

УДК 004.855.5

doi: 10.15622/rcai.2025.072

ГЛУБОКИЙ ЛЕС ДЛЯ АНАЛИЗА ВЫЖИВАЕМОСТИ В РАМКАХ МНГОВОВАРИАНТНОГО ОБУЧЕНИЯ¹

А.В. Константинов (*konstantinov_av@spbstu.ru*)

Л.В. Уткин (*utkin_lv@spbstu.ru*)

Санкт-Петербургский политехнический университет
Петра Великого, Санкт-Петербург

В работе рассматривается задача на стыке машинного обучения и анализа выживаемости в условиях многовариантного обучения с табличными данными. Проблема возникает, когда каждый объект характеризуется не одним, а набором векторов признаков, однако событие происходит для объекта в целом. Для решения данной задачи предложен новый метод на основе глубокого леса. Ключевой особенностью метода является построение модели, обрабатывающей множество векторов признаков в качестве входа, элементами которой являются классические ансамбли деревьев решения. Численные эксперименты с использованием реальных данных демонстрируют применимость предложенного метода.

Ключевые слова: анализ выживаемости, многовариантное обучение, деревья решений, ансамбли, глубокий лес.

Введение

Анализ выживаемости является важной задачей в различных областях науки, включая медицину, инженерию и экономику, позволяющей оценивать вероятность события (например, смерти пациента, отказа оборудования или оттока клиента) в зависимости от времени. В отличие от классификационных моделей, методы анализа выживаемости позволяют оценивать функцию распределения или функцию выживаемости для каждого отдельного объекта, характеризующегося вектором признаков. В последние годы наблюдается растущий интерес к применению методов машинного обучения для решения задач анализа выживаемости, особенно в условиях, когда число примеров мало, а данные представлены в табличном виде с большим

¹ Работа выполнена при финансовой поддержке РФФ (проект № 25-21-00103).

количеством признаков. Традиционные методы анализа выживаемости, такие как модель Кокса, случайный лес выживаемости (Random Survival Forest, RSF), успешно применяются на практике, однако не дают возможности использовать в качестве входа множества векторов признаков.

Вместе с тем возникает специфическая проблема при работе с данными, где каждый объект характеризуется не одним набором признаков, а множеством экземпляров, каждый из которых представляется в виде отдельного вектора признаков. Эти векторы могут представлять собой характеристики однородных элементов системы, результаты различных анализов одного пациента в разных условиях или параметры взаимодействующих объектов. При этом событие (например, отказ) относится к объекту в целом, а не к конкретному вектору признаков из множества. Этот сценарий соответствует парадигме многовариантного обучения (МВО) [Carbonneau et al., 2018], которая традиционно применялась для задач классификации изображений и распознавания объектов, анализа гистологических изображений высокого разрешения, прогнозирования активности препаратов, и так далее. Интеграция методов анализа выживаемости с подходом МВО представляет собой актуальную и сложную задачу, требующую разработки новых алгоритмов, способных эффективно обрабатывать данные в виде множеств векторов признаков в условиях цензурированности.

В настоящей работе предлагается новый метод решения задачи анализа выживаемости в условиях МВО на основе композиции случайных лесов выживаемости в формате глубокого леса [Zhou, 2019]. Предложенный подход заключается в построении модели, которая принимает на вход множество векторов признаков и использует ансамбль деревьев решений для оценки функции риска. Ключевой особенностью разработанного метода является его способность эффективно агрегировать информацию из различных векторов признаков, представляющих один объект, для прогнозирования времени до события. Результаты численных экспериментов на реальных данных демонстрируют перспективность предложенного подхода и его применимость в задачах анализа выживаемости с многовариантными данными.

1. Анализ выживаемости и многовариантное обучение

В данном разделе приведены постановки задачи выживаемости, МВО и рассматриваются способы определения задачи анализа выживаемости в рамках МВО.

1.1. Анализ выживаемости

Задача анализа выживаемости заключается в моделировании условного распределения времени до наступления определенного события (например, смерти, отказа оборудования или рецидива заболевания) при за-

данном векторе признаков объекта \mathbf{x} . А именно, требуется оценить функцию выживаемости $S(t | \mathbf{x}) = P(T > t | \mathbf{x})$. В отличие от традиционных задач классификации и регрессии, анализ выживаемости учитывает возможность цензурирования данных – ситуации, когда событие не было зафиксировано в течение периода наблюдения за объектом. Поэтому набор обучающих данных, по которому требуется построить оценку, состоит из троек $(\mathbf{x}_i, y_i, \delta_i)$, где помимо вектора признаков \mathbf{x}_i и времени y_i содержится метка цензурирования δ_i , принимающая либо значение 1, обозначающее что событие в данный момент произошло, либо 0, если событие не произошло до этого момента.

