# Интеллектуальная система выбора лечения на основе каскада случайных лесов в рамках анализа выживаемости

Л. В. Уткин[1], М. А. Рябинин[2]
Санкт-Петербургский политехнический университет Петра Великого
[1]lev.utkin@gmail.com, [2]mihail-ryabinin@yandex.ru

А. А. Мелдо
Санкт-Петербургский клинический научно-практический центр специализированных видов медицинской помощи (онкологический)
anna.meldo@yandex.ru

***Аннотация*. В работе предлагается новая архитектура рекомендательной интеллектуальной системы, позволяющей выбирать персонифицированное оптимальное лечение для пациента на основе его медицинских показателей, с точки зрения минимизации функции риска. В основе предлагаемого подхода для реализации системы лежит математический аппарат статистического анализа выживаемости. Вычисление функции риска осуществляется при помощи каскада случайных лесов выживаемости, каждый из которых реализует расширение известной модели пропорционального риска Кокса. Случайные леса выживаемости обучаются с использованием данных о пациентах и отличаются от обычных регрессионных случайных лесов специфической функцией расщепления, максимизирующей разность показателей выживаемости двух расщепленных подмножеств обучающих данных. Особенностью предлагаемой архитектуры является организация каскадов случайных лесов, которые можно рассматривать как одну из модификаций глубокого леса.**

***Ключевые слова: анализ выживаемости; случайный лес; искусственный интеллект; функция риска; пациент; модель Кокса***

## I. Введение

Большое число систем диагностирования заболеваний было разработано для обеспечения успешного обнаружения заболеваний и для поддержки принятия решения о выборе метода лечения. Однако только малая часть таких систем учитывает аспект выживаемости пациентов, особенно онкологических больных. Большое количество данных собрано о пациентах и их характеристиках, их заболеваниях в медицинских учреждениях. Одной из важнейших задач является использование этих данных для создания систем, которые бы помогли врачу выбирать оптимальное лечение для каждого пациента на основе его состояния и симптомов болезни. Так правильное и раннее диагностирование рака может сохранить жизнь пациента.

Основой для таких систем диагностирования может являться анализ выживаемости, используемый в самых различных областях. Одна из наиболее важных областей – область медицинских исследований, где модели выживаемости применяются для оценки значимости прогностических факторов возникновения таких событий, как смерть или рецидив рака, и для информирования о выборе лечения [14]. Наборы данных, используемые в анализе выживаемости, отличаются от обычных наборов данных тем, что время наступления события, например, смерти, для части пациентов неизвестно, так как это событие может еще не наступить в течении периода наблюдений. Если наблюдаемое время жизни меньше действительного времени жизни, то имеем частный случай цензурированных справа данных.

Модели выживаемости можно разделить на три группы: параметрические, непараметрические и полупараметрические. В параметрических моделях предполагается, что вид закона распределения вероятностей времени жизни известен, например, экспоненциальное распределение, распределение Вейбулла или гамма распределение. К одной из простейших непараметрических моделей относится оценка Каплана-Мейера, которая используется для построения распределения вероятностей времени жизни по однородным данным, т.е. модель не учитывает то, что пациенты различаются по своим характеристикам.

Популярной регрессионной моделью анализа выживаемости является полупараметрическая модель Кокса пропорционального риска [4]. Предположение пропорционального риска в модели Кокса означает, что различные пациенты имеют функции риска, которые пропорциональны друг другу, т.е. отношение функций риска для двух пациентов с различными признаками является постоянной величиной и не зависит от времени. Модель Кокса является достаточно эффективным методом обработки данных выживаемости. Следствием этого является большое количество подходов для работы с моделью и ее модификаций, предложенных в последние десятилетия. Структура методов анализа выживаемости и их детальный обзор представлен в работе [21].

Исследование выполнено за счет гранта Российского научного фонда (проект № 18-11-00078)

Следует отметить, что модель Кокса может приводить к неудовлетворительным результатам при большой размерности данных и малом количестве наблюдений, например, в задачах анализа данных экспрессии генов. Для частичного устранения этого недостатка в работе [18] была предложена модификация модели Кокса на основе метода Лассо. Другой недостаток модели Кокса заключается в предположении линейной зависимости между признаками и временем наступления события. Различные модификации были предложены, учитывающие нелинейный характер этой зависимости. Так в работе [6] был представлен подход, использующий простую нейронную сеть, которая является основой для модели нелинейного пропорционального риска. Модель стала также базовой для создания моделей выживаемости на основе глубоких нейронных сетей [14], [17]. Часть таких моделей рассмотрена в обзоре [21]. Однако использование нейронных сетей требует наличия большого количества данных для обучения. Поэтому в работе [20] было предложено использовать метод опорных векторов для улучшения модели при малом объеме обучающих данных.

