

Классификацию становится возможным осуществить на каждом из подмножеств $\mathbf{X}_1$, ..., $\mathbf{X}_n$. Формируется тестовая и обучающая выборка с учетом действующих факторов v_j. В зависимости от обрабатываемого подмножества классификатор $\varphi(x, \mathbf{w})$ может быть дополнен функцией $\psi(v_j)$. Функция $\psi(v_j)$ учитывает фактор v_j, влияющий на подмножество $\mathbf{X}_j$, и определяет по его значению весовую матрицу $\mathbf{w}_j = \psi(v_j)$. Классификатор примет вид $\varphi(x, \psi(v_j))$.

В качестве одной из мер оценки модели может быть применена функция потерь.

Функция потерь $L(v_j)$ для подмножества $\mathbf{X}_j$ определяется выражением

$$L(v_j) = \frac{1}{N}\sum\nolimits_i L_i(\varphi(x_i, \psi(v_j)), c_i) + \lambda R(\psi(v_j)), \quad (3)$$

где $R(\psi(v_j))$ — функция регуляризации; λ — коэффициент регуляризации. Они влияют на дисперс-

сию и смещение ответов классификатора. Регуляризация предназначена для добавления дополнительных ограничений, предотвращающих переобучение.

Средняя сумма потерь для данных множества **X**

$$L = \frac{1}{M}\sum_{j=1}^{M} L(v_j). \qquad (4)$$

Применяя (3) и (4) и минимизируя среднюю сумму потерь, можно найти оптимальные параметры на основе выражения

$$L = \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}(L_i(\varphi(x_i,\ \psi(v_j)),\ c_i) + \\ + \lambda R(\psi(v_j)) \to \min. \qquad (5)$$

В предлагаемой методике с использованием выражения (5) появляется возможность определить разбиение выборки на подмножества с учетом факторов *V*.

Эксперимент

Эксперимент проводился на множестве датасета PowerSupply [22], содержащего информацию о почасовой подаче электроэнергии итальянской энергетической компании. В нем представлены данные о потребляемой мощности из электрических сетей каждый час с 1995 по 1998 год. Различные факторы (рабочее и нерабочее время дня, будни и выходные, сезонность, перемена погоды) вызывают явление дрейфа концепта.

В эксперименте рассматривалось предсказание рабочего и нерабочего времени по входным данным потребляемой мощности. В качестве воздействующего фактора, на основе которого производилась сегментация выборки данных, была выбрана сезонность. В первом случае разбиение производилось по датам перехода на летнее и зимнее время, во втором — по календарным временам года.

Общий вид датасета представлен на рис. 3. В горизонтальной плоскости на осях показаны дни наблюдений и часы, по вертикальной оси отложена потребляемая мощность.

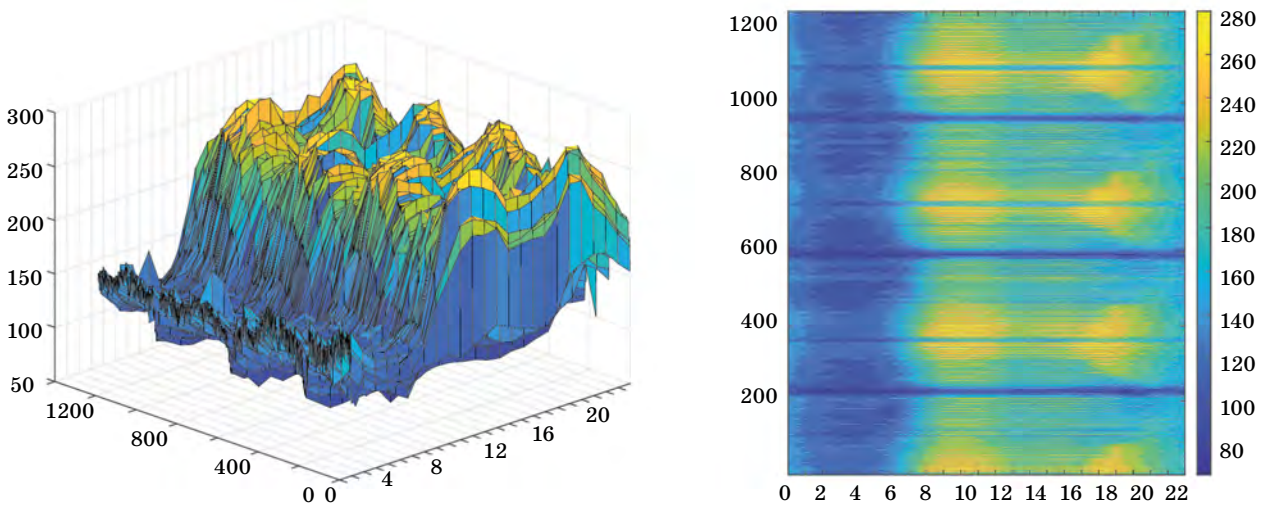
Даты перехода на летнее и зимнее время дают возможность реализовать правило или функцию принадлежности, с помощью которых выборка делится на два непересекающихся подмножества (рис. 4, *а* и *б*).

