УДК 621.396 **DOI** 10.15622/ia.22.3.1

И.С. Лебедев

ПРИМЕНЕНИЕ МНОГОУРОВНЕВЫХ МОДЕЛЕЙ В ЗАДАЧАХ КЛАССИФИКАЦИИ И РЕГРЕССИОННОГО АНАЛИЗА

Лебедев И.С. **Применение многоуровневых моделей в задачах классификации и регрессионного анализа.**

Аннотация. Применение моделей машинного обучения обуславливает необходимость создания методов, направленных на повышение качественных показателей обработки информации. В большинстве практических случаев диапазоны значений целевых переменных и предикторов формируются под воздействием внешних и внутренних факторов. Такие явления, как дрейф концепций, приводят к тому, что модель со временем понижает показатели полноты и точности результатов. Целью работы является повышение качества анализа выборок и информационных последовательностей на основе многоуровневых моделей для задач классификации и регрессии. Предлагается двухуровневая архитектура обработки данных. На нижнем уровне происходит анализ поступающих на вход информационных потоков и последовательностей, осуществляется решение задач классификации или регрессии. На верхнем уровне выполняется разделения выборок на сегменты, определяются текущие свойства данных в подвыборках и назначаются наиболее подходящие по достигаемым качественным показателям модели нижнего уровня. Приведено формальное описание двухуровневой архитектуры. В целях повышения показателей качества решения задач классификации и регрессии производится предварительная обработка выборки данных, вычисляются качественные показатели моделей, определяются классификаторы, имеющие лучшие результаты. Предложенное решение позволяет реализовывать постоянно обучающиеся системы обработки данных. Оно направлено на снижение затрат на переобучение моделей в случае трансформации свойств данных. Проведены экспериментальные исследования на ряде наборов данных. Численные эксперименты показали, что предложенное решение позволяет повысить качественные показатели обработки. Модель может быть рассмотрена как совершенствование ансамблевых методов обработки информационных потоков и выборок данных. Обучение отдельного классификатора, а не группы сложных классификационных моделей дает возможность уменьшить вычислительные затраты.

Ключевые слова: машинное обучение, многоуровневые модели, назначение классифицирующих алгоритмов.

1. Введение. В настоящее время применение методов искусственного интеллекта в отдельных задачах направлено на повышение достигаемых качественных показателей обработки информации. Алгоритмы машинного обучения дают возможность выявлять характеристики, статистические свойства, неявные знания, необходимые для достижения заданного результата, путем систематического анализа достаточного количества соответствующих выборок данных.

Алгоритмы машинного обучения требуют предварительного извлечения значений характеристик объектов наблюдения для представления входных последовательностей данных. Показатели

качества алгоритмов машинного обучения в значительной степени зависят от характеристик анализируемых выборок. Крайне важно выбрать правильную группу признаков, оптимально представляющих наиболее значимые свойства входных данных [1]. После этого модель позволяет получить сопоставление между извлеченными характеристиками объектов и желаемым результатом.

Тем не менее, для многих задач выявление основных характеристик и признаков данных, которые позволят достичь лучших качественных показателей, является сложным и трудоемким процессом.

Одновременно с этим на практике при обработке информационных потоков возникают явления, связанные с дрейфом концепции, когда происходит смещение диапазонов целевых переменных и изменения распределений данных. Все это приводит к тому, что с течением времени любая модель может ухудшить свои качественные показатели обработки. В статье предложено одно из возможных решений повышения качественных показателей, где рассматривается использование многоуровневых моделей для решения задач классификации и регрессии. На нижнем уровне происходит обработка поступающих на вход информационных потоков и последовательностей, а на верхнем – решаются задачи разделения потоков на части, определения текущих свойств данных в сегментах и, исходя из их свойств, назначения наиболее подходящей по достигаемым качественным показателям моделей нижнего уровня.

2. Современные методы обработки выборок. Большинство методов машинного обучения используют «централизованные данные», где в выборках хранится вся информация по объектам наблюдения. Процессы сбора осуществляются в течение определенного периода времени и, обычно, содержат кортежи значений, когда наблюдаемая система находится в разных состояниях и на нее воздействует множество разнородных факторов. В результате возникают явления, связанные с трансформацией свойств, смещения диапазонов значений, полученных от регистрирующих элементов. Это приводит к возникновению неоднородности данных в выборках. В отдельных последовательностях внутри выборки может появляться дисбаланс классов, происходить смена распределений, изменение вероятностей возникновения событий и частоты появления объектов наблюдения.

Решение задач классификации и регрессии методами машинного обучения может быть затруднено при возникновении различных статистических эффектов. Например, если в данных

наблюдается парадокс Симпсона [2], стандартный подход к централизованному интеллектуальному анализу выборки может не позволять достигать заданных качественных показателей обработки данных, а результат обработки не соответствовать истинному состоянию.

Современные подходы к построению моделей обработки связаны с формированием, анализом и объединением локальных результатов, при котором используются методы агрегирования.

Методы и алгоритмы, решающие задачи классификации и регрессии, могут иметь различные результаты выбранного показателя качества на одном и том же наборе данных, например, в случае разных обучающих примеров. Полученные при обработке объектов наблюдения результаты различных классификаторов могут отличаться. Их ответы объединяются агрегирующей функции для формирования общего ответа. За счет интеграции нескольких моделей в ряде случаев становится возможным повысить качество классификации.

На сегодняшний день доминирующими являются ансамблевые методы. Среди них наиболее известны подходы на основе простого, комбинированного голосования [3, 4], а так же применение ряда агрегирующих функций, вычисляющих максимальные, средние, медианные и другие вероятности классов, усредняющих результат предсказания по совокупности ответов.

В качестве альтернативы используются различные агрегаторы, основанные на рейтингах классифицирующих алгоритмов, арбитры, комбинаторы [4, 5], которые могут быть применены как к бинарным, так и к многоклассовым задачам.

Еще одно направление связано с формированием выборок. В [3, 5 – 9] были исследованы различные аспекты вертикально разделенных данных, предложены технологии, базовые алгоритмы и комбинированные стратегии, направленные на выбор объектов наблюдения, позволяющих получить основные характеристики последовательностей и выборок, исключить из рассмотрения значения, приводящие к искажению свойств данных [10 – 12].

В последние годы в ряде задач обработки информации применяются гибридные классификаторы. Используются комбинации методов, где различные модели на основе относительно простых классифицирующих алгоритмов и сложных нейронных сетей достигают высоких показателей полноты и точности [13, 14]. Однако возможности негибридных моделей зависят от свойств обучающей выборки и, в случае изменения характеристик данных, качественные

показатели могут существенно снизиться [15 – 19]. Точность, полнота результатов обработки зависит от многих факторов.

Применение подобных подходов часто приводит к возникновению различных ситуаций, когда агрегация разных моделей не только не способствует повышению качественных показателей, а наоборот ухудшает результаты [20, 21]. Причем подобные эффекты часто нивелируются на большой выборке данных, но явно прослеживаются на ее отдельных сегментах. Это приводит к тому, что при обработке потоков данных возможны ошибки вследствие разных настроек классифицирующих моделей.