Другим подходом к анализу цензурированных данных является использование деревьев и случайных лесов выживаемости. Благодаря большому числу достоинств деревьев решений как инструмента классификации и регрессии, разработан ряд модификаций, решающих задачи анализа цензурированных данных [3], [7], [15]. Детальный обзор деревьев и случайных лесов выживаемости представлен в работе [1]. Случайные леса [2] были разработаны в целях устранения недостатков деревьев решений. Оказалось, что случайные леса также стали одним из наиболее эффективных и популярных инструментов анализа выживаемости. Популярность случайных лесов выживаемости обусловлена целым рядом факторов. Прежде всего, как отмечено в работе [12], случайные леса требуют определения только трех настраиваемых параметров: числа случайно выбираемых признаков, числа деревьев в лесу и правила расщепления.

Учитывая преимущества случайных лесов, большое число моделей на их основе было разработано для работы с цензурированными данными [1], [10], [11]. Большинство моделей различаются только критериями расщепления и правилами композиции отдельных оценок. Основная часть моделей случайных лесов выживаемости используют усреднение оценок накопленных рисков по всем деревьям.

Так как случайный лес выживаемости является одним из наиболее эффективных моделей в анализе выживаемости, то в работе уделяется основное внимание этой модели и предлагается подход для ее улучшения. Основная идея, лежащая в основе этого подхода, заключается в построении каскада случайных лесов выживаемости, который можно рассматривать как частный случай глубокого леса, предложенного в [23]. Фактически в представленной работе разрабатывается глубокий лес выживаемости. Ключевым моментом предлагаемой каскадной структуры является алгоритм стекинга. Изначально стекинг [22] был разработан для реализации идеи того, что каждый последующий слой многослойной структуры может обнаруживать и корректировать ошибки обучения, появляющиеся в предыдущих слоях. Поэтому в предлагаемой работе также предлагается модификация алгоритма стекинга, учитывающая особенности случайного леса выживаемости.

## II. Модель Кокса

В анализе выживаемости пациент $i$ представлен тройкой $(\mathbf{x}_i, \delta_i, T_i)$, где $\mathbf{x}_i = (x_{i1}, ..., x_{im})$ – вектор характеристик (признаков) пациента; $T_i$ – время до появления события (смерти пациента). Если событие наблюдается, то $T_i$ – время между некоторой точкой отсчета и наступлением события. В этом случае индикатор события равен $\delta_i = 1$, и имеет место нецензурированное наблюдение. Если событие не наблюдается, то $T_i$ – время между точкой отсчета и концом наблюдения, и $\delta_i = 0$. Это – цензурированное наблюдение. Пусть обучающее множество $D$ состоит из $n$ троек $(\mathbf{x}_i, \delta_i, T_i)$, $i = 1, ..., n$. Цель анализа выживаемости заключается в оценке времени $T$ до события для нового пациента, характеризуемого вектором признаков $\mathbf{x}$ с использованием $D$.

Функции выживаемости и риска являются ключевыми элементами анализа выживаемости. Функция выживаемости $S(t)$ – это вероятность выживания до этого момента времени $t$, т.е., $S(t) = \Pr\{T > t\}$. Функция риска $h(t)$ - это вероятность погибнуть в момент $t$ при условии, что до него дожили. Функция выживаемости определяется через функцию риска следующим образом:

$$S(t) = \exp\left(-\int_0^t h(x)\mathbf{d}x\right).$$

В соответствии с моделью Кокса [4],[9], функция риска при заданном векторе признаков $\mathbf{x}$ определяется как

$$h(t \mid \mathbf{x}) = h_0(t) \cdot \Psi(\mathbf{x}, \mathbf{b}) = h_0(t) \cdot \exp\left(\psi(\mathbf{x}, \mathbf{b})\right).$$

Здесь $h_0(t)$ – базовая функция риска; $\Psi(\mathbf{x})$ – функция риска, не зависящая от времени; $\mathbf{b} = (b_1, ..., b_m)$ – вектор параметров регрессии. Функция $\psi(\mathbf{x}, \mathbf{b})$ является линейной, т.е. $\psi(\mathbf{x}, \mathbf{b}) = \mathbf{x}\mathbf{b}^{\mathbf{T}}$.

Идея, лежащая в основе использования нейронных сетей в анализе выживаемости, заключается в замене линейной функции $\psi(\mathbf{x})$ нелинейной, реализованной нейронной сетью [6].