Второе разбиение было осуществлено для анализа изменений показателей качества классификаторов. Общая выборка была разделена на четыре части и содержала значения потребляемой энергии в весенние, летние, осенние, зимние месяцы. В дальнейшем два способа разбиения использовались для сравнения результатов классифицирующих алгоритмов.

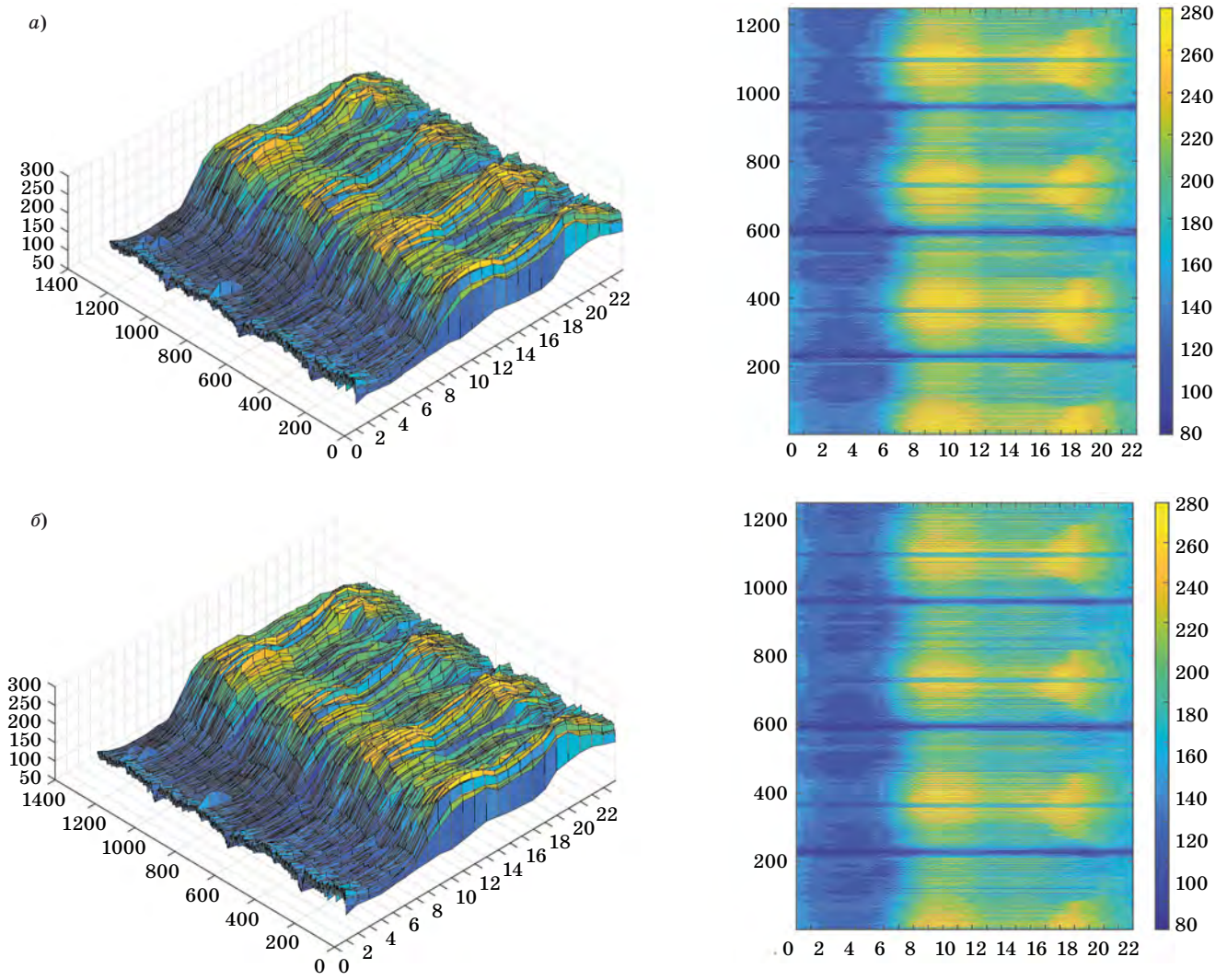
В качестве обучающих данных использовался переменный объем каждого подмножества, составляющий от 5 до 35 %.