Таким образом, необходимо разрабатывать новые и адаптировать существующие стратегии, дающие возможность проводить точное и надежное обучение в рамках разделения функций и выборок.

3. Построение многоуровневой модели обработки данных. Практически все предлагаемые подходы, методы, алгоритмы машинного обучения на сегодняшний день являются узко специализированными. Каждая модель достигает определенных качественных показателей для тех предметных областей, где она оптимизировалась, и на данных которых она обучалась.

Одна из основных проблем достижения качественных показателей в методах машинного обучения связана с тем, что при изменении свойств поступающих данных, возникает необходимость в дополнительном обучении. Большинство моделей, решающие задачи классификации и регрессии, обучаются на заранее определенной совокупности объектов наблюдения. В случае появления трансформации свойств информационных последовательностей качество обработки снижается.

В статье предлагается решение, направленное на дальнейшее усовершенствование и расширение ансамблевых методов, где в зависимости от свойств данных происходит выбор модели с лучшими качественными показателями, а также формирование многоуровневых структур, осуществляющих анализ поступающих информационных потоков и назначение наиболее подходящей модели для обработки последовательностей в рамках решения текущей задачи.

В основе предлагаемого решения лежит возможность построения иерархий, когда модель верхнего уровня применяется для назначения наиболее эффективной модели нижнего уровня на отдельный сегмент выборки.

Для этого на различных уровнях иерархии разделяются роли и выполняемые функции моделей.

На нижних уровнях располагаются классификаторы и алгоритмы, реализации которых предназначены для решения узкоспециализированных задач.

На верхнем уровне применяются модели, решающие задачи управления алгоритмами нижнего уровня. Они "оценивают" поступающие информационные последовательности, определяют их свойства, сегментируют данные и назначают наиболее подходящую для решения текущей задачи модель нижнего уровня. В дальнейшем на верхнем уровне осуществляется мониторинг ошибок и отклонений результатов от реальных значений, организуется непрерывное обучение моделей нижнего уровня.

Применение иерархии направлено на обеспечение параллельного функционирования алгоритмов. Иерархическая система координирует действия для достижения заданных качественных показателей обработки данных, производит анализ текущих свойств данных, используемых для назначения эффективных алгоритмов.

В основе реализации многоуровневых моделей для классификации и регрессии лежит разделение их функций обработки.

В целях управления классификаторами, алгоритмами, работающими непосредственно с данными, формируются модели верхнего уровня. Они, в зависимости от свойств данных и требуемых качественных показателей, назначают наиболее подходящую модель нижнего уровня. В их функционал входят решение ряда перечисленных ниже задач:

– идентификация и сегментирование поступающей на вход последовательности;

– формирование пула предобученных моделей для ее обработки;

– выбор показателей и назначения наиболее подходящей модели для обработки текущей последовательности;

– анализ результатов обработки;

– принятие решения о формировании новой выборки и дополнительного обучения моделей.

Для их решения моделями верхнего уровня осуществляется управление и запуск стандартных алгоритмов кластеризации, поиска точек разладки для разделения поступающей последовательности, выполнение процессов обучения заранее предопределенного пула моделей нижнего уровня, определение их свойств и назначения на сегменты тех, которые имеют лучшие качественные показатели обработки. Они также отслеживают свойства поступающих данных и,

в случае обнаружения изменений и ухудшения результатов обработки, принимают решение о запуске процедур обучения.

Функции классификаторов и обрабатывающих алгоритмов нижнего уровня направлены на решение заранее определенных задач обработки и вычисления информации.

Для их обучения на вход подаются сформированные моделями верхнего уровня подвыборки, где свойства объектов наблюдения различаются. В процессе обучения определяются модели, достигающие лучших показателей качества на сегментах.

В результате в зависимости от свойств выборки данных определяются модели обработки, которые используются при появлении близких по свойствам сегментов.

Схема иерархии представлена на рисунке 1.

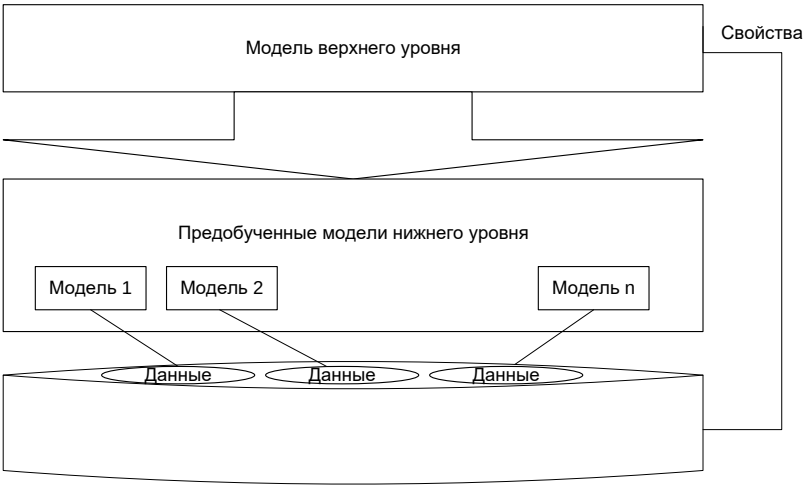

Рис. 1. Иерархия моделей

Использование нескольких моделей для повышения качества результатов предсказания известно как ансамблевые методы. Несмотря на различные комбинации, объединяющие отдельные алгоритмы в модель, возникают разные ситуации, где такое объединение не только может не повысить, но и даже ухудшить результат. В связи с этим необходимо предотвращать подобные ситуации.

В отличие от классических подходов, когда методы ансамбля [20] пытаются объединить несколько моделей с определенными стратегиями для улучшения качественных показателей классификации, в предложенном решении делается выбор наиболее

подходящей модели для текущих свойств данных. Это позволяет избегать различных ситуаций, когда слабые классификаторы могут ухудшить качественные показатели более сильной модели обработки данных.

Для этого рассматривается применение двухуровневых моделей, где на нижнем уровне идет обработка последовательности, а на верхнем – происходят процедуры оценивания данных и назначения алгоритмов.

4. Формальное описание модели. Финальную обработку данных для решения задач выполняют модели нижнего уровня. Они обрабатывают множество объектов наблюдения $\{x_1,...,x_z\} \in X$, заданных в метрическом пространстве, где каждому объекту x_i ставится в соответствие метка $y_i \in Y$, определенная на множестве меток.

Однако в различных локальных областях множества объектов наблюдения возможно изменение свойств (частоты появления, размаха данных, диапазонов предикторов и целевых переменных, распределений). В связи с этим, возникает необходимость реализации функции разделения множества $\mu: X \to \{1,...,m\}$, которая любому объекту $x_i \in X$ ставит в соответствие одно из подмножеств $X^1,...,X^m$.