Так как пациенты подвержены различному уровню риска в соответствии с их характеристиками и их лечением, то интересно выбрать наилучшее лечение. Рассмотрим подход, предложенный в работе [14]. Предположим, что все пациенты принадлежат одной из $s$ групп лечения $\tau \in \{1, ..., s\}$. Также предполагается, что каждый вид лечения $i$ имеет независимую функцию риска $\exp(\psi_i(\mathbf{x}))$, и базовая функция риска $h_0(t)$ одинакова для всех видов лечения. Один из способов определения наилучшего вида лечения заключается в вычислении рекомендательной функции, которая определяется как

$$rec_{ij}(\mathbf{x}) = \log\left(\frac{h(t; \mathbf{x} \mid \tau = i)}{h(t; \mathbf{x} \mid \tau = j)}\right) = \psi_i(\mathbf{x}) - \psi_j(\mathbf{x}).$$

Для обеспечения персональных рекомендаций по лечению в соответствии с этой функцией вычисляются $\psi_i(\mathbf{x})$ и $\psi_j(\mathbf{x})$ для разных групп лечения. Если полученная функция $rec_{ij}(\mathbf{x})$ положительна, то лечение *j* предпочтительнее по сравнению с лечением *i*. В случае отрицательной функции лечение *i* является более эффективным и ведет к меньшему риску, чем лечение *j*.

Для сравнения моделей выживаемости используется C-индекс, предложенный в работе [8]. Он оценивает, насколько хороша модель при ранжировании времени выживания. Фактически, это вероятность того, что времена событий пары пациентов корректно ранжируется. Рассмотрим допустимые пары $\{(\mathbf{x}_i,\delta_i,T_i),(\mathbf{x}_j,\delta_j,T_j)\}$ для $i \le j$ в *D*. C-индекс рассчитывается как отношение числа пар, корректно упорядоченных моделью, к общему числу допустимых пар *M*. Пара не является допустимой, если самое раннее время в паре является цензурированным. Если C-индекс равен 1, то соответствующая модель выживаемости является идеальной. Если C-индекс равен 0,5, то модель не лучше, чем случайное угадывание. Пусть $t_1^*,...,t_q^*$ – заранее определенные моменты времени, например, $t_1,...,t_N$, где *N* – число различных событий. Если выход алгоритма – прогнозируемая функция выживаемости $S(t)$, то C-индекс рассчитывается как [21]:

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

Здесь $\mathbf{1}[a]$ – индикаторная функция, принимающая значение 1, если условие *a* выполняется, и 0 иначе; *S* – оцененная функция выживаемости.

## III. Случайные леса выживаемости

Было отмечено, что случайный лес выживаемости является одной из лучших моделей анализа выживаемости благодаря его свойствам. Это основная причина его модификации для улучшения результатов анализа выживаемости и повышения точности прогнозирования.

Общий алгоритм построения случайных лесов выживаемости имеет следующий вид [13]:

1. Случайным образом выбирается Q подмножеств из обучающих данных. Каждый выбор исключает 37% данных, называемых данными вне пакета (данные OOB).
2. Строится дерево выживаемости для каждого выбора. На каждом узле дерева случайно выбираются $m^{1/2}$ признаков. Узел расщепляется с использованием признака, который максимизирует разницу значений выживаемости между дочерними узлами.
3. Дерево строится так, что листья должны иметь не менее, чем $d > 0$ различных событий (смертей).
4. Вычисляется накопленная функция риска для каждого дерева и среднее значение функции для леса.
5. Используя данные OOB, вычисляется ошибка прогнозирования для усредненной функции риска.

Важным вопросом случайных лесов выживаемости, который определяет их различные реализации, является правило расщепления. Как показано в работе [13], оптимальное расщепление максимизирует разницу в выживаемости по двум наборам данных. Существует много правил расщепления, но ниже приводятся три основные правила: (1) правило, основанное на логранговом критерии; (2) сохранение расщепленных событий; (3) приближенное разбиение по логранговому критерию. У каждого правила есть плюсы и минусы. Подробный обзор правил можно найти в работах [13], [21].

Пусть $\{t_{j,k}\}$ – времена смерти в вершине *k* для *q*-го дерева, $Z_{j,k}$ и $Y_{j,k}$ – значения числа смертей и пациентов, доживших до момента $t_{j,k}$. Накопленная функция риска для вершины *k* определяется как $H_k(t) = \sum_{t_{j,k}\le t} Z_{j,k}/Y_{j,k}$.

Если *i*-ый пациент с признаками $\mathbf{x}_i$ попадает в вершину *k*, то можно сказать, что $H(t\,|\,\mathbf{x}_i) = H_k(t)$. Накопленная функция риска всего леса для *i*-го пациента получается усреднением соответствующих функции по всем *Q* деревьям, т.е.

$$H_{\text{forest}}(t\,|\,\mathbf{x}_i) = \frac{1}{Q}\sum_{q=1}^{Q} H_q(t\,|\,\mathbf{x}_i).$$

Функция выживаемости может быть получена накопленной функции риска. Другая оценка композиции рассмотрена в. [13], где используются данные OOB.