Статистические свойства предсказываемой целевой переменной изменяются с течением времени. В рассматриваемых случаях на значения входящих в кортеж данных влияет заранее определенный сезонный фактор.

В первой части эксперимента для оценки влияния на классифицирующий алгоритм разбиения данных на подмножества была реализована



■ ***Рис. 3.*** Поверхность датасета потребления электроэнергии
■ ***Fig. 3.*** The dataset surface of electricity consumption



■ ***Рис. 4.*** Поверхность подмножества датасета потребления электроэнергии летнего (*а*) и зимнего (*б*) времени
■ ***Fig. 4.*** The dataset surface of electricity consumption “Summer Time” (*а*) and “Winter Time” (*б*)

двухслойная нейросеть. Значение точности сети (Accuracy) составляло около 0,75. Задача сети состояла в том, чтобы по входному кортежу признаков потребления электроэнергии пользователями спрогнозировать, происходит данная ситуация в рабочее или нерабочее время.

В качестве меры была выбрана среднеквадратичная функция потерь (MSE). Она является одним из основных показателей в задачах регрессии и чувствительна к выбросам данных.

Среднеквадратичная функция потерь для всего множества, представленного на рис. 3:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2. \qquad (6)$$

Для разделенного датасета, состоящего из двух или четырех частей ($M = 2$, $M = 4$), выражение (6) будет выглядеть следующим образом:

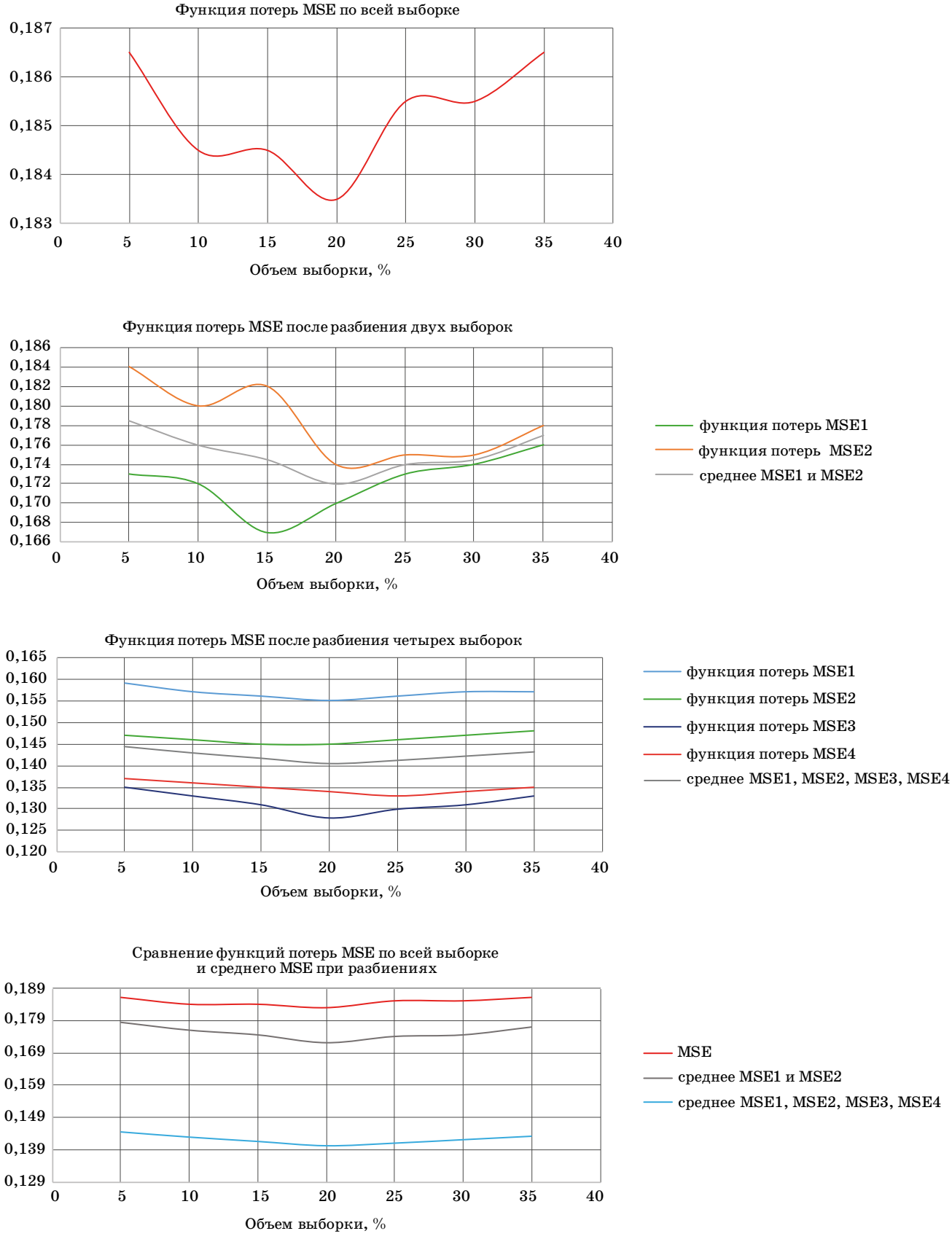
$$L_{MSE_{\mathbf{x}_1,\mathbf{x}_2}} = \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}(x_{ji} - \hat{x}_{ji})^2. \qquad (7)$$

В результате проведения эксперимента по обучению и оценки модели были получены следующие графики функции потерь (рис. 5), определяемой выражениями (6) и (7).

Графики показывают более низкие значения функции потерь для модели, когда пространство сегментировано, что позволяет говорить о целесообразности разделения на подмножества на основе информации о влияющем факторе.

Кроме того, на графиках виден момент переобучения нейросети для выбранных подмножеств, что позволяет оценить свойства данных для дальнейших стратегий обучения на подмножествах.

Во второй части эксперимента для общей оценки влияния применения подмножеств на ка-



■ ***Рис. 5.*** Значения функций потерь
■ ***Fig. 5.*** Loss functions values

чество результатов моделей машинного обучения были выбраны различные алгоритмы: линейного дискриминантного анализа (LD), квадратичного дискриминантного анализа (QD), наивного байесовского классификатора (NB), ближайших соседей (KNN).

На каждый классифицирующий алгоритм подавались наборы данных из двух подмножеств зимнего и летнего времени (см. рис. 3, 4) и четырех подмножеств весенних, летних, зимних, осенних месяцев.

Результаты тестирования классификаторов (AUC — площадь под ROC-кривой, точность Accuracy, F-мера) для целого множества $\mathbf{X}$ и усредненные значения для множеств $\mathbf{X}_{11}$ и $\mathbf{X}_{12}$ и множеств $\mathbf{X}_{21}$, $\mathbf{X}_{22}$, $\mathbf{X}_{23}$, $\mathbf{X}_{24}$ приводятся в таблице.