В результате множество X сегментируется на подмножества $X^1 \cup X^2 \cup ... \cup X^m = X$ и $X^i \cap X^j \neq 0 \quad \forall i \neq j$. Объединение подмножеств совпадает с множеством X, подмножества не пересекаются.

Объекты подмножеств используются для вычисления центроидов $\hat{x}_c$ и определения метрики близости $\rho(x_i, \hat{x}_c)$, например, на основе евклидового расстояния. При выборе функции разделения множества необходимо добиваться, чтобы после обработки выборки подмножества состояли из объектов близких по $\rho(x_i, \hat{x}_c)$, а сами множества отличались.

После применения функции μ обучающая выборка разделяется на подмножества и представляется парами значений $(x_j^i, y_j^i)_{j=1}^{n_i} \in X^i$, где $\{X^1,...,X^m\} \in X$ – множество сегментов выборки данных X, X^i – множество описаний объектов наблюдения в сегменте, n_i – количество объектов в i-ом сегменте.

Для каждого сегмента X^i из заранее предопределенного множества моделей классификации $\{a_1, a_2, ..., a_n\} \in A$ определяется

решающая функция $a_i : X^i \to Y$, имеющая наилучшие выбранные качественные показатели для сегмента.

Метод построения решающих функций на подмножествах X^i ставит в соответствие сегменту X^i алгоритм обработки, принадлежащий заданному классу решающих функций. Он основывается на том, что каждая функция характеризуется своими показателями качества.

Пусть $L(y_j, a_k(x_j))$ – функция потерь. Тогда показатель качества для сегмента X^i определяется выражением:

$$Q(a_k(x), X^i) = \frac{1}{n_i} \sum_{j=1}^{n_i} L(y_j, a_k(x_j)) . \quad (1)$$

Используя (1), обозначим $q_0^i = \min_{a_k \in A} Q(a_k(x), X^i)$ полученное оптимальное значение, выбранное из всех значений функционала качества для моделей $a_k(x)$ на сегменте X^i.

Тогда для множества X, разделенного на ряд сегментов $\{X^1, ..., X^m\} \in X$, возможно определить матрицу R_X, показывающую лучшие значения показателей качества для каждого сегмента.

	X^1	...	X^m
a_1	q_1^1	...	q_1^m
...	...	...	...
a_n	q_n^1	...	q_n^m

. (2)

Анализируя значения (2) можно осуществить выбор лучшей по определенному показателю качества модели для сегмента X^i:

$$a_k(x) = \arg\min_{a_k \in A} Q(a_k(x), X^i) . \quad (3)$$

Выражение (3) дает возможность определить модели для сегментов.

В дальнейшем, рассматривая в динамике изменяющиеся значения матрицы R_X, становится возможным в автоматическом режиме определять модели для сегментов.

Качество результатов решения любой задачи методами машинного обучения зависит от свойств данных. Модели верхнего уровня предназначены для отслеживания изменений в информационных последовательностях вследствие воздействия внешних и внутренних факторов. Для этого они запускают процессы, разбиения на отдельные сегменты, выполняют алгоритмы определения точки разладки для временных рядов, определяют изменения свойств, а так же производят анализ и осуществляют назначение алгоритмов, моделей, обладающих лучшими характеристиками для выбранных сегментов.

В общем случае решение задач классификации и регрессии осуществляется на многомерных данных информационных последовательностей, из которых формируются выборки. Все информационные последовательности подвергаются процедурам обработки и сегментируются. На каждом из сегментов определяются качественные показатели таким образом, что выборке X ставится в соответствие матрица R_X, которая дает возможность назначить модели верхнего уровня наиболее подходящую модель.

В матрице R_X, содержится информация о значениях качественных показателей на выборке X каждой модели $a_1, a_2, ..., a_n$, что дает возможность назначать на сегменты выборки X моделей $\{a_1, a_2, ..., a_n\} \in A$, исходя из значений их показателей.

5. Реализация модели. Построение системы, использующей предложенную модель, происходит в несколько этапов.

Вначале выполняется настройка моделей нижнего уровня на решение требуемых задач обработки информации. Для них (рисунок 2) происходит первоначальное обучение. Определяется предварительная информация о совокупности объектов наблюдения $\{x_1, ..., x_z\} \in X$ и формируется первоначальное обучающее множество.

В выборке X осуществляется сегментирование областей. Формирование сегментов может происходить как с помощью заданной заранее системы правил, так и в автоматическом режиме с помощью алгоритмов, осуществляющих обнаружение точек, где изменяются свойства последовательностей. В этих целях могут быть применены методы поиска точек разладки, кластеризации, обнаружения и идентификации смены концепции и т.д. Выбор способа разделения последовательности происходит с учетом требований по обработке. В результате их применения определяются сегменты с различными свойствами. В дальнейшем все действия алгоритмов машинного

обучения происходят с отдельными областями, выделенными на предварительном этапе.

Обработка последовательности, в которой происходят периодические изменения, может дать возможность сформировать эвристические правила, позволяющие выделить сегменты, кластеры, имеющие разные свойства и диапазоны значений. Такой подход очень часто определяет довольно эффективную систему правил, направленную на достижение заданных показателей качества. Однако в дальнейшем приходится иметь дело со статичной и сложной в настройке системой [22].

В случае автоматического разделения последовательности объектов наблюдения обнаруживается момент θ, где происходит изменение характеристик наблюдаемого процесса [23]:

$$x_t^i = \begin{cases} x_t^i, 0 < t < \theta_i \\ x_t^{i+1}, t \geq \theta_i \end{cases}. \quad (4)$$

Исходная выборка в моменты изменения характеристик наблюдаемого процесса (4) делится на несколько частей $X^1,...,X^m$. Их свойства анализируются, совпадающие по свойствам сегменты объединяются, им присваиваются одинаковые индексы. В дальнейшем происходит обучение заранее определенных моделей $a_1, a_2, ..., a_n$ на подвыборках $X^1,...,X^m$, и анализируется достигаемый каждой моделью функционал качества $Q(a_k(x), X^i)$. С его помощью осуществляется ранжирование моделей $\{a_1, a_2, ..., a_n\} \in A$, и для каждого сегмента определяются те, которые имеют наиболее высокие качественные показатели.

Поступающий на вход информационный поток подвергается такой же обработке, как и обучающая последовательность. На нем выделяются сегменты, определяются их свойства, и, в зависимости от них, назначается модель, имеющая лучшие значения функционала качества на схожем по свойствам сегменте обучающей выборки [22]. Затем выбранная модель $a_k(x)$ используется для решения задач обработки потока. Предсказанные моделью результаты спустя некоторое время сравниваются с реальными значениями объектов наблюдения, полученными от регистрирующих систем и устройств. В случае увеличения ошибок выше заранее определенного порога принимается решение о формировании данных для уточнения

алгоритма, которые впоследствии добавляются в обучающую выборку. Происходит дальнейшее обучение модели.

Таким образом, для каждого информационного потока формируется постоянно обучающаяся модель, где процессы обучения и обработки информационных потоков могут выполняться параллельно. Предварительное обучение на выборках со сходными свойствами может сократить временные затраты, когда необходимо в режиме реального времени осуществить назначение модели в условиях изменения свойств.

