Адаптивное обучение моделей для прогнозирования потоков данных в информационно-телекоммуникационных системах

Models adaptive training for forecasting data flows in information and telecommunication systems

Сухопаров / Sukhoparov M.
Михаил Евгеньевич
(mail@sukhoparovm.ru)
доктор технических наук.
ФГБОУ ВО «Российский государственный гидрометеорологический университет» (РГГМУ),
доцент кафедры информационных технологий и систем безопасности.
г. Санкт-Петербург

Лебедев / Lebedev I.
Илья Сергеевич
(isl_box@mail.ru)
доктор технических наук, профессор.
ФГБУН «Санкт-Петербургский Федеральный исследовательский центр РАН»,
заведующий лабораторией интеллектуальных систем.
г. Санкт-Петербург

Сикарев / Sikarev I.
Игорь Александрович
(sikarev@yandex.ru)
доктор технических наук, профессор.
РГГМУ, профессор кафедры информационных технологий и систем безопасности.
г. Санкт-Петербург

Ключевые слова: машинное обучение – machine learning; информационно-телекоммуникационные системы – information and telecommunication systems; потоки данных – data flows; воздействующие факторы – influencing factors; сетевой трафик – network traffic.

Предлагается метод адаптивного обучения моделей для обработки потоков данных в информационно-телекоммуникационных системах, использующий автоматическое сегментирование данных. Целью предлагаемого решения является повышение качественных показателей классифицирующих моделей за счет разделения выборок данных с учетом их свойств на подмножества. Основное отличие предлагаемого метода заключается в том, что используется разделение выборки на подвыборки. Приведено разбиение выборки данных на конечное число не пересекающихся измеримых подмножеств.

A method of adaptive training of models for processing data flows in information and telecommunication systems using automatic data segmentation is proposed. The purpose of the proposed solution is to improve the quality indicators of classifying models by dividing data samples, considering their properties, into subsets. The main difference of the proposed method is that the sample is divided into subsamples. The data sample is divided into a finite number of non-overlapping measurable subsets.

Введение

Решение задач классификации и прогнозирования состояний объектов сетевой инфраструктуры информационно-телекоммуникационных систем (ИТКС) связано с необходимостью решения ряда проблемных вопросов обработки многомерных информационных потоков. Это обусловлено тем, что устройства сетей подвержены внешним и внутренним воздействиям, вызывающим быстрые переходы из одного состояние в другое. Распределение анализируемых данных может изменяться с течением времени [1]. Возникает необходимость в создании методов адаптивного обучения и использования моделей, позволяющих идентифицировать состояние телекоммуникационных устройств и узлов в динамически меняющихся и нестационарных средах [2, 3].

Увеличение количества источников информации о протекающих процессах в определенных случаях позволяет повысить качественные показатели обрабатываемой информации. Кроме того, возможно вычисление дополнительных статистических характеристик временных последовательностей и потоков, которые позволяют в явном или неявном виде учитывать различные воздействующие факторы. Выявление

границ и точек «разладки» для временных рядов дают возможность осуществить анализ не всей выборки, а ее отдельных частей [4]. Определив свойства сегментов и обучив на них алгоритм, возможно осуществить выбор наиболее подходящих моделей для каждого сегмента с их последующим объединением.

Современные информационные потоки в ИТКС характеризуются высокой интенсивностью поступления информации, что накладывает на модель обработки потока ограничения по скорости вычисления требуемых значений. Необходимо анализировать, обрабатывать, прогнозировать на данных, имеющих бесконечные длины и поступающих в реальном масштабе времени. Предсказываемые моделью статистические свойства предикторов и целевых переменных информационного потока изменяются в различных временных интервалах непредсказуемым образом. Полученные ранее «исторические данные» устаревают, и их использование не позволяет достичь заданных качественных показателей при обучении. Возникает необходимость изменения и адаптации существующих моделей интеллектуального анализа данных и разработки новых алгоритмов, позволяющих производить адаптивное обучение. Необходимо осуществлять автоматическую обработку выборок данных для определения свойств отдельных сегментов и назначения наиболее подходящих классификационных моделей.