1.2. Многовариантное обучение

В парадигме МВО набор данных состоит из множеств или групп экземпляров, каждый из которых содержит отдельные экземпляры-векторы признаков. Стандартной задачей в рамках МВО является бинарная классификация, где класс группы считается положительным, если хотя бы один экземпляр в этой группе является ее представителем, то есть соответствует определенному условию. Задача заключается в построении модели, которая, принимая на вход множество экземпляров, определяет класс для всей группы. В отличие от традиционных задач обучения с учителем, где каждый объект имеет однозначную метку, в многовариантном обучении метка доступна только для группы [Carboneau et al., 2018]. Поскольку целью настоящей работы является построение модели выживаемости, рассматривается более общая постановка, где метка группы может соответствовать не одному конкретному экземпляру группы, а части или всем экземплярам в совокупности. Такая общая постановка рассмотрена в работе [Yao et al., 2019], где для анализа выживаемости в рамках многовариантного обучения проводился анализ гистологических образцов высокого разрешения, фрагменты которых представляли собой отдельные экземпляры, а времена событий были доступны только для целых изображений. Таким образом, требуется построить не зависимость от одного экземпляра и определить активный экземпляр в группе, а функцию множества экземпляров. Альтернативный подход к определению МВО для анализа выживаемости, заключающийся во включении явной агрегации функций выживаемости в постановку задачи, требует разработки специализированных правил для каждой конкретной прикладной области и поэтому не рассматривается в данной работе.

1.3. Модели на основе деревьев решений

Наиболее часто применяемыми для обработки табличных данных являются модели на основе ансамблей деревьев решений. Для задач выживаемости таким ансамблем является случайный лес выживаемости [Ishwaran, 2008]. Он существенно отличается от классических лесов де-

ревью решений. В листьях каждого дерева такого леса строится безусловная модель Каплана-Мейера, а правила расщепления в узлах определяются таким образом, чтобы максимизировать логранговый критерий. Стоит отметить, что несмотря на высокий потенциал деревьев решений для обработки табличных данных и успешное использование таких моделей для МВО [Leistner et al., 2010], модели выживаемости на основе деревьев в рамках МВО ранее не применялись.

Для реализации модели на основе деревьев решений для анализа выживаемости в рамках МВО, в настоящей работе предлагается строить более сильную, композиционную модель, включающую множество этапов обработки данных, то есть модель, представляющую собой глубокий лес (Deep Forest) [Zhou, 2019]. Такой лес, в отличие от классического алгоритма случайного леса, где каждое дерево строится независимо, строится послойно, используя выходные данные предыдущих деревьев в качестве входных признаков для последующих. Это позволяет модели реализовывать обучение представлениям, учитывая более сложные зависимости между признаками и повышая за счет этого обобщающую способность. Глубокий лес демонстрирует высокую точность и эффективность на различных задачах машинного обучения, включая задачи классификации, регрессии, а также анализа выживаемости [Utkin et al., 2020], однако ранее глубокий лес для анализа выживаемости в рамках МВО разработан не был.

2. Метод построения глубокого леса выживаемости в условиях МВО

Ансамбли деревьев решений позволяют строить наиболее точные модели для табличных данных, в том числе в рамках анализа выживаемости. Поэтому в данной работе предлагается реализовать подход на основе таких ансамблей, опирающийся на их преимущества, и вместе с тем позволяющий обрабатывать данные в формате МВО.

Для применения ансамбля деревьев решений к группам примеров можно рассмотреть два подхода. Первый подход предполагает выбор репрезентативных объектов (представителей) из каждой группы для формирования набора данных в формате “объект – время – метка события”. Однако данный подход оказывается неприменим в случае, если у группы отсутствует один представитель, т.е. если более одного экземпляра в группе влияют на функцию выживаемости. Поэтому предлагается второй подход, состоящий в создании псевдо-разметки путем повторения метки группы для каждого экземпляра и последующему применению ансамблевого метода ко всем экземплярам каждой группы. На рис. 1 представлена схема предлагаемого метода, где указано как для всех экземпляров первой группы $x_{1,1}$, $x_{1,2}$, $x_{1,3}$ повторяются метки y_1 , δ_1 . Сформированный набор данных, где каждому экземпляру сопоставлены метки времени и цензурирования, используется для обучения случайного леса выживаемости.