Модели верхнего уровня являются надстройкой, задачи которой – анализировать свойства входящего потока и назначать наиболее подходящую модель нижнего уровня.

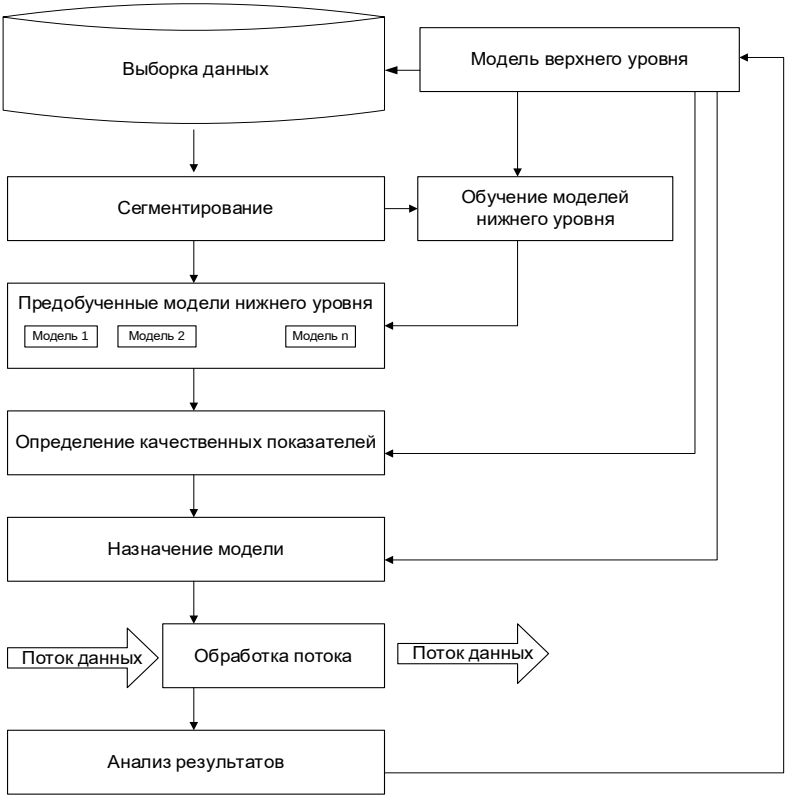

Рис. 2. Взаимодействие моделей верхнего и нижнего уровня

6. Эксперимент на модельных данных. Данные, поступающие от источников, часто представляют трудно разделимые последовательности. В них содержится множество классов, неоднородность областей, различные диапазоны значений. Один из примеров таких данных приведен на рисунке 3.

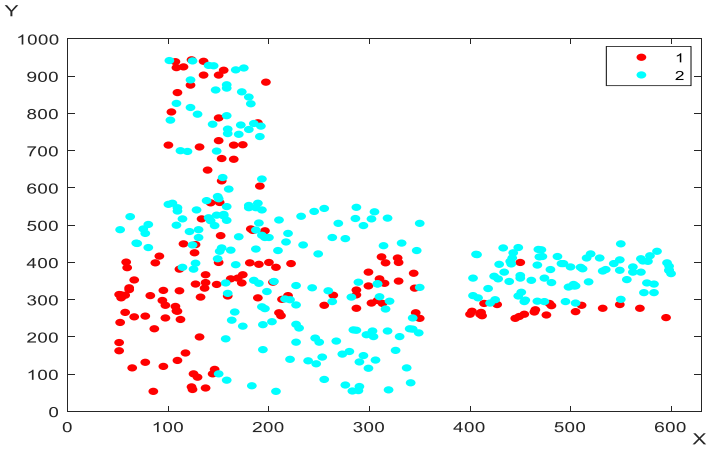

Рис. 3. Модельный ряд данных значений X, Y для двух классов

Для реализации предлагаемой модели осуществляется разделение набора на отдельные кластеры, обладающие разными свойствами. Задача такого разбиения состоит в том, чтобы уменьшить разброс данных, выбросы и использовать изменение диапазонов значений при формировании областей анализа для повышения показателей качества моделей.

Одним из проблемных вопросов является определение кластеров данных. В эксперименте их количество вычислялось на основе коэффициента силуэта, который показывает, насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров.

Полученные результаты коэффициента силуэта на рисунке 4 дают возможность выбрать наиболее подходящее количество кластеров. Оптимальное значение было получено при разбиении выборки на четыре подмножества.

Кластеризация дает возможность выделить группы схожих объектов, учесть выбросы, определить нетипичные объекты. Анализ групп может быть проведен с помощью графика функции плотности вероятности. При разделении множества необходимо добиваться

в подмножествах уменьшения пересекающихся площадей между классами. Поэтому в целях улучшения разделимости данных могут применяться различные алгоритмы.

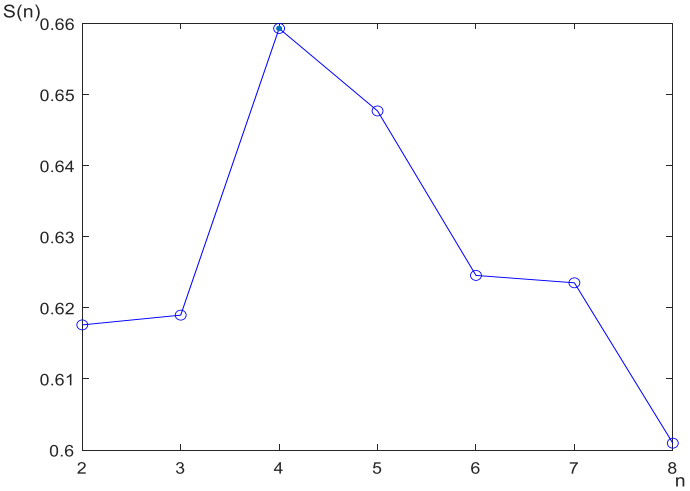

Рис. 4. Результаты значений коэффициента силуэта S(n) для n кластеров

На рисунке 5 приведены графики функций плотности вероятностей для всей выборки и кластеров, определенных методом К-средних. Полученные результаты кластеризации для разбиения выборки явились основой для дальнейшего исследования характеристик и качественных показателей различных классифицирующих алгоритмов.

В рассматриваемом примере разделение выборки данных, использующее коэффициент силуэта, позволяет для бинарной классификации построить такие группы, где суммарно площади пересечений функций плотности вероятностей в относительных единицах уменьшаются по сравнению со всей выборкой целиком.

На следующем этапе рассматривались отдельные модели и их ансамбли на кластерах и выборки целиком. В качестве базовых алгоритмов были выбраны: наивный байесовский классификатор с гауссовым ядром (NBK), линейный дискриминантный анализ (LD), квадратичный дискриминантный анализ с параболической функцией (QD), деревья решений с максимальным количеством ветвлений 100 (DT), метод К ближайших соседей (KNN) с количеством соседей 10, метод опорных векторов (SVM) с линейным ядром, метод

случайного леса с 30 деревьями (RF). Алгоритмы были поделены по своим качественным показателям на два подмножества: «сильных» и «слабых» классификаторов.

