Адаптивный весовой глубокий лес выживаемости

Л. В. Уткин[1], А. В. Константинов[2], А. А. Лукашин[3], В. А. Мулюха[4]
Высшая школа прикладной математики и вычислительной физики
Санкт-Петербургский политехнический университет Петра Великого
Санкт-Петербург, Россия
[1]lev.utkin@gmail.com, [2]andrue.konst@gmail.com, [3]alexey.lukashin@spbstu.ru, [4]mulyukha_va@almazovcentre.ru

***Аннотация*. Предложена адаптивная взвешенная модель глубокого леса для анализа выживаемости пациентов. Модель является расширением адаптивного взвешенного глубокого леса. Во-первых, она основана на том, что глубокий лес представляет собой композицию моделей, включающую набор случайных лесов, организованных в виде уровней каскада глубокого леса, аналогично слоям в нейронных сетях. Во-вторых, модель использует специальную схему назначения весов обучающим примерам в глубоком лесу, которая позволяет адаптировать случайные леса выживаемости на каждом уровне к данным обучения. Одна из основных идей, лежащих в основе предлагаемой модели, состоит в том, чтобы ввести и применить маргинальный индекс согласованности (MC-индекс) в качестве меры качества предсказания, связанного с примером, и вычислить весовые коэффициенты как функции MC-индексов. Численные примеры с реальными данными иллюстрируют предлагаемую адаптивную модель.**

Ключевые слова: случайный лес; C-индекс; анализ выживаемости; модели композиций; стекинг

I. Введение

Важной задачей в медицине является выбор подходящего лечения для определенного пациента, который может отличаться от других пациентов по своим клиническим или другим характеристикам. Основой для правильного решения являются цензурированные данные, характерные для медицины и исследуемые в рамках анализа выживаемости [1]. Можно рассмотреть три типа моделей выживаемости. Первый тип охватывает параметрические модели, основанные на известных распределениях вероятностей времени до интересующего нас события, например, смерти пациента. Второй тип включает полупараметрические модели, которые не предполагают знания распределения времени до определенного события, но делают предположения о том, как признаки, характеризующие пациента, изменяют данные о выживаемости. Хорошо известной полупараметрической моделью является модель пропорциональных рисков Кокса [2], основанная на использовании условия линейной зависимости признаков. Третий тип – это непараметрические модели, используемые, когда теоретические распределения вероятностей не соответствуют данным. Одной из известных непараметрических моделей выживаемости является модель Каплана-Мейера, где предполагаемая функция выживаемости постоянна между событиями.

Некоторые непараметрические модели выживаемости основаны на использовании случайных лесов выживаемости (СЛВ), которые можно рассматривать как расширение случайных лесов (СЛ) [4] с учетом цензурированности обучающих данных [5]. Эти модели дают превосходящие результаты, когда обучающих данных мало или когда в обучающей выборке имеется много цензурированных примеров [6,7].

Для расширения и улучшения СЛВ был предложен глубокий лес выживаемости (ГЛВ) [8] в качестве расширения так называемого глубокого леса (ГЛ) [9] на случай цензурированных данных для анализа выживаемости. ГЛ является моделью на основе композиции СЛ, которая включает в себя набор СЛ, организованных в виде уровней каскада лесов, аналогично слоям в нейронных сетях. Однако нейроны заменяются различными СЛ, которые играют ту же роль, что и нейроны. СЛ в ГЛВ заменяются на СЛВ. Однако в отличие от исходного ГЛ ГЛВ использует специальную схему стекинга, которая реализует связи между уровнями ГЛВ.

Следует отметить, что авторы работы [10] предложили сократить время обучения и тестирования с помощью механизма проверки достоверности, который ограничивает количество обучающих и тестирующих примеров, проходящих через уровни ГЛ. Это было сделано путем классификации примеров на каждом уровне ГЛ на два подмножества: одно состоит от примеров, которые классифицируются с высокой вероятностью правильной классификации, а другое подмножество – с малой вероятностью. Этот подход можно рассматривать как улучшение ГЛ. Следуя идее механизма доверительного скрининга, новая его модификация была предложена в [11]. Она основана на адаптивном взвешивании каждого обучающего примера на каждом уровне ГЛ в зависимости от распределения вероятностей его классов на предыдущем уровне ГЛ. Это модификация называется адаптивным взвешенным глубоким лесом (АВГЛ).