Результаты тестирования демонстрируют, что разбиение общей выборки на отдельные подмножества в основном позволяет повысить ряд показателей качества классификации для выбранных алгоритмов.

Разделение на два подмножества позволяет улучшить результаты на 1 %. При разбиении на четыре подмножества наблюдается повышение значений AUC, Accuracy, F-меры для каждого отдельного классификатора уже на 5–8 %, что сопоставимо с результатами ансамблевых моделей.

Можно осуществить дальнейшую операцию разбиения на подмножества, используя информацию о выходных и праздничных днях, погодных условиях и т. д. Предложенное решение можно использовать как дополнение к различным классификационным моделям. Допустима и более сложная сегментация, учитывающая дополнительные параметры, которая за счет уменьшения явления «выбросов» данных позволит повысить качественные показатели.

Заключение

Основные проблемные вопросы методов машинного обучения лежат в области формирования выборок данных, определяющих достижение заданных показателей. Качество обучающих подмножеств повышается путем устранения шумов, удаления дисбаланса классов, обнаружения дрейфа концепта.

Предложена методика сегментации выборок данных, основанная на учете факторов, которые влияют на изменение диапазонов целевых переменных. Выявление воздействий дает возможность сформировать сегментированные выборки данных исходя из текущих предполагаемых ситуаций. Для каждого полученного множества возможен поиск лучшей модели, реализующей классификационную задачу.

Наборы данных, тестовые множества имеют свои свойства. Применение предложенного решения позволяет уменьшить влияние шумовых данных, избежать введения дополнительных затрат на борьбу с явлением дрейфа концепта, повысить показатели полноты и точности за счет уменьшения разброса параметров.

Однако имеется несколько моментов, которые необходимо учитывать для реализации предлагаемой методики. Внутри сформированных сегментов выборок могут происходить изменения целевых переменных под влиянием других факторов. В обрабатываемых множествах могут быть другие концепции, влияющие на конечный результат. Для эффективного применения классифицирующей модели необходимо исследовать объемы, свойства обучаемых выборок, ограничения, связанные с длинами сегментов, пороговыми значениями диапазонов целевых переменных.

■ Результаты классифицирующих алгоритмов

■ Results of classifying algorithms

Модель	Объем выборки, %	AUC			Accuracy			F-мера		
		$\mathbf{X}$	Среднее $\mathbf{X}_{11}+\mathbf{X}_{12}$	Среднее $\mathbf{X}_{21}+\mathbf{X}_{22}+$ $+\mathbf{X}_{23}+\mathbf{X}_{24}$	$\mathbf{X}$	Среднее $\mathbf{X}_{11}+\mathbf{X}_{12}$	Среднее $\mathbf{X}_{21}+\mathbf{X}_{22}+$ $+\mathbf{X}_{23}+\mathbf{X}_{24}$	$\mathbf{X}$	Среднее $\mathbf{X}_{11}+\mathbf{X}_{12}$	Среднее $\mathbf{X}_{21}+\mathbf{X}_{22}+$ $+\mathbf{X}_{23}+\mathbf{X}_{24}$
LD	5	0,76	0,78	0,81	0,727905	0,730246	0,783536	0,790824	0,775388	0,821392
	35	0,77	0,78	0,82	0,717489	0,717681	0,784194	0,757563	0,758894	0,825022
QD	5	0,95	0,95	0,97	0,834698	0,872516	0,924455	0,872141	0,895602	0,943755
	35	0,96	0,96	0,99	0,873597	0,861341	0,925849	0,896387	0,886896	0,947586
NB	5	0,89	0,91	0,94	0,803321	0,811307	0,866226	0,844574	0,84626	0,892745
	35	0,91	0,92	0,96	0,812483	0,812493	0,864015	0,844612	0,84462	0,898299
KNN	5	0,88	0,95	0,98	0,827748	0,878408	0,925379	0,85605	0,898913	0,953179
	35	0,95	0,96	0,99	0,881582	0,895829	0,942752	0,901138	0,913712	0,968623