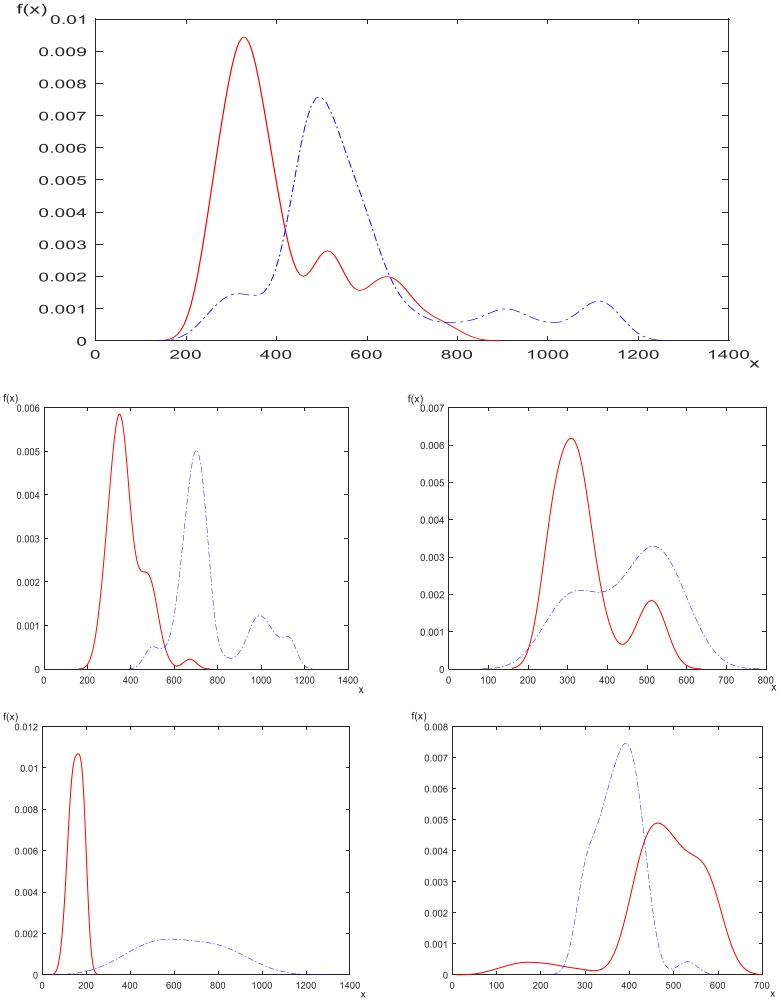

Рис. 5. Функции плотности вероятности f(x) для всей выборки (верхний рисунок) и кластеров (нижние 4 рисунка)

Каждая модель обучалась на всей последовательности целиком и на полученных в результате кластеризации подмножеств данных.

Результаты значения доли правильных ответов (accuracy) для каждого выбранного классификатора для всей выборки и при кластеризации приведены на рисунке 6.

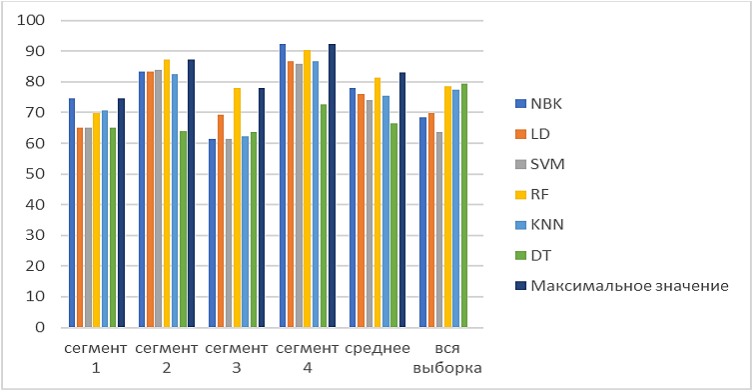

Рис. 6. Результаты значений доли правильных ответов (accuracy) для выборки и кластеров

Гистограмма показывает значения доли правильных ответов для каждого выбранного классификатора. Несмотря на отсутствие явного дисбаланса классов (соблюдаются пропорции 60:40), возникает ситуация, когда лучшие достигаемые качественные показатели на каждом сегменте и выборке показывают разные классификаторы. Выбирая для каждого сегмента классификатор, имеющий лучшее значение, становится возможным повысить качественные показатели обработки.

Таким образом, выделение сегментов данных и оценка их свойств позволяют осуществлять поиск и назначении моделей машинного обучения, обладающих лучшими характеристиками.

Аналогичным образом возможно сравнение ансамблей, состоящих из нескольких сложных классифицирующих моделей или простейших алгоритмов. В прилагаемом эксперименте был рассмотрен бэггинг. Его особенностью является возможность обучаться на множестве выборок, полученных из исходной. Каждая модель строится на основе случайного подмножества данных.

На практике не всегда удаётся сделать разнообразные независимые модели классификации. В рассмотренном примере классификаторы обучаются на одинаковых множествах, что снижает

их разнообразие. Не всегда возможно реализовать разделение обучающей выборки данных, чтобы данные оказались случайными, однородными и независимыми.

В результате может возникнуть ситуация, когда имеется, например, один «хороший» и один «плохой» по качественным показателям алгоритмы, а это будет приводить к тому, что качество результатов ансамбля будет хуже, чем у «хорошего» алгоритма. Результаты работы классификаторов и отдельных алгоритмов для полученных сегментов приведены на рисунке 7.

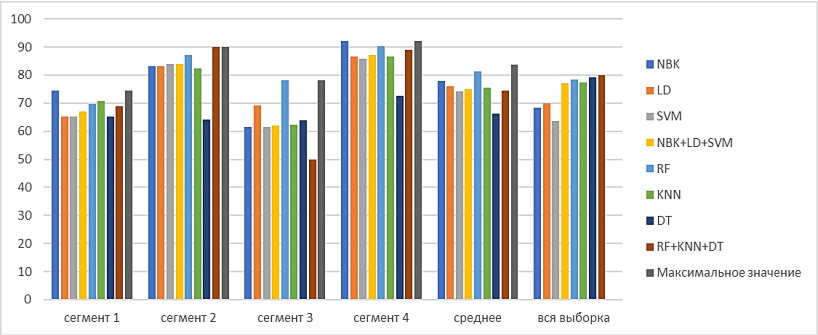

Рис. 7. Значения доли правильных ответов (accuracy) классификаторов и их моделей

На гистограмме видно, что на различных сегментах отдельные классифицирующие алгоритмы показывают результаты лучше, чем ансамбли. Применение усреднения по максимальным значениям на каждом сегменте дает возможность получить выигрыш по точности. Подходы, связанные с усреднением модели хорошо работают только тогда, когда отдельные классификаторы показывают результаты, имеющие высокую дисперсию. Ансамбль полезен в случае применения классификаторов, когда незначительные изменения свойств данных в выборке могут приводить к существенным изменениям классификации [4].