Постановка задачи

На сегодняшний день не существует универсальных алгоритмов, позволяющих достичь заданного качества обработки данных с различными свойствами. Определение оптимального алгоритма обучения для конкретной выборки является трудоемким, многоитерационным процессом. В связи с этим возникает необходимость реализации модели, объединяющей несколько алгоритмов, для улучшения качественных показателей принятых решений [5].

Особенности алгоритмов классификации, свойств информации оказывают влияние на эффективность применения различных классификаторов. В зависимости от требуемых показателей полноты, точности, скорости обработки выбираются алгоритмы, обладающие своими уникальными характеристиками. Для достижения требуемых показателей приходится проводить ряд процедур, направленных на настройку, обработку данных, связанных с удалением выбросов и шумов.

Уточнение локальной информации является одним из подходов к разработке, адаптации и использованию моделей, направленных на повышение качественных показателей процесса классификации состояния [6, 7].

Анализ обучающей выборки предполагает ее обработку как единой последовательности. Однако возможна сегментация данных на подмножества в момент воздействия различных внешних и внутренних факторов. Например, рост интенсивности появления служебных сообщений в сетях происходит из-за появления ряда управляющих команд. Некоторые факторы либо известны заранее, либо поддаются прогнозированию, что дает возможность учесть их воздействие на значения переменных. В связи с этим возможно произвести разделение выборки на подмножества и определить кортежи, полученные в момент различных воздействий на ИТКС [8, 9].

Рассмотрим X^q – выборку размером q. $\{a_1,a_2,...,a_r\} \in A$ – множество классифицирующих алгоритмов, $a(x)$ – результирующая модель, решающая задачу классификации. В момент формирования выборки воздействует множество известных факторов $\{v_1,v_2,...,v_m\} \in V$, влияющих на значения целевых переменных в кортежах $x \in X^q$. Для оценки качества классифицирующих алгоритмов $a_1,a_2,...,a_r$ используется функция потерь L, рассчитывающаяся как разница между фактическим выходным значением и прогнозируемым выходным значением. Определим $L(a,x,v)$ как функцию потерь в момент действия фактора v. Тогда с учетом v средняя величина ошибки определяется выражением:

$$Q(a,X^q,v)=\frac{1}{q}\sum_{x\in X} L(a,x,v). \quad (1)$$

Необходимо минимизировать функционал качества по параметру a:

$$Q(a,X^q,v)\to \min. \quad (2)$$

Такая постановка позволяет учитывать эффекты, влияющие на разброс, смещение ответов классификаторов, образуемых воздействием известных факторов [10].

Предлагаемый подход

В рамках предлагаемого подхода для обработки динамических информационных потоков предлагается агрегировать алгоритмы машинного обучения в ансамбль, настраиваемый на свойства данных. Структурная схема формирования ансамбля представлена на рис. 1 (дрейф концепта – эффект, при котором статистические свойства целевой переменной, которую пытается предсказать модель, меняются со временем непредвиденным образом; верификация модели классификации – процесс проверки истинности, адекватности и корректности работы модели).

Информационный поток в ИТКС, представленный последовательностью данных, поступает на предварительную обработку. Происходит оценка свойств данных и разбиение выборки на отдельные сегменты с совпадающими свойствами данных. В результате множество X^q с учетом действия множества факторов V разбивается на подмножества $X_1,X_2,...,X_m$, каждое из которых имеет свойства, отличные от других.