Основной целью представленной работы является расширение АВГЛ на случай анализа выживаемости путем введения специальной схемы назначения весов примерам и исследования их эффективности. Расширение называется адаптивным весовым глубоким лесом выживаемости (АВГЛВ). Основная идея, лежащая в основе предлагаемой модели, состоит в том, чтобы ввести и применить

Работа выполнена при финансовой поддержке РФФИ, проект №19-29-01004

маргинальный индекс согласованности (МС-индекс), который получен из хорошо известного С-индекса [12]. Множество числовых примеров с реальными данными иллюстрируют предлагаемый АВГЛВ.

II. Определения анализа выживаемости

Рассмотрим обучающее множество D, состоящее из n троек $(\mathbf{x}_i, \delta_i, T_i)$, $i = 1,...,n$, где каждая тройка характеризует пациента, $\mathbf{x}_i = (x_{i1},...,x_{im}) \in \mathbf{X} \subseteq \mathbf{R}^m$ – вектор признаков; $T_i \in \mathbf{R}_+$ – время смерти i-го пациента, предполагается, что оно не отрицательное и непрерывное; $\delta_i \in \{0,1\}$ – индикатор события, $\delta_i = 1$, если событие наблюдалось (не цензурированное наблюдение), $\delta_i = 0$, если соответствующее событие не наблюдалось (цензурированное наблюдение). Цель – оценить время до события T на основе D для нового пациента, имеющего вектор признаков $\mathbf{x}$.

Ключевыми понятиями в анализе выживаемости являются функции выживаемости и риска [1]. Функция выживаемости $S(t)$ есть вероятность того, что пациент выжил до момента времени t, т.е. $S(t) = \Pr\{T > t\}$. Функция риска $h(t)$ – плотность вероятности погибнуть в момент t при условии, что до него дожили. Функция вычисляется как $h(t) = f(t)/S(t)$, где $f(t)$ – плотность вероятности события. Оценивание функции $S(t)$ выполняется с использованием оценки Каплана-Мейера [1]. Важным понятием является кумулятивная функция риска (КФР) $H(t)$, которая определяется как интеграл от функции риска $h(t)$ и может интерпретироваться как вероятность события в момент времени t, при условии, что пациент дожил до момента времени t.

Существует несколько показателей для сравнения различных моделей выживаемости. Наиболее популярным показателем, характеризующим эффективность прогнозирования модели, является С-индекс [12]. Он оценивает, насколько правильно модель ранжирует времена жизни. Его также определяют как меру согласия между прогнозируемой и наблюдаемой функциями выживаемости. Для оценки С-индекса рассмотрим часть множества D, состоящего из допустимых пар $\{(\mathbf{x}_i, \delta_i, T_i), (\mathbf{x}_j, \delta_j, T_j)\}$ для $i \le j$. Пара является недопустимой, если оба события цензурируются справа или, если наименьшее время в паре является цензурированным. Тогда С-индекс – это отношение количества пар, правильно упорядоченных моделью, к общему количеству допустимых пар. Если С-индекс равен 1, то соответствующая модель выживаемости считается идеальной. Если это 0.5, то модель не лучше, чем случайное угадывание. Пусть $t_1,...,t_n$ – заранее определенные моменты времени. Если выход алгоритма – прогнозируемая функция выживаемости то С-индекс рассчитывается как [12]:

$$C = \frac{1}{M} \sum_{i\,:\,\delta_i=1} \sum_{j\,:\,t_i<t_j} \mathbf{1}\left[\hat{S}(t_i \mid \mathbf{x}_i) > \hat{S}(t_j \mid \mathbf{x}_j)\right].$$

Здесь M – число всех допустимых пар; $\mathbf{1}[a]$ – индикаторная функция, принимающая значение 1, если условие a выполняется, и 0, иначе.

III. Случайные леса выживаемости и глубокие леса выживаемости

СЛВ аналогичен регрессионному СЛ, но каждое дерево решений в СЛВ использует специальные правила расщепления, которые устанавливают процедуру деления подмножества точек обучающих данных на два подмножества. Существует несколько правил расщепления [12]. Алгоритм построения СЛВ приведен в [4]. В отличие от СЛ, выход которого – распределение вероятностей классов, СЛВ прогнозирует КФР, обозначенную $H_{RF}(t \mid \mathbf{x})$, которая рассчитывается путем усреднения оценок КФР, полученных всеми деревьями СЛВ.