В то же время вычислительные затраты на агрегацию и обучение группы сложных классификационных моделей выше, чем затраты на обучение отдельного классификатора. Это может вызывать увеличение времени и вычислительных затрат, когда происходит смена концепции или изменение свойств данных, по сравнению с «подстановкой» готовой модели. Не всегда модели могут быть

построены с использованием различных комбинаций признаков, например при анализе одномерного ряда. А это, в свою очередь, влечет невозможность достижения их различности. Среднее значение моделей будет улучшением только в том случае, если модели независимы друг от друга.

В информационных потоках в условиях постоянно поступающих данных последовательности может происходить трансформация свойств данных. В результате, обученные на исторических данных слабые классификационные модели в различные временные промежутки могут становиться сильными и наоборот. Такие изменения свойств классифицирующих моделей происходят за очень короткий период, что приводит к тому, что качество решения задач ансамблем классифицирующих моделей оказывается хуже, чем одним из классификаторов. Это позволяет назначать лучшие классификаторы на подмножества.

7. Результаты обработки реальных данных. Следующим этапом эксперимента был рассмотрен ряд наборов данных, содержащих значения периодической выработки электроэнергии. В качестве экспериментальных данных рассматривались несколько наборов данных [24, 25], которые содержали данные о производстве электроэнергии в различных регионах Италии и Испании с 1995 по 2020 год. Объем первого набора около 30000 записей, 3 предиктора и 1 целевая переменная, второго – 250000 записей, 16 предикторов и 1 целевая переменная. Данные представлялись временными рядами. Для формирования обучающей выборки использовалось около 30% записей. Запись об объектах наблюдений определялась вектором *x*. Во время процессов обучения обрабатываемые последовательности были разделена на сегменты X^i. В каждом сегменте вычислялось значение его центра $\widehat{x}_c$. Были заранее определены модели обработки данных. Они обучались на выделенных сегментах, где определялся оптимальный функционал качества $Q(a_k(x), X^i) \to \min\limits_{a_k \in A}$.

В дальнейшем оставшаяся часть записей использовалась для имитации потока информации. По поступающим значениям с помощью функции расстояния между объектами $\rho(x_i, \widehat{x}_c)$ определялась принадлежность точек кластеру. В целях усреднения результатов было реализовано «окно», внутри которого постоянно определялись усредненные значения. Они сравнивались с заранее вычисленными центрами кластеров последовательности. Задавая ширину окна, возможно вычислять среднее значение

последовательности для оценки близости к предполагаемому центру, а также определять правила формирования выборок, основанные на частоте, для постоянного обучения сегментов.

В рассмотренном эксперименте были выбраны наборы данных [24, 25], описывающие генерацию «зеленой» электроэнергии в течение 3 лет. Предобучение заключалось в предварительном анализе выборок за 1 год. На обучающем множестве были выделены сегменты. На основе оценки коэффициента силуэта количество рекомендованных сегментов оказалось равным четырем. Наиболее вероятно, что это связано с погодными условиями, возникающими вследствие смены времен года.

В качестве анализируемых показателей были выбраны F-мера, доля правильных ответов (accuracy), площадь под кривой ошибок (AUC). Все выборки подвергались обработке целиком и алгоритмами, перечисленным в разделе 6.

Результаты качественных показателей, полученных при предсказании для сегментов и всей выборки целиком, представлены в таблицах 1–3.

Таблица 1. Результаты обработки набора данных

Классификатор	**Показатель**	**Вся выборка**	**Сегменты Power Supply dataset**				**среднее**
			1	2	3	4	
LD	**F мера**	0,80	0,84	0,81	0,82	0,81	0,82
	Accuracy %	70,3	78,9	73,1	74,0	72,8	74,7
	AUC	0,74	0,79	0,73	0,75	0,76	0,76
QD	**F мера**	0,81	0,84	0,81	0,82	0,81	0,82
	Accuracy %	71,3	78,2	72,2	74,2	73,4	74,5
	AUC	0,74	0,79	0,73	0,75	0,76	0,76
KNN	**F мера**	0,70	0,81	0,51	0,80	0,78	0,73
	Accuracy %	70,5	77,6	72,4	72,9	72,3	73,8
	AUC	0,71	0,79	0,72	0,74	0,76	0,76
NB	**F мера**	0,79	0,84	0,81	0,81	0,81	0,82
	Accuracy %	69,4	76,1	73,3	73,6	72,6	73,9
	AUC	0,72	0,78	0,73	0,75	0,76	0,76
SVM	**F мера**	0,81	0,81	0,83	0,83	0,82	0,83
	Accuracy %	71,4	77,5	77,6	77,6	77,5	77,6
	AUC	0,75	0,74	0,76	0,75	0,75	0,75
RF	**F мера**	0,83	0,83	0,84	0,85	0,84	0,84
	Accuracy %	71,5	74,5	77,1	73,6	73,5	74,7
	AUC	0,73	0,75	0,76	0,75	0,76	0,76

Таблица 2. Результаты обработки набора данных

Классификатор	Показатель	Вся выборка	Сегменты Valencia Sun energy generation dataset				среднее
			1'	2'	3'	4'	
LD	**F мера**	0,80	0,83	0,82	0,82	0,80	0,82
	Accuracy %	72,1	76,0	72,5	73,5	72,3	73,6
	AUC	0,71	0,73	0,74	0,74	0,76	0,75
QD	**F мера**	0,81	0,83	0,82	0,81	0,81	0,82
	Accuracy %	72,3	75,3	72,7	73,3	73,9	73,6
	AUC	0,74	0,74	0,75	0,74	0,77	0,75
KNN	**F мера**	0,75	0,80	0,62	0,79	0,78	0,75
	Accuracy %	70,5	75,7	71,0	71,6	72,1	72,6
	AUC	0,71	0,76	0,72	0,71	0,76	0,74
NB	**F мера**	0,79	0,83	0,82	0,80	0,80	0,83
	Accuracy %	70,3	73,5	73,4	71,5	72,2	72,7
	AUC	0,72	0,73	0,74	0,74	0,76	0,75
SVM	**F мера**	0,81	0,80	0,84	0,83	0,83	0,83
	Accuracy %	71,3	76,9	76,3	76,1	76,5	76,5
	AUC	0,73	0,72	0,76	0,73	0,74	0,74
RF	**F мера**	0,83	0,83	0,84	0,85	0,84	0,84
	Accuracy %	71,3	72,6	75,1	73,1	72,1	73,3
	AUC	0,74	0,74	0,76	0,73	0,75	0,75