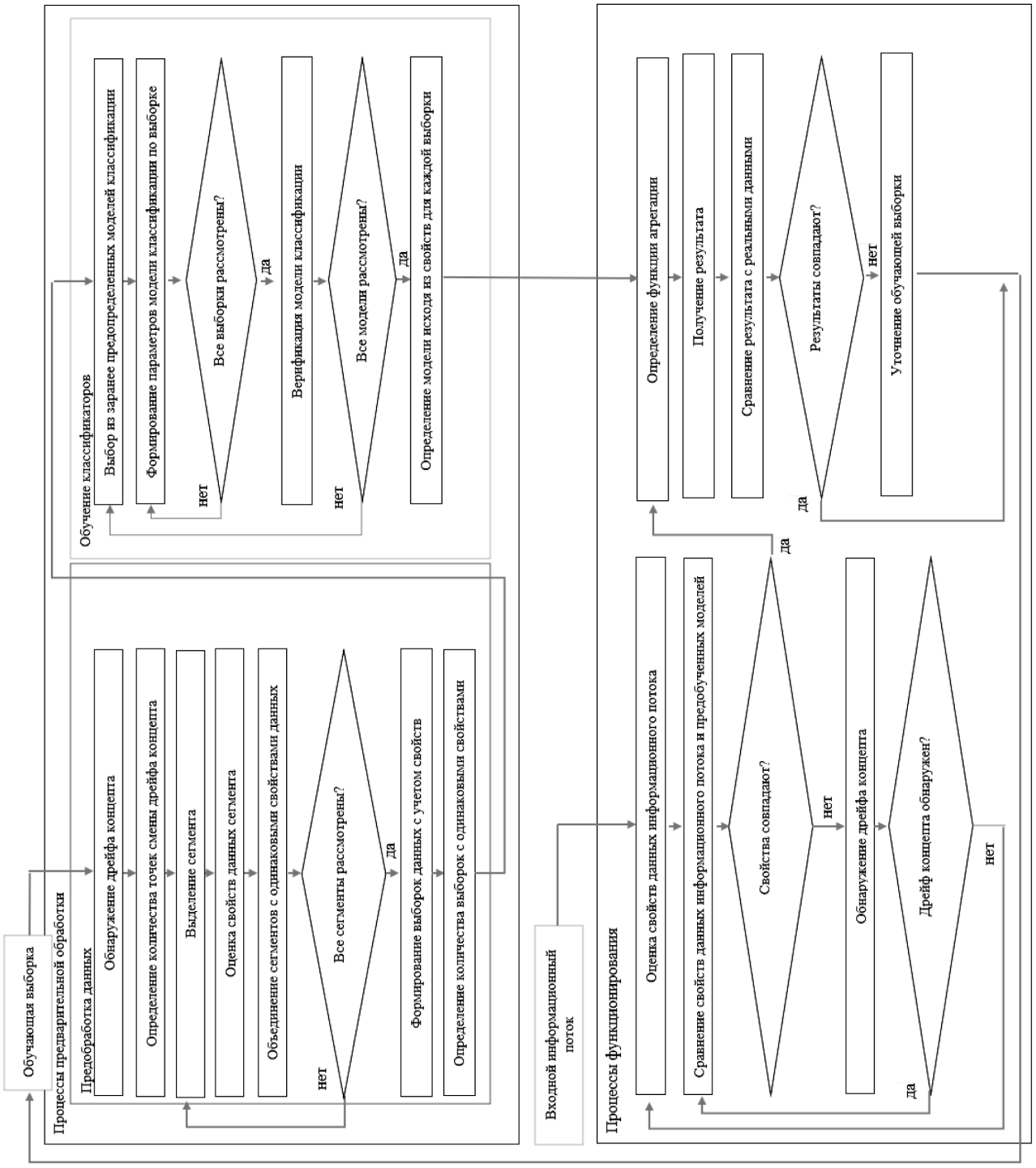

Рис. 1. Структурная схема формирования ансамбля классифицирующих алгоритмов, настраиваемого на свойства данных