Идея, лежащая в основе ГЛВ, аналогична идее обычного ГЛ. Архитектура ГЛ [8] может быть представлена набором каскадов ГЛ, так что каждый каскад содержит некоторое количество СЛ. Каскадная структура реализует идею обучения посредством послойной обработки векторов признаков. Каждый каскад (слой) каскадной структуры получает информацию об особенностях классификации на предыдущем каскаде и выдает результат своей обработки на следующий уровень.

Важной идеей, лежащей в основе ГЛ, является распределение вероятностей классов, создаваемое каждым деревом решений СЛ для каждого входного примера или каждого вектора признаков. Оценка распределения вероятностей классов выполняется путем подсчета доли обучающих примеров различных классов в конечном узле дерева, в который попадает соответствующий пример, и затем усреднения по всем деревьям в одном и том же лесу. Чтобы использовать результаты классификации на некотором уровне ГЛ, распределения вероятностей классов СЛ в качестве дополненных признаков объединяются с исходным вектором для его передачи на следующий уровень каскада. Важно отметить, что это объединение можно рассматривать как тип хорошо известной схемы стекинга. ГЛВ можно рассматривать как модификацию ГЛ, где случайные леса заменяются на СЛВ. В отличие от ГЛ, ГЛВ использует расширенные признаки специальной формы [10], а именно, эти признаки состоят из среднего времени a_i до события i-го объекта и $\nu - 1$ квантилей $t_i(p_1),...,t_i(p_{\nu-1})$, которые имеют значения

$$t_i(p_k) = \inf\{t \,:\, p_k \le 1 - S_f(t \mid \mathbf{x}_i)\},$$
$$p_k = k/\nu,\ k = 1,...,\nu - 1.$$

Таким образом, $\nu - 1$ квантилей и среднее время до события объединяются с исходным вектором признаков. Значение ν является параметром настройки, который получается с помощью процедуры кросс-валидации.

Следует отметить, что квантили в ГЛВ играют ту же роль, что и распределения вероятностей классов в ГЛ. Мы также используем среднее время, чтобы уменьшить смещение, которое может иметь место, когда число квантилей мало. С одной стороны, количество квантилей должно быть максимально большим, чтобы учесть всю информацию о выходе СЛВ. С другой стороны, расширеные признаки не должна маскировать свойства исходных примеров для правильного построения деревьев решений в СЛВ на следующем уровне каскада лесов.

IV. Адаптивный весовой глубокий лес выживаемости

Архитектура АВГЛВ показана на рис. 1. Ее можно рассматривать как расширение ГЛВ. Из рис. 1 видно, что АВГЛВ, как и ГЛВ, имеет несколько уровней. Каждый уровень состоит из нескольких СЛВ, которые генерируют расширенные признаки в форме средних значений и квантилей. Они показаны на рис. 1 тремя черными квадратиками для каждого СЛВ. Первая идея, лежащая в основе АВГЛВ, которая отличает его от ГЛВ, заключается в назначении веса каждому примеру $\mathbf{x}_i$ на текущем уровне каскада АВГЛВ в соответствии с результатами обучения, достигнутыми $\mathbf{x}_i$ на этом уровне. Если достигается высокая точность на уровне АВГЛВ, то пример $\mathbf{x}_i$ в этом случае не должен проходить следующий уровень АВГЛВ. Это можно сделать, присвоив ему вес, близкий к 0. В противном случае, когда для примера достигается низкая точность на уровне каскада АВГЛВ, этот пример должен участвовать в построении деревьев, и мы должны попытаться построить с ним СЛВ на следующем уровне, так что обучающий пример $\mathbf{x}_i$ будет обеспечивать более высокую точность прогнозирования. Это можно сделать, присвоив ему вес, близкий к 1. В целом, мы присваиваем вес, который можно определить как убывающую функцию показателя точности прогнозирования.

Следующий вопрос – как определить соответствующий показатель точности прогнозирования. Проблема в том, что самый популярный С-индекс не может быть использован здесь, потому что он характеризует полностью модель анализа выживаемости и не измеряет влияние каждого примера в отдельности. Поэтому необходимо ввести другой показатель, который может характеризовать каждый пример в отдельности. Одним из таких показателей может быть МС-индекс C_i в качестве нового показателя точности прогнозирования для примера $\mathbf{x}_i$, который определяется следующим образом. Если $\delta_i = 0$, тогда $C_i = 0$. Если $\delta_i = 1$, то

$$C_i = \frac{1}{M_i} \sum_{j\,:\,t_i<t_j} \mathbf{1}\left[S_f(t_i \mid \mathbf{x}_i) - S_f(t_j \mid \mathbf{x}_j) > 0 \right].$$

Здесь M_i – число допустимых пар для *i*-го пациента. Из приведенного выше выражения видно, что МС-индекс показывает, как предсказание, касающееся *i*-го пациента, согласуется со всеми пациентами, которые относятся к набору допустимых пар с *i*-ым пациентом. Фактически это мера качества прогноза для одного пациента.