## IV. Каскад случайных лесов выживаемости

Рассмотрим каскад случайных лесов как частный случай глубокого леса, предложенный в [23]. Глубокий лес представляет собой каскадную структуру, где каждый уровень каскада получает информацию о характеристиках предыдущего уровня и выводит результат обработки на следующий уровень. Предполагаем, что количество уровней в каскаде $K$.

Одной из важных идей, лежащих в основе структуры каскадного леса, является конкатенация выхода леса с каждого уровня каскада с исходным вектором для использования на следующем уровне. Эту идею можно рассматривать как реализацию метода стекинга [22]. В отличие от стандартного алгоритма стекинга, глубокий лес использует исходный вектор и выход на следующем каскадном уровне посредством их конкатенации. Таким образом, вектор признаков увеличивается после каждого каскадного уровня. Основная сложность для реализации алгоритма стекинга заключается в том, чтобы выбрать некоторое представление о выходе случайного леса

выживаемости. Это представление должно быть компактным и информативным одновременно.

Выходом случайного леса выживаемости является функция риска или функция выживаемости. Мы не можем делать конкатенацию ее с исходным вектором $\mathbf{x}_i$ для реализации схемы стекинга, так как большое число расширенных признаков будет маскировать исходные данные. Поэтому предлагается добавлять только два признака. Первый признак – среднее время $m_i$ до события *i*-го пациента, которое просто вычисляется интегрированием функции выживаемости. Второй признак называется маргинальным C-индексом *i*-го пациента и обозначается $C_i$. Этот индекс определяется следующим образом. Если $\delta_i = 0$, то $C_i = 0$. Если $\delta_i = 1$, то

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

Здесь $M_i$ - число допустимых пар для *i*-го пациента. Из приведенного выше выражения видно, что маргинальный C-индекс показывает, как прогноз относительно *i*-го пациента согласуется с прогнозами по всем пациентам, которые являются допустимыми парами с *i*-м пациентом. Фактически, это показатель качества прогнозирования одного пациента. В итоге, вход следующего слоя лесного каскада можно представить следующим образом:

$$\mathbf{x}_i^{k+1} \leftarrow (\mathbf{x}_i^k, m_i^k, C_i^k),\ k = 1, ..., K.$$

Верхний индекс *k* означает номер слоя каскада.

Для нового пациента с признаками $\mathbf{x}$, вычисляется функция риска на последнем слое обученного каскада. Используя выражение для рекомендательной функции $rec_{ij}(\mathbf{x})$, получим рекомендации по персональному лечению нового пациента. Теперь используем неупрощенное выражение для $rec_{ij}(\mathbf{x})$, т.е.

$$rec_{ij}(\mathbf{x}) = \log\left( \frac{h(t;\mathbf{x} \mid \tau = i)}{h(t;\mathbf{x} \mid \tau = j)} \right).$$

Функция риска $h(t;\mathbf{x} \mid \tau = i)$ является производной функции $H_{\text{forest}}(t \mid \mathbf{x}^K)$, полученной для группы $\tau = i$.

## V. Заключение

В работе представлена интеллектуальная система рекомендаций, которая позволяет выбрать оптимальное лечение для пациента на основе применения каскада случайных лесов выживаемости. Основной вклад заключается в том, что предлагается новый алгоритм стекинга в каскаде случайных лесов. Он состоит в конкатенации оригинального вектор признаков пациента и двух признаков, полученных на выходе предыдущего слоя лесов. Предлагаемую реализацию системы можно рассматривать как первую попытку построить глубокий лес выживаемости. Чтобы улучшить предлагаемую структуру, можно также использовать дополнительную процедуру, которая оптимальным образом назначает веса деревьям в каждом лесу, аналогично той же процедуре, что была реализована в работе [19]. Однако это является направлением для дальнейших исследований.

## Список литературы


[1] Bou-Hamad I., Larocque D., and Ben-Ameur H. A review of survival trees // Statistics Surveys, 5:44-71, 2011.

[2] Breiman L. Random forests // Machine learning, 45(1):5-32, 2001.

[3] Ciampi A. Generalized regression trees // Computational Statistics & Data Analysis, 12:57-78, 1991.

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

[5] Devarajn K. and Ebrahimi N. A semi-parametric generalization of the cox proportional hazards regression model: Inference and applications // Computational Statistics & Data Analysis, 55(1):667-676, 2011.

[6] Faraggi D. and Simon R. A neural network model for survival data // Statistics in medicine, 14(1):73-82, 1995.

[7] Gordon L. and Olshen R.A. Tree-structured survival analysis // Cancer treatment reports, 69(10):1065-1069, 1985.

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

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

[10] Hu C. and Steingrimsson J