Предобработка позволяет осуществить адаптивную настройку базовых классификаторов $\{a_1, a_2, ..., a_r\} \in A$ на каждом подмножестве $X_i \in X^q$. Происходит их обучение, направленное на формирование параметров и весовых матриц алгоритмов классификации, а затем проводится анализ их результатов. Выбор осуществляется исходя из минимизации эмпирического риска, направленного на поиск алгоритма, для которого при $x \in X^q$ соблюдается условие:

$$a(x) = \arg\min_{a_i \in A} Q(a_i(x), X_i). \quad (3)$$

Используя выражение (3) для каждой подвыборки, становится возможным выбирать алгоритм с лучшими качественными показателями. Простая агрегирующая функция, вычисляющая лучшие алгоритмы на подвыборках $X_1, X_2, ..., X_m$ для выражения (3), примет вид:

$$a(x) = F(\arg\min_{a_i \in A}(Q(a_1(x), X_1), ..., a_r(x), X_1)), ...,$$
$$\arg\min_{a_i \in A}(Q(a_1(x), X_m), ..., a_r(x), X_m))). \quad (4)$$

В зависимости от решаемой задачи, заданных качественных показателей, а также от свойств данных и особенностей обучающих подмножеств, возможно формирование более сложных агрегирующих функций, учитывающих веса классификаторов и использующих аддитивные коэффициенты, изменяющие «важность голоса» алгоритма.

Полученный результат обработки в дальнейшем поступает в обучающую выборку и учитывается алгоритмом предобработки для уточнения модели.

Предложенный многоуровневый подход направлен на осуществление оценки возможных алгоритмов на предварительном этапе, и последующий выбор алгоритма и его агрегации с другими.

В ходе реализации сложных моделей машинного обучения возникает ряд проблемных вопросов, связанных с эффективностью применения отдельных ее компонент. Каждый базовый алгоритм обладает различными качественными показателями на данных с разными свойствами. В реальных системах с течением времени может происходить изменение частоты наблюдаемых событий, смещение диапазона значений, появление дисбаланса в совокупности данных. В связи с этим необходима разработка эффективных методов предобработки информации, направленных на вычисление и анализ свойств, поступающих на вход анализатора. Они должны выполнять не только обычные задачи фильтрации, удаления шумов и выбросов, но и вычислять свойства данных, формировать их сегменты. Комплекс таких методов должен позволять осуществлять выбор и определение наиболее подходящих моделей для задач классификации и регрессии.

Применение методов основывается на предварительной обработке, где в первоначальной выборке происходит разграничение отдельных сегментов с данными, обладающими сходными свойствами. При автоматическом способе разделения вычисляются точки, в которых происходит изменение направления тренда, либо анализируются ситуации смены концепции.

Рассмотрим задачу машинного обучения, определенную в компактном пространстве X и меток C. Определим функционал качества Q, который позволяет оценить результат прогноза.

При обучении в рассматриваемом случае имеется дополнительная информация для ряда объектов наблюдений из генеральной совокупности, определенной в пространстве X', которая доступна из обучающей выборки. То есть имеется доступ к объектам выборки, где распределение данных представляется в виде $x_i, c_i \sim p(x|c)$.

Учитывая класс функций $a(x, w)$, параметризованный весовыми коэффициентами w, необходимо найти решение следующей задачи оптимизации:

$$\min_w E_{x,c \boxtimes p(x|c)}[Q(c, a(x, w))]. \quad (5)$$

На практике решение обозначенной задачи может быть осуществлено следующим образом. Первоначальная обучающая выборка обрабатывается классифицирующим алгоритмом, в сопряжении с которым работает метод детектора дрейфа концепта. Задача такой системы состоит в том, чтобы обнаружить ситуации, где происходит изменение плотности вероятности в анализируемых данных $p_t(x|c) \neq p_{t+1}(x|c)$. Задание порогов плотностей позволяет скомпоновать «близкие» области, свойства которых практически одинаковы. В результате исходная выборка делится на несколько частей с плотностями вероятностей $p_1(x|c), p_2(x|c), ..., p_n(x|c)$. Эти выборки поступают на вход классификационным алгоритмам $a_1, a_2, ..., a_r$, где возможно их обучение и последующий выбор для каждого обучающего подмножества с учетом условия $Q(a_i, X_i) \to \min$

После этого первоначальная классификация, позволяющая выбрать модели обработки данных, заканчивается.