Можно предложить множество функций для вычисления весов. Основным условием для этих функций является то, что они должны уменьшаться с показателем C_i. Мы предлагаем использовать простейшую функцию $w_i = 1 - C_i$, для упрощения вычислений. Кроме того, мы применяем стратегию использования весов, которая случайным образом извлекает примеры из обучающей выборки с возвращением в соответствии с распределением вероятностей, полученным из весов. Если вес *i*-го примера очень близок к 0, то примера не участвует в построении деревьев решений. Это означает, что механизм проверки достоверности, предложенный в работе [9] для сокращения времени обучения и тестирования посредством ограничения количества обучающих и тестирующих примеров, проходящих через уровни АВГЛВ, здесь частично реализован.

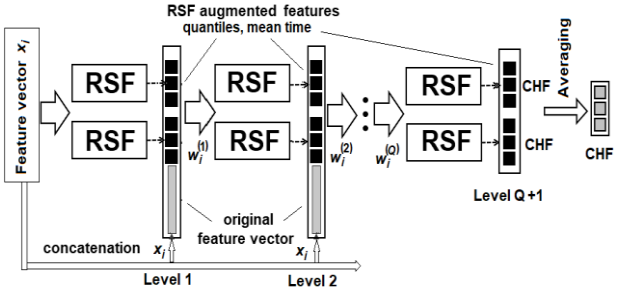

Рис. 1. Архитектура АВГЛВ

Вторая идея по улучшению вышеупомянутой взвешенной процедуры заключается в явном использовании идей механизма проверки достоверности непосредственно в АВГЛВ. В соответствии с механизмом проверки достоверности экземпляр перемещается на следующий уровень каскада, только если определено, что он требует более высокого уровня обучения; в противном случае прогнозируется использование модели на текущем уровне [9]. Введем порог η_q, чтобы сравнить его со значением $1 - w_i$. Если $1 - w_i \geq \eta_q$, то соответствующий пример прогнозируется с использованием модели на текущем уровне, и ему не нужно проходить на следующий уровень, в противном случае пример должен пройти на следующий уровень. В итоге, количество примеров для обучения сокращается на каждом уровне, что упрощает весь процесс обучения. В результате мы имеем комбинацию весов и пороговой процедуры, применяемой к АВГЛВ для анализа выживаемости. Этот подход используется на этапе тестирования таким же образом. В этом случае МС-индекс определяется для тестового примера с учетом всех допустимых обучающих примеров.

V. Численные эксперименты

Чтобы исследовать предложенный АВГЛВ и сравнить его с ГЛВ, мы используем следующие открытые данные:

1) German Breast Cancer Study Group 2 (GBSG2) содержит данные о 686 пациентах (пакет TH.data R).

2) Lupus Nephritis Dataset (LND) содержит данные о 87 пациентах

АВГЛВ реализован с использованием Python. Для оценки С-индекса используется кросс-валидация, так что 80 % данных берут для обучения, а 20 % – для тестирования.

Результаты тестирования для данных LND показаны на рис. 2, где значения С-индекса изображены в зависимости от номера каскада и порога η_q, который принимает значения 0.5, 0.65, 0.75, 0.85. Из рис. 2 видно, что порог 0.75 позволяет улучшить прогнозирование на 6% для LND. При этом три уровня АВГЛВ обеспечивают наилучшие результаты. Порог 0.65 также позволяет нам улучшить прогнозировани, но это улучшение наблюдается на пятом уровне каскада, что требует дополнительного времени для обучения. Интересно отметить, что введение весов и порога может привести к худшим результатам. Из рис. 2 видно, что порог 0.5 делает С-индекс слишком малым на четвертом уровне каскада. Причиной ухудшения является то, что слишком много примеров удаляются из обучения, а СЛВ на следующих уровнях обучаются только на «плохих» примерах.