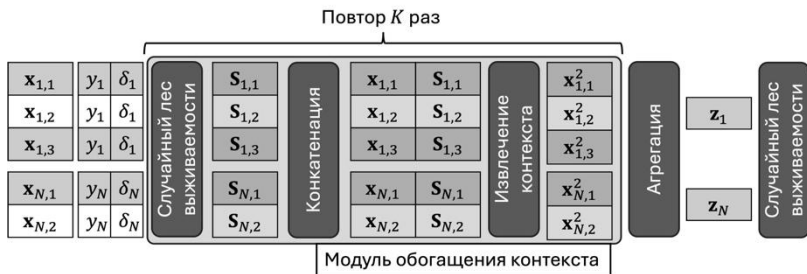


Рис. 1. Общая схема предлагаемого глубокого леса

Далее с помощью построенного леса выживаемости для каждого экземпляра строится оценка функции выживаемости: $S_{1,1}, \dots, S_{N,2}$. Поскольку число уникальных времен событий E в обучающем наборе данных конечно, функции выживаемости представляются в виде векторов, размерность которых совпадает с E , а каждое значение задает вероятность, что событие произойдет после соответствующего временного интервала. Полученные оценки можно рассматривать как дополнение представления исходных векторов экземпляров. С целью снижения объёма обрабатываемых данных производится понижение разрешения временной сетки, путем усреднения значений выживаемости в соседних интервалах. Для получения нового представления исходные векторы конкатенируются к соответствующим оценкам выживаемости, таким образом формируются новые векторы $(x_{1,1}, S_{1,1}), \dots, (x_{N,2}, S_{N,2})$.

Полученные новые представления экземпляров комбинируются в блоке извлечения контекста, для расширения представления экземпляра информации о группе, в которую он включён. К каждому вектору признаков конкатенируется среднее значение представлений всех остальных экземпляров в группе. С целью снижения вычислительной сложности данной процедуры, которая в простейшей реализации является квадратичной по числу экземпляров в группе, для каждой группы размера g сперва рассчитывается сумма всех представлений, затем для каждого экземпляра из суммы вычитается его вектор-представление, и полученный вектор умножается на $1/(g - 1)$. Сложность при такой реализации становится линейной. Блок извлечения контекста позволяет учесть взаимодействия отдельных экземпляров и их влияние в совокупности на распределение времени события.

После применения вышеописанной процедуры, составляющей модуль обогащения контекста, для каждого экземпляра формируется новое представление, содержащее, помимо информации о данном экземпляре, контекст, то есть необходимую информацию о других экземплярах группы,

или ее часть. Данный модуль применяется K раз (с различными случайными лесами). Затем для построения модели на уровне групп применяется агрегация за счет усреднения представлений в каждой группе. В результате формируется обучающий набор данных, где вместо экземпляров содержатся векторы-представления групп $\mathbf{z}_1, \dots, \mathbf{z}_N$, которым соответствуют исходные времена y_1, \dots, y_N . С помощью данного обучающего набора строится финальный случайный лес выживаемости, который позволяет строить функцию выживаемости для групп, а не экземпляров в отдельности.

3. Численные эксперименты

Предложенный метод был реализован на языке Python с использованием фреймворка для разработки глубоких лесов Bosk и случайных лесов выживаемости из пакета scikit-survival. Для проведения численных экспериментов с реальными данными были использованы четыре широко применяемых для анализа выживаемости набора данных: Breast Cancer (BC), German Breast Cancer Study Group 2 (GBSG2), Veterans и Worcester Heart Attack Study (WHAS500).

Для исследования предложенного метода в рамках МВО исходные наборы данных были приведены к формату группа – время – метка события с помощью следующего преобразования. Для каждого примера вектор признаков разбивался на множество новых векторов признаков, описывающее объект (группу) обучающей выборки в рамках МВО. Каждый такой вектор признаков соответствует отдельной компоненте исходного вектора, и состоит из значения и номера данной компоненты. Таким образом, размер множества (группы) равен числу признаков в исходной задаче, а сами векторы признаков в новой задаче принадлежат дизъюнктному объединению областей определения признаков исходной задачи.

Тестирование проводилось с помощью перекрёстной проверки (кросс-валидации) с 5 разбиениями, сбалансированными по доле цензурированных примеров, повторяющейся с различными разбиениями 20 раз. Для оценки точности применялся индекс конкордации (С-индекс), позволяющий оценить корректность ранжирования ожидаемого времени события с помощью модели, и интегрированная оценка Брайера (IBS), позволяющая оценить соответствие предсказанного распределения тестовой выборке. Далее представлены результаты сравнения предложенного глубокого леса (ГЛ-МВО), решающего поставленную проблему в рамках МВО, с выступающим в качестве базового метода классическим случайным лесом выживаемости (RSF) и RSF со рандомизированным построением деревьев решений (RSF-Extra), применёнными к исходной выборке. Результаты для С-индекса приведены в табл. 1, где каждая строка соответствует отдельной тестируемой модели. Чем выше значение С-индекса, тем ниже ошибка ранжирования ожидаемого времени моделями.