Функционирование системы в динамике предполагает наличие поступающего на вход информационного потока данных x'. Он принимается детектором потока, в котором задается ширина окна и происходит сравнение временных рядов $x'_{t_1}, ..., x'_{t_1+T}$ и $x_{t_1}, ..., x_{t_1+T}$, осуществляется оценка их свойств $p_{X'}(x'|c)$ и $p_X(x|c)$. Затем выбирается наиболее подходящая классификационная модель, обученная на подмножестве, свойства которого максимально совпадают со свойствами информационного потока данных.

На последнем этапе модель используется для решения задач, полученные результаты обработки

сравниваются с имеющимися, производится анализ результатов и формируются данные для выборки, которая может быть использована для уточнения процесса предварительного обучения.

В реальных системах возникают ситуации, когда появляются случайные факторы, изменяющие ее состояния. Часть из таких факторов возможно определить заранее, другая часть возникает случайно и не поддается прогнозированию. Однако в любом случае цена ошибки может быть очень высока.

В связи с этим необходимо прогнозирование и проактивное управление, направленное на оценку развития ситуации в динамике. Проактивное управление ИТКС позволяет осуществлять оценку состояния системы в динамике на несколько дискретов времени вперед, подразумевает оперативное реагирование и последующее недопущение возможных нештатных и аварийных ситуаций, предполагает предотвращение возникновения указанных ситуаций за счет создания в соответствующей системе управления принципиально новых прогнозирующих и упреждающих возможностей при формировании и реализации управляющих воздействий. Разбиение временных рядов на *n* отдельных сегментов позволяет определить их свойства $p_{X_i}(x^{'} \mid c)$, например плотности вероятностей появления событий.

Определяя интервальные допуски ε возможно сопоставление и принятие решения о совпадении анализируемых сегментов $p_{X_i}(x^{'} \mid c) <> \varepsilon$. В результате можно выявить их последовательности, порядок следования друг за другом.

Обработка последовательностей дает возможность строить графы и матрицы переходов. Анализ переходов состояний делает возможным в текущий дискретный момент времени определять наиболее вероятные переходы из текущего состояния в последующие.

Для повышения вероятностей переходов возможно применение трендов полученных при анализе сегментов, описанных временными рядами.

Применение метода

Предлагаемый подход был реализован на актуальных наборах данных (CIC IoT Dataset 2022, Bot-IoT Dataset, TON_IoT Datasets, NSL-KDD), полученных из открытых источников. Анализируемые наборы данных представляют собой нормальный и аномальных траффик, содержащий различные сценарии кибератак на промышленные ИТКС, информацию о процессах функционирования IoT устройств и сенсоров в штатных и нештатных условиях. Анализу перечисленных наборов данных посвящен ряд актуальных работ различных зарубежных и отечественных исследователей, что позволяет сравнить полученные в рамках исследования результаты. Алгоритмы обработки тестовых данных были реализованы на языке Python с использованием библиотек pandas, numpy, sklearn. Для оценки изменений показателей качества было осуществлено два разбиения последовательности сетевого трафика: на основе индикаторной функции и на основе метода определения точек разладки. В дальнейшем для анализа алгоритмов классификации использовались выборки, полученные двумя способами разбиения.

С течением времени под воздействием внутренних и внешних факторов происходит изменение диапазонов целевой переменной. В рассматриваемых случаях заранее известный фактор влияет на значения входящих в кортеж данных.

Оценка статистических свойств подмножеств на качество результатов моделей машинного обучения производилась с помощью различных алгоритмов: линейного дискриминантного анализа (LD), квадратичного дискриминантного анализа (QD), наивного байесовского классификатора (NB), ближайших соседей (KNN).

Ниже в таблице 1 приводятся результаты тестирования классификаторов (AUC – площадь под ROC-кривой, Accuracy – точность, F-мера – комплексный показатель эффективности). Сегментация была произведена по функции принадлежности и с помощью метода поиска точек разладки в автоматическом режиме осуществляет обнаружения момента, в котором происходит изменение направления тренда наблюдаемого процесса.