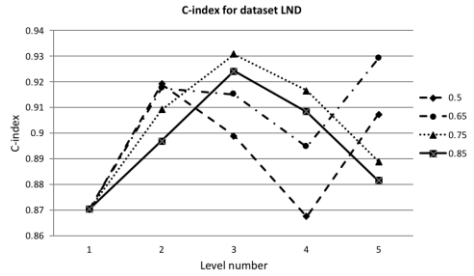

Рис. 2. С-индекс как функция номера каскада и порога для LND

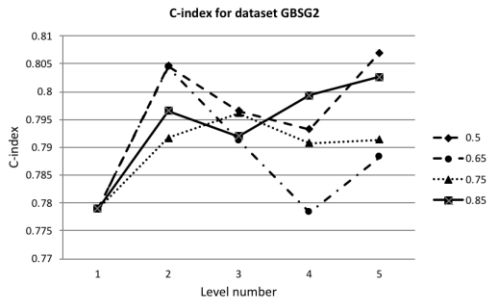

Рис. 3. С-индекс как функция номера каскада и порога для GBSG2

Аналогичные результаты показаны на рис. 3 для набора данных GBSG2. Из рис. 3 видно, что наилучшие результаты получены для порога 0.5. При этом пять уровней АВГЛВ обеспечивают наилучшие результаты для рассматриваемого набора данных.

Эксперименты предполагают, что введение весов и схемы адаптации может привести к лучшим результатам по сравнению с обычным ГЛВ. Более того, из численных результатов видно, что невозможно предсказать оптимальные значения количества уровней и порога. Они сильно зависят от наборов данных и могут быть получены только путем проведения экспериментов.

Следует отметить, что основной целью численных экспериментов является демонстрация того, что введение адаптивных весов и пороговых значений может привести к превосходящим результатам.

VI. Заключение

В статье была предложена новая модель выживаемости на основе ГЛВ. Основные идеи, лежащие в основе модели, заключаются в том, чтобы модифицировать ГЛ путем замены СЛ на СЛВ и назначить веса примеров в зависимости от качества их прогнозирования на каждом уровне ГЛ. Алгоритм сокращает время на обучение и тестирование ГЛВ и позволяет нам получить лучшие результаты анализа выживаемости.

Однако узким местом алгоритма является достаточно сложное вычисление МС-индекса. Следовательно, проблема поиска нового показателя для каждого примера, имеющего простой алгоритм его вычисления, является направлением для дальнейших исследований. Мы изучили только одну функцию, определяющую веса примеров. Другим направлением дальнейших исследований является изучение различных функций, которые могут привести к лучшим результатам прогнозирования.

Список литературы

[1] Hosmer D., Lemeshow S., May S. Applied Survival Analysis: Regression Modeling of Time to Event Data. John Wiley & Sons, New Jersey, 2008.

[2] Cox D. Regression models and life-tables // Journal of the Royal Statistical Society, Series B (Methodological). 1972. vol. 34, pp. 187-220.

[3] Breiman L. Random forests // Machine learnig. 2001. vol. 45, pp. 5-32

[4] Ishwaran H., Kogalur U. Random survival forests for R // R News. 2007. vol. 7., pp. 25-31.

[5] Biau G., Scornet E. A random forest guided tour // Test. 2016. vol. 25, pp. 197-227.

[6] Bou-Hamad I., Larocque D., Ben-Ameur H. Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data // Statistical Modelling. 2011. vol. 11, pp. 429-446.

[7] Wang, H., Zhou, L. Random survival forest with space extensions for censored data // Artificial intelligence in medicine. 2017. vol. 79, pp. 52-61.

[8] Zhou Z.-H., Feng J. Deep forest: Towards an alternative to deep neural networks, in Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017, pp. 3553-3559.

[9] Pang M., Ting K., Zhao P., Zhou Z.H. Improving deep forest by confidence screening, in Proceedings of the 18th IEEE International Conference on Data Mining. Singapore. 2018. pp. 1-6.

[10] Utkin L., Konstantinov A., Meldo A., Ryabinin M., Chukanov V. A deep forest improvement by using weighted schemes, in Proceedings of the 24th Conference of Open Innovations Association FRUCT. Moscow, Russia, IEEE, 2019. pp. 451-456.

[11] Harrell F., Califf R., Pryor D., Lee K., Rosati R. Evaluating the yield of medical tests // Journal of the American Medical Association. 1982. vol. 247, pp. 2543-2546.

[12] Wang P., Li Y., Reddy C. Machine learning for survival analysis: A survey // ACM Computing Surveys. 2019. vol. 51, pp. 1-36.