Литература

1. **Schlimmer J. C., Granger R. H.** Incremental learning from noisy data. *Machine Learning*, 1986, no. 1, pp. 317–354. doi:10.1023/A:1022810614389
2. **Widmer G., Kubat M.** Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 1996, no. 23(1), pp. 69–101. doi:10.1007/BF00116900
3. **Gama J., Žliobait I., Bifet A., Pechenizkiy M., Bouchachia A.** A survey on concept drift adaptation. *ACM Computing Surveys*, 2014, no. 46(4), pp. 1–37.
4. **Sung-Yu T., Jen-Yuan C.** Parametric study and design of deep learning on leveling system for smart manufacturing. *IEEE International Conference on Smart Manufacturing, Industrial & Logistics Engineering (SMILE)*, February 8–9, 2018, pp. 48–52. doi:10.1109/SMILE.2018.8353980
5. **Lu J., Liu A., Dong F., Gu F., Gama J., Zhang G.** Learning under concept drift: a review. *IEEE Transactions on Knowledge and Data Engineering*, 2019, no. 31(12), pp. 2346–2363.
6. **Wang L. Y., Park C., Yeon K., Choi H.** Tracking concept drift using a constrained penalized regression combiner. *Computational Statistics & Data Analysis*, 2017, no. 108, pp. 52–69.
7. **Khan S., Yairi T.** A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 2018, no. 107, pp. 241–265. doi:10.1016/j.ymssp.2017.11.024
8. **Salehi H., Burgueño R.** Emerging artificial intelligence methods in structural engineering. *Engineering Structures*, 2018, no. 171, pp. 170–189. doi:10.1016/j.engstruct.2018.05.084
9. **Zhou Z.-H., Feng J.** Deep forest. *National Science Review*, 2019, vol. 6, no. 1, pp. 74–86. doi:10.1093/nsr/nwy108
10. **Sethi T., Kantardzic M.** Handling adversarial concept drift in streaming data. *Expert Systems with Applications*, 2018, vol. 97, pp. 18–40. doi:10.1016/j.eswa.2017.12.022
11. **Takacs A., Toledano-Ayala M., Dominguez-Gonzalez A., Pastrana-Palma A., Velazquez D. T., Ramos J. M., Rivas-Araiza A. E.** Descriptor generation and optimization for a specific outdoor environment. *IEEE Access*, 2020, vol. 8, pp. 2169–3536. doi:10.1109/ACCESS.2020.2975474
12. **Saadallah A., Moreira-Matias L., Sousa R., Khiari J., Jenelius E., Gama J.** Bright–drift-aware demand predictions for taxi networks. *IEEE Transactions on Knowledge and Data Engineering*, 2020, vol. 32, pp. 234–245.
13. **Шелухин О. И., Симонян А. Г., Ванюшина А. В.** Влияние структуры обучающей выборки на эффективность классификации приложений трафика методами машинного обучения. *T-Comm: Телекоммуникации и транспорт*, 2017, т. 11, № 2, с. 25–31.
14. **Lughofer E., Weigl E., Heidl W., Eitzinger C., Radauer T.** Recognizing input space and target concept drifts in data streams with scarcely labeled and unlabelled instances. *Information Sciences*, 2016, vol. 35, pp. 127–151.
15. **Sethi T., Kantardzic M.** Handling adversarial concept drift in streaming data. *Expert Systems with Applications*, 2018, vol. 97, pp. 18–40.
16. **Рзаев Б. Т., Лебедев И. С.** Применение бэггинга при поиске аномалий сетевого трафика. *Научно-технический вестник информационных технологий, механики и оптики*, 2021, т. 21, № 2, с. 50–56. doi:10.17586/2226–1494-2021-21-2-50-56
17. **Maletzke A., dos Reis D., Cherman E., Batista G.** (2019). DyS: A frame work for mixture models in quantification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 1, pp. 4552–4560. doi:10.1609/aaai.v33i01.33014552
18. **Gomes H. M., Bifet A., Read J., Barddal J. P., Enembreck F., Pfharinger B., Holmes G.** Adaptive random forests for evolving data stream classification. *Machine Learning*, 2017, vol. 106, iss. 9–10, pp. 1469–1495. doi:10.1007/s10994-017-5642-8
19. **Бажаев Н. А., Лебедев И. С., Кривцова И. Е.** Анализ статистических данных мониторинга сетевой инфраструктуры для выявления аномального поведения локального сегмента системы. *Научно-технический вестник информационных технологий, механики и оптики*, 2017, т. 17. № 1, с. 92–99. doi:10.17586/2226-1494-2017-92-99
20. **Oikarinen E., Tiittanen H., Henelius A.** Detecting virtual concept drift of regressors without ground truth values. *Data Mining and Knowledge Discovery*, 2021, vol. 35, iss. 3, pp. 821–859. doi:10.1007/s10618-021-00739-7
21. **Maletzke A., dos Reis D., Batista G.** Combining instance selection and self-training to improve data stream quantification. *Journal of the Brazilian Computer Society*, 2018, vol. 24, no. 12, pp. 123–141. doi:10.1186/s13173-018-0076-0
22. PowerSupply dataset. http://www.cse.fau.edu/~xqzhu/stream.html (дата обращения: 27.03.2021).