В анализируемых подвыборках наблюдался рост показателей качества классификации по сравнению с классификацией всего обучающего множества.

Результаты подтверждают, что сегментирование общей выборки, в основном, позволяет повысить определенные качественные показатели отдельных алгоритмов. Однако иногда возникают ситуации, ухудшающие результаты. Возникновение большого разброса, сложность построения разделяющей поверхности, могут негативно влиять на результаты отдельных классификаторов.

Над набором данных возможно осуществить дальнейшее разбиение на подмножества, повышая порог, идентифицирующий смену концепции или, для эвристических методов, используя информацию о выходных и праздничных днях, условиях и т.д. Решение можно применять как дополнение к различным моделям.

Предложенный метод подготовки данных можно применить для улучшения качественных показатели регрессии. Для оценки влияния подмножеств на качество результатов моделей машинного обучения были выбраны различные алгоритмы: линейной регрессии (LD), гауссовой регрессии (GR), деревьев решений (DT), метода опорных векторов (SVM).

Рассматривались задача предсказания интенсивности сетевого трафика, при условии известных значений рабочего или нерабочего времени. Значения функции потерь (RMSE (Root Mean Square Error) – среднеквадратическая ошибка) для различных сегментов представлено в таблице 2.

Таблица 1

Результаты классифицирующих алгоритмов

		Вся выборка	Сегментации по функции принадлежности				среднее	Сегментация по методу поиска точек разладки				среднее
			1	2	3	4		1'	2'	3'	4'	
LD	F-мера	0,81	0,84	0,81	0,82	0,81	0,82	0,83	0,82	0,82	0,8	0,82
	Accuracy %	72,3	78,9	73,1	74	72,8	74,70	76	72,5	73,5	72,3	73,58
	AUC	0,74	0,79	0,73	0,75	0,76	0,76	0,73	0,74	0,74	0,76	0,74
QD	F-мера	0,81	0,84	0,81	0,82	0,81	0,82	0,83	0,82	0,81	0,81	0,82
	Accuracy %	72,3	78,2	72,2	74,2	73,4	74,50	75,3	72,7	73,3	73	73,58
	AUC	0,74	0,79	0,73	0,75	0,76	0,76	0,74	0,75	0,74	0,77	0,75
KNN	F-мера	0,79	0,81	0,51	0,8	0,78	0,73	0,8	0,62	0,79	0,78	0,75
	Accuracy %	70,5	77,6	72,4	72,9	72,3	73,80	75,7	71	71,6	72,1	72,60
	AUC	0,71	0,79	0,72	0,74	0,76	0,75	0,76	0,72	0,71	0,76	0,74
NB	F-мера	0,79	0,84	0,81	0,81	0,81	0,82	0,83	0,82	0,8	0,8	0,81
	Accuracy %	71,4	76,1	73,3	73,6	72,6	73,90	73,5	73,4	71,5	72,2	72,65
	AUC	0,74	0,78	0,73	0,75	0,76	0,76	0,73	0,74	0,74	0,76	0,74

Таблица 2

Результаты функции потерь RMSE для разных классификаторов

RMSE	Вся выборка	Сегментации по функции принадлежности				Среднее	Сегментация по методу поиска точек разладки				Среднее
		1	2	3	4		1	2	3	4	
LR	6,36	6,55	5,65	5,33	4,60	5,53	7,25	4,93	5,08	4,52	5,45
GR	6,03	6,34	5,60	5,31	4,59	5,46	7,03	4,90	5,06	4,49	5,37
DT	6,37	7,10	6,10	5,77	5,01	5,99	7,83	5,40	5,55	4,97	5,94
SVM	5,90	6,56	5,65	5,33	4,61	5,54	7,26	4,93	5,08	4,53	5,45

В случае применения сегментации в среднем наблюдается уменьшение значений функции потерь по сравнению с полной не сегментированной выборкой.