Как видно из табл. 1, предложенная модель не уступает по точности базовому RSF в трех из четырех случаев, а для набора данных Veteran позволяет существенно превзойти его по точности, в то время как задача в многовариантной постановке является более сложной, чем исходная.

Таблица 1

Модель	BC	GBSG2	Veterans	WHAS500
RSF	0,650	0,689	0,675	0,767
RSF-Extra	0,689	0,673	0,653	0,745
ГЛ-MBO	0,677	0,687	0,702	0,760

Для интегрированной оценки Брайера результаты приведены в табл. 2. Чем меньше значение оценки, тем лучше прогнозы модели соотносятся тестовой выборкой.

Таблица 2

Модель	BC	GBSG2	Veterans	WHAS500
RSF	0,162	0,173	0,129	0,167
RSF-Extra	0,154	0,179	0,133	0,177
ГЛ-MBO	0,155	0,173	0,127	0,175

Аналогично, в соответствии с табл. 2, предложенная модель превосходит базовый RSF на всех наборах данных, кроме WHAS500, а базовый RSF-Extra на всех, кроме BC.

Таким образом, можно сделать вывод, что предложенный метод позволяет решать задачи выживаемости в условиях многовариантного обучения с достаточной точностью. С вычислительной точки зрения, затраты для построения глубокого леса превосходят таковые у базового метода (RSF) приблизительно в $(K + 1)$ раз, что не является существенным ограничением для применимости данного метода в условиях MBO.

Заключение

В настоящей работе предложен новый метод решения задачи анализа выживаемости в условиях MBO с табличными данными на основе глубокого леса. Разработанный подход позволяет эффективно обрабатывать в качестве входа множество векторов признаков, представляющих один объект, и агрегировать информацию из них для прогнозирования распределения времени до события. Ключевой особенностью метода является построение многослойной модели, состоящей из ансамблей деревьев решений и блоков извлечения контекста, способной учитывать сложные зависимости между признаками и обеспечивающей высокую точность оценки функции выживаемости.

Численные эксперименты с использованием реальных данных продемонстрировали перспективность предложенного подхода и его применимость в задачах анализа выживаемости с многовариантными данными. Полученные результаты показали, что разработанный метод демонстрирует сопоставимую точность по сравнению с ансамблями, примененными к классической задаче анализа выживаемости на исходных данных, либо превосходит базовые методы по показателям индекса конкордации и интегрированной оценки Брайера.

Несмотря на достигнутые результаты, следует отметить некоторые ограничения предложенного метода. В частности, эффективность работы алгоритма может зависеть от качества и структуры входных данных, а также от выбора оптимальных гиперпараметров модели. Кроме того, дальнейшие исследования могут быть направлены на разработку методов автоматической настройки гиперпараметров для повышения устойчивости и обобщающей способности предложенного подхода на каждом слое глубокого леса. В перспективе представляется интересным изучение возможности применения разработанного метода к другим типам данных, таким как изображения, разбитые на фрагменты – патчи, или текстовые данные, а также расширение его функциональности для обеспечения интерпретируемости модели в целом с целью выявления экземпляров и их признаков, вносящих наиболее значимый вклад в предсказания.

Список литературы

- [Carbonneau et al., 2018] Carbonneau M.A., Cheplygina V., Granger E., Gagnon G. Multiple instance learning: A survey of problem characteristics and applications // *Pattern recognition*. – 2018. – Vol. 77. – P. 329-353. – doi: 10.1016/j.patcog.2017.10.009.
- [Ishwaran, 2008] Ishwaran H., Kogalur U.B., Blackstone E.H., Lauer M.S. Random Survival Forests // *The Annals of Applied Statistics*. – 2008. – P. 841-860. – doi: 10.1214/08-AOAS169.
- [Leistner et al., 2010] Leistner C., Saffari A., Bischof H. MIForests: Multiple-instance learning with randomized trees // *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11*. – Springer Berlin Heidelberg, 2010. – P. 29-42. – doi: 10.1007/978-3-642-15567-3_3.
- [Utkin et al., 2020] Utkin L.V., Konstantinov A.V., Lukashin A.A., Muliukha V.A. An adaptive weighted deep survival forest // *2020 XXIII International Conference on Soft Computing and Measurements (SCM)*. – IEEE, 2020. – P. 198-201. – doi: 10.1109/SCM50615.2020.9198755.
- [Yao et al., 2019] Yao J., Zhu X., Huang J. Deep multi-instance learning for survival prediction from whole slide images // *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*. – Springer International Publishing, 2019. – P. 496-504. – doi: 10.1007/978-3-030-32239-7_55.
- [Zhou, 2019] Zhou Z.H., Feng J. Deep forest // *National science review*. – 2019. – Vol. 6, No. 1. – P. 74-86. – doi: 10.1093/nsr/nwy108.