UDC 621.396
doi:10.31799/1684-8853-2021-3-29-38

Dataset segmentation considering the information about impact factors

I. S. Lebedev[a], Dr. Sc., Tech., Professor, orcid.org/0000-0001-6753-2181, isl_box@mail.ru
[a]Saint-Petersburg Federal Research Center of the RAS, 39, 14 Line V. O., 199178, Saint-Petersburg, Russian Federation

Introduction: The application of machine learning methods involves the collection and processing of data which comes from the recording elements in the offline mode. Most models are trained on historical data and then used in forecasting, classification, search for influencing factors or impacts, and state analysis. In the long run, the data value ranges can change, affecting the quality of the classification algorithms and leading to the situation when the models should be constantly trained or readjusted taking into account the input data. **Purpose:** Development of a technique to improve the quality of machine learning algorithms in a dynamically changing and non-stationary environment where the data distribution can change over time. **Methods:** Splitting (segmentation) of multiple data based on the information about factors affecting the ranges of target variables. **Results:** A data segmentation technique has been proposed, based on taking into account the factors which affect the change in the data value ranges. Impact detection makes it possible to form samples based on the current and alleged situations. Using PowerSupply dataset as an example, the mass of data is split into subsets considering the effects of factors on the value ranges. The external factors and impacts are formalized based on production rules. The processing of the factors using the membership function (indicator function) is shown. The data sample is divided into a finite number of non-intersecting measurable subsets. Experimental values of the neural network loss function are shown for the proposed technique on the selected dataset. Qualitative indicators (Accuracy, AUC, F-measure) of the classification for various classifiers are presented. **Practical relevance:** The results can be used in the development of classification models of machine learning methods. The proposed technique can improve the classification quality in dynamically changing conditions of the functioning.

Keywords — machine learning, segmentation of multiple data, affecting factors, changing conditions.

For citation: Lebedev I. S. Dataset segmentation considering the information about impact factors. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 3, pp. 29–38 (In Russian). doi:10.31799/1684-8853-2021-3-29-38