Таблица 3. Результаты обработки набора данных

Классификатор	Показатель	Вся выборка	Сегменты Valencia Wind energy generation dataset				среднее
			1'	2'	3'	4'	
LD	**F мера**	0,81	0,84	0,81	0,85	0,83	0,84
	Accuracy %	71,3	77,0	72,1	73,9	72,9	74,0
	AUC	0,74	0,76	0,73	0,76	0,77	0,76
QD	**F мера**	0,82	0,84	0,85	0,81	0,82	0,83
	Accuracy %	72,3	75,1	73,2	74,1	74,1	74,1
	AUC	0,72	0,75	0,73	0,74	0,79	0,75
KNN	**F мера**	0,80	0,83	0,74	0,83	0,89	0,82
	Accuracy %	70,5	76,4	73,0	71,8	72,9	73,5
	AUC	0,71	0,77	0,73	0,72	0,78	0,75
NB	**F мера**	0,78	0,85	0,83	0,84	0,84	0,84
	Accuracy %	70,4	73,7	74,5	72,4	72,3	73,2
	AUC	0,73	0,74	0,76	0,79	0,78	0,77
SVM	**F мера**	0,82	0,83	0,82	0,84	0,83	0,83
	Accuracy %	71,9	75,3	72,3	73,1	74,2	73,4
	AUC	0,71	0,72	0,73	0,72	0,72	0,72
RF	**F мера**	0,82	0,83	0,81	0,84	0,83	0,83
	Accuracy %	71,4	71,9	73,3	72,9	71,4	72,4
	AUC	0,76	0,74	0,80	0,75	0,77	0,77

Результаты эксперимента показывают, что на отдельных сегментах значения показателя качества отдельных алгоритмов лучше, чем при обработке всей выборки целиком. Реализовав функцию, производящую анализ качественных показателей, можно назначать на сегмент модель, имеющую на нем лучшее значение, что позволяет в зависимости от показателя получить выигрыш около 3-5%.

8. Заключение. Применение моделей и методов искусственного интеллекта возможно в проблемных областях, где возникает неопределенность и трудно определить оптимальные решения.

В целях повышения показателей качества классификационных и регрессионных моделей существует возможность реализации процессов предварительной обработки выборки данных, чтобы извлечь определенные атрибуты, которые представляют наиболее важные характеристики информации. В различных сегментах приходится реализовывать разделяющие поверхности разной сложности, что приводит к тому, что на разных подвыборках лучше работают различные модели.

Сбор объектов наблюдения является трудоемкой задачей, где внутри кортежей могут возникать различные смещения из-за настроек регистрирующих устройств или потерь точности значений отдельных параметров. Извлечение признаков может терять свою актуальность в случае возникновения дрейфа концепции. В связи с этим необходимо постоянно обрабатывать поступающие на вход выборки данных и анализировать каждый сегмент.

Информация о свойствах данных в сегментах сильно зависит от способа сегментации и разделения выборки. Обработка этих данных необходима, чтобы получить информацию о разделимости классов, сформировать разделяющую поверхность и повысить качественные показатели классифицирующего алгоритма.

Использование нескольких моделей для повышения качества результатов предсказания в виде ансамблевых методов, приводит к тому, что, несмотря на различные комбинации, объединяющие отдельные алгоритмы в модель, возникают ситуации, где такое объединение не только может не повысить, но и даже ухудшить результат. В связи с этим необходимо предотвращать подобные ситуации. Для этого предлагаются использовать многоуровневые модели, где на нижнем уровне идет обработка последовательности, а на верхнем – происходят процедуры оценивания данных и назначения алгоритмов.

Выполнение процесса сегментации и вычисления характеристик должно осуществляться постоянно для настройки моделей обработки.

Литература

1. Pouyanfar S., Sadiq S., Yan Y., Tian H., Tao Y., Reyes M.P., Shyu M.L., Chen S.C., Iyengar S.S. A survey on deep learning: algorithms, techniques, and applications // ACM Computing Surveys. 2019. vol. 51. no. 5. pp. 1–36.
2. Blyth C.R. On Simpson's Paradox and the Sure-Thing Principle // Journal of the American Statistical Association. 1972. vol. 67. pp. 364–387.
3. McConnell S., Skillicorn D.B. Building predictors from vertically distributed data // Proceedings of the 2004 Conference of the Centre for Advanced Studies on Collaborative Research (CASCON '04). 2004. pp. 150–162.
4. Trevizan B., Chamby-Diaz J., Bazzan A.L.C., Recamonde-Mendoza M. A comparative evaluation of aggregation methods for machine learning over vertically partitioned data // Expert Systems with Applications. 2020. vol. 152. pp. 113–126.
5. Li Y., Jiang Z.L., Yao L. et al. Outsourced privacy-preserving C4.5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties // Cluster Computation. 2019. vol. 22. no. 1. pp.1581–1593.
6. Mendoza M.R., Bazzan A.L.C. On the Ensemble Prediction of Gene Regulatory Networks: a Comparative Study // Proceedings of the Brazilian Symposium on Neural Networks. 2012. pp. 55–60.
7. Chan P.K., Stolfo S.J. On the Accuracy of Meta-learning for Scalable Data Mining // Journal of Intelligent Information Systems. 1997. no. 8. pp. 5–28.
8. Sun L., Mu W.S., Qi B. et al. A new privacy-preserving proximal support vector machine for classification of vertically partitioned data // International journal of machine learning and cybernetics. 2015. vol. 3. no. 6. pp. 109–118.
9. Zhou Z.-H., Feng J. Deep forest // National Science Review. 2019. vol. 6. no. 1. pp. 74–86.
10. Ho T.K. The random space method for constructing decision forests // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998. vol. 20. no. 8. pp. 832–844.
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.
12. Liu J., Li Y., Song S., Xing J., Lan C., Zeng W. Multi-modality multi-task recurrent neural network for online action detection // IEEE Transactions on Circuits and Systems for Video Technology. 2018. vol. 29. no. 9. pp. 2667–2682.
13. Salehi H., Burgueno R. Emerging artificial intelligence methods in structural engineering // Engineering Structures. 2018. no. 171. pp. 170–189.
14. 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. vol. 31. no. 12. pp. 2346–2363.
15. Zhang X., Wang M. Weighted Random Forest Algorithm Based on Bayesian Algorithm // Journal of Physics: Conference Series. 2021. vol. 1924. pp. 1–6.
16. Scanagatta M., Salmeron A., Stella F. A survey on Bayesian network structure learning from data // Progress in Artificial Intelligence. 2019. no. 8, pp. 425–439.
17. Wright M., Dankowski T., Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics // Statistics in Medicine. 2017. vol. 36. no. 8. pp. 1272–1284.
18. Zheng X., Aragam B., Ravikumar P., Xing E. DAGs with no tears: Continuous optimization for structure learning // Advances in Neural Information Processing Systems. 2018. vol. 43. pp. 9492–9503.
19. Di Franco G., Santurro M. Machine learning, artificial neural networks and social research // Qual Quant. 2021. no. 5. pp. 1007–1025.