Заключение

Появление различных статистических эффектов в данных, наличие процессов, приводящих к трансформации свойств, увеличивает временные затраты на обучение и приводит к необходимости реализации сложных моделей. В то же время существует достаточно большое количество воздействующих на данные факторов, которое можно определить и вычислить заранее. Такой анализ можно осуществить на основе эвристических или автоматических методов.

Предлагаемое решение направлено на разбиение данных на подмножества, что в ряде случаев позволяет получить однородные свойства данных и назначить для них алгоритм классификации. В текущий момент времени возможно применение классифицирующего алгоритма на связанных с ним подмножествах данных.

Однако в реальных системах учитывать воздействие факторов очень сложно. Существуют как явные, так и скрытые факторы. В связи с этим возникает необходимость их выявления. Предложенный метод позволяет выявлять факторы автоматически на основе обнаружения точек смены концепта во временных рядах и сегментирования выборок данных. Оценивая свойства полученных сегментов выборок, становится возможным подобрать наиболее эффективные алгоритмы и классифицирующие модели.

Различные свойства объектов наблюдения, вызванные влиянием внешних и внутренних факторов, дают возможность строить разделяющую поверхность разными алгоритмами классификации. Это позволяет назначать и определять разные области компетенции для различных алгоритмов.

Предложенный метод представляет собой дополнение к различным методам классификации. Его эффективное применение в различных условиях зависит от свойств данных в отдельных сегментах, которые должны соответствовать классифицирующим алгоритмам. В дальнейшем на основе предложенного решения предполагается реализовывать двухуровневую модель для решения задач классификации и регрессии. На нижнем уровне будут применяться модели, обрабатывающие разделяемые подвыборки поступающих информационных последовательностей, а на верхнем – модели, назначающие предобученные классификаторы, на текущие сегменты. Такой подход предполагается рассмотреть, как усовершенствование ансамблиевых мотодов.

Литература

1. Descriptor generation and optimization for a specific outdoor environment / A. Takacs, M. Toledano-Ayala, A. Dominguez-Gonzalez [et al.] // IEEE Access. – 2020. – Vol. 8. – P. 2169–2182.

2. Лебедев, И. С. Сегментирование множества данных с учетом информации воздействующих факторов / И.С. Лебедев // Информационно-управляющие системы. – 2021. – № 3 (112). – С. 29–38.

3. Zhou, Z.-H. Deep Forest / Z.-H. Zhou, J. Feng // National Science Review. – 2019. – Vol. 6, No. 1. – P. 74–86.

4. Лебедев, И. С. Адаптивное применение моделей машинного обучения на отдельных сегментах выборки в задачах регрессии и классификации / И.С. Лебедев // Информационно-управляющие системы. – 2022. – № 3 (118). – С. 20–30.

5. kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data / J. Maillo, S. Ramrez, I. Triguero, F. Herrera // Knowledge-Based Systems. – 2017. – Vol. 117. – P. 3–15.

6. Maletzke, A. Combining instance selection and self-training to improve data stream quantification / A. Maletzke, D. dos Reis, G. Batista // Journal of the Brazilian Computer Society. – 2018. – Vol. 24, No. 12. – P.123–141.

7. Wu, Z. Improving generalization via scalable neighborhood component analysis / Z. Wu, A. Efros, S. Yu // Proceedings of the European Conference on Computer Vision (ECCV). – 2018. – P. 685–701.

8. The analysis of abnormal behavior of the system local segment on the basis of statistical data obtained from the network infrastructure monitoring / I. Lebedev, I. Krivtsova, V. Korzhuk [et al.] // Lecture Notes in Computer Science. – 2016. – No. 9870. – P. 503–511.

9. Meiseles, A. Source Model Selection for Deep Learning in the Time Series Domain / A. Meiseles, L. Rokach // IEEE Access. – 2020. – Vol. 8. – P. 6190–6200.

10. Дьяконов, А. Методы решения задач классификации с категориальными признаками / А. Дьяконов // Прикладная математика и информатика. Труды факультета Вычислительной математики и кибернетики МГУ имени М.В. Ломоносова. – 2014. – № 46. – С. 103–127.