References

1. Schlimmer J. C., Granger R. H. Incremental learning from noisy data. *Machine Learning*, 1986, no. 1, pp. 317–354. doi:10.1023/A:1022810614389
2. Widmer G., Kubat M. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 1996, no. 23(1), pp. 69–101. doi:10.1007/BF00116900
3. Gama J., Zliobait I., Bifet A., Pechenizkiy M., Bouchachia A. A survey on concept drift adaptation. *ACM Computing Surveys*, 2014, no. 46(4), pp. 1–37.
4. Sung-Yu T., Jen-Yuan C. Parametric study and design of deep learning on leveling system for smart manufacturing. *IEEE International Conference on Smart Manufacturing, Industrial & Logistics Engineering* (*SMILE*), February 8–9, 2018, pp. 48–52. doi:10.1109/SMILE.2018.8353980
5. Lu J., Liu A., Dong F., Gu F., Gama J., Zhang G. Learning under concept drift: a review. *IEEE Transactions on Knowledge and Data Engineering*, 2019, no. 31(12), pp. 2346–2363.
6. Wang L. Y., Park C., Yeon K., Choi H. Tracking concept drift using a constrained penalized regression combiner. *Computational Statistics & Data Analysis*, 2017, no. 108, pp. 52–69.
7. Khan S., Yairi T. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 2018, no. 107, pp. 241–265. doi:10.1016/j.ymssp.2017.11.024
8. Salehi H., Burgueño R. Emerging artificial intelligence methods in structural engineering. *Engineering Structures*, 2018, no. 171, pp. 170–189. doi:10.1016/j.engstruct.2018.05.084
9. Zhou Z.-H., Feng J. Deep forest. *National Science Review*, 2019, vol. 6, no. 1, pp. 74–86. doi:10.1093/nsr/nwy108
10. Sethi T., Kantardzic M. Handling adversarial concept drift in streaming data. *Expert Systems with Applications*, 2018, vol. 97, pp. 18–40. doi:10.1016/j.eswa.2017.12.022
11. Takacs A., Toledano-Ayala M., Dominguez-Gonzalez A., Pastrana-Palma A., Velazquez D. T., Ramos J. M., Rivas-Araiza A. E. Descriptor generation and optimization for a specific outdoor environment. *IEEE Access*, 2020, vol. 8, pp. 2169–3536. doi:10.1109/ACCESS.2020.2975474
12. Saadallah A., Moreira-Matias L., Sousa R., Khiari J., Jenelius E., Gama J. Bright–drift-aware demand predictions for taxi networks. *IEEE Transactions on Knowledge and Data Engineering*, 2020, vol. 32, pp. 234–245.
13. Sheluhin O. I., Simonyan A. G., Vanyushina A. V. Influence of training sample structure on traffic application efficiency classification using machine-learning methods. *T-Comm*, 2017, vol. 11, no. 2, pp. 25–31 (In Russian).
14. Lughofer E., Weigl E., Heidl W., Eitzinger C., Radauer T. Recognizing input space and target concept drifts in data streams with scarcely labeled and unlabelled instances. *Information Sciences*, 2016, vol. 35, pp. 127–151.
15. Sethi T., Kantardzic M. Handling adversarial concept drift in streaming data. *Expert Systems with Applications*, 2018, vol. 97, pp. 18–40.
16. Rzayev B. T., Lebedev I. S. Applying bagging in finding network traffic anomalies. *Scientific and TechnicalJournal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 2, pp. 50–56 (In Russian). doi:10.17586/2226–1494-2021-21-2-50-56
17. Maletzke A., dos Reis D., Cherman E., Batista G. (2019). DyS: A frame work for mixture models in quantification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 1, pp. 4552–4560. doi:10.1609/aaai.v33i01.33014552
18. Gomes H. M., Bifet A., Read J., Barddal J. P., Enembreck F., Pfharinger B., Holmes G. Adaptive random forests for evolving data stream classification. *Machine Learning*, 2017, vol. 106, iss. 9–10, pp. 1469–1495. doi:10.1007/s10994-017-5642-8
19. Bazhayev N. A., Lebedev I. S., Krivtsova I. E. Analysis of statistical data from network infrastructure monitoring to detect abnormal behavior of system local segments. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2017, vol. 17, no. 1, pp. 92–99 (In Russian). doi:10.17586/2226-1494-2017-17-1-92-99
20. Oikarinen E., Tiittanen H., Henelius A. Detecting virtual concept drift of regressors without ground truth values. *Data Mining and Knowledge Discovery*, 2021, vol. 35, iss. 3, pp. 821–859. doi:10.1007/s10618-021-00739-7
21. Maletzke A., dos Reis D., Batista G. Combining instance selection and self-training to improve data stream quantification. *Journal of the Brazilian Computer Society*, 2018, vol. 24, no. 12, pp. 123–141. doi:10.1186/s13173-018-0076-0
22. *PowerSupply dataset*. Available at: http://www.cse.fau.edu/~xqzhu/stream.html (accessed 27 March 2021).