20. Scanagatta M., Corani G., Zaffalon M., Yoo J., Kang U. Efficient learning of bounded-treewidth Bayesian networks from complete and incomplete data sets // International Journal of Approximate Reasoning. 2019. vol. 95. pp. 152–166.
21. Kheyreddine D., Kadda B.-B., Abdenour A. A new adaptive sampling algorithm for big data classification // Journal of Computational Science. 2022. vol. 61. pp. 101–116.
22. Лебедев И.С. Сегментирование множества данных с учетом информации воздействующих факторов // Информационно-управляющие системы. 2021. № 3. С. 29–38.
23. Лебедев И.С. Адаптивное применение моделей машинного обучения на отдельных сегментах выборки в задачах регрессии и классификации // Информационно-управляющие системы. 2022. № 3. С. 20–30.
24. Power Supply dataset. URL: http://www.cse.fau.edu/~xqzhu/stream.html (Дата обращения 27.10.2022).
25. Energy generation dataset. URL: https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather/data?select=energy_dataset.csv (Дата обращения 27.10.2022).

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 линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-3311.

DOI 10.15622/ia.22.3.1

I. LEBEDEV

APPLICATION OF MULTILEVEL MODELS IN CLASSIFICATION AND REGRESSION PROBLEMS

Lebedev I. **Application of Multilevel Models in Classification and Regression Problems.**

Abstract. There is a constant need to create methods for improving the quality indicators of information processing. In most practical cases, the ranges of target variables and predictors are formed under the influence of external and internal factors. Phenomena such as concept drift cause the model to lose its completeness and accuracy over time. The purpose of the work is to improve the processing data samples quality based on multi-level models for classification and regression problems. A two-level data processing architecture is proposed. At the lower level, the analysis of incoming information flows and sequences takes place, and the classification or regression tasks are solved. At the upper level, the samples are divided into segments, the current data properties in the subsamples are determined, and the most suitable lower-level models are assigned according to the achieved qualitative indicators. A formal description of the two-level architecture is given. In order to improve the quality indicators for classification and regression solving problems, a data sample preliminary processing is carried out, the model's qualitative indicators are calculated, and classifiers with the best results are determined. The proposed solution makes it possible to implement constantly learning data processing systems. It is aimed at reducing the time spent on retraining models in case of data properties transformation. Experimental studies were carried out on several datasets. Numerical experiments have shown that the proposed solution makes it possible to improve the quality processing indicators. The model can be considered as an improvement of ensemble methods for processing information flows. Training a single classifier, rather than a group of complex classification models, makes it possible to reduce computational costs.

Keywords: machine learning, multilevel models, purpose of classifying algorithms.

References

1. Pouyanfar S., Sadiq S., Yan Y., Tian H., Tao Y., Reyes M.P., Shyu M.L., Chen S.C., Iyengar S.S. A survey on deep learning: algorithms, techniques, and applications. ACM Computing Surveys. 2019. vol. 51. no. 5. pp. 1–36.
2. Blyth C.R. On Simpson's Paradox and the Sure-Thing Principle. Journal of the American Statistical Association. 1972. vol. 67. pp. 364–387.
3. McConnell S., Skillicorn D.B. Building predictors from vertically distributed data. Proceedings of the 2004 Conference of the Centre for Advanced Studies on Collaborative Research (CASCON '04). 2004. pp. 150–162.
4. Trevizan B., Chamby-Diaz J., Bazzan A.L.C., Recamonde-Mendoza M. A comparative evaluation of aggregation methods for machine learning over vertically partitioned data. Expert Systems with Applications. 2020. vol. 152. pp. 113–126.
5. Li Y., Jiang Z.L., Yao L. et al. Outsourced privacy-preserving C4.5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties. Cluster Computation. 2019. vol. 22. no. 1. pp.1581–1593.
6. Mendoza M.R., Bazzan A.L.C. On the Ensemble Prediction of Gene Regulatory Networks: a Comparative Study. Proceedings of the Brazilian Symposium on Neural Networks. 2012. pp. 55–60.
7. Chan P.K., Stolfo S.J. On the Accuracy of Meta-learning for Scalable Data Mining. Journal of Intelligent Information Systems. 1997. no. 8. pp. 5–28.

8. Sun L., Mu W.S., Qi B. et al. A new privacy-preserving proximal support vector machine for classification of vertically partitioned data. International journal of machine learning and cybernetics. 2015. vol. 3. no. 6. pp. 109–118.
9. Zhou Z.-H., Feng J. Deep forest. National Science Review. 2019. vol. 6. no. 1. pp. 74–86.
10. Ho T.K. The random space method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998. vol. 20. no. 8. pp. 832–844.
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.
12. Liu J., Li Y., Song S., Xing J., Lan C., Zeng W. Multi-modality multi-task recurrent neural network for online action detection. IEEE Transactions on Circuits and Systems for Video Technology. 2018. vol. 29. no. 9. pp. 2667–2682.
13. Salehi H., Burgueno R. Emerging artificial intelligence methods in structural engineering. Engineering Structures. 2018. no. 171. pp. 170–189.
14. 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. vol. 31. no. 12. pp. 2346–2363.
15. Zhang X., Wang M. Weighted Random Forest Algorithm Based on Bayesian Algorithm. Journal of Physics: Conference Series. 2021. vol. 1924. pp. 1–6.
16. Scanagatta M., Salmeron A., Stella F. A survey on Bayesian network structure learning from data. Progress in Artificial Intelligence. 2019. no. 8, pp. 425–439.
17. Wright M., Dankowski T., Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. Statistics in Medicine. 2017. vol. 36. no. 8. pp. 1272–1284.
18. Zheng X., Aragam B., Ravikumar P., Xing E. DAGs with no tears: Continuous optimization for structure learning. Advances in Neural Information Processing Systems. 2018. vol. 43. pp. 9492–9503.
19. Di Franco G., Santurro M. Machine learning, artificial neural networks and social research. Qual Quant. 2021. no. 5. pp. 1007–1025.
20. Scanagatta M., Corani G., Zaffalon M., Yoo J., Kang U. Efficient learning of bounded-treewidth Bayesian networks from complete and incomplete data sets. International Journal of Approximate Reasoning. 2019. vol. 95. pp. 152–166.
21. Kheyreddine D., Kadda B.-B., Abdenour A. A new adaptive sampling algorithm for big data classification. Journal of Computational Science. 2022. vol. 61. pp. 101–116.
22. Lebedev I.S. [Dataset segmentation considering the information about impact factors]. Informacionno-upravljajushhie sistemy – Information and Control Systems. 2021. no. 3. pp. 29–38. (In Russ.).
23. Lebedev I.S. [Adaptive application of machine learning models on separate segments of a data sample in regression and classification problems]. Informacionno-upravljajushhie sistemy – Information and Control Systems. 2022. no. 3. pp. 20–30. (In Russ.).
24. Power Supply dataset. Available at: http://www.cse.fau.edu/~xqzhu/stream.html (accessed 27.10.2022).
25. Energy generation dataset. Available at: https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather/data?select=energy_dataset.csv (accessed 27.10.2022).

Lebedev Ilya — Ph.D., Dr.Sci., Professor, Chief researcher, 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)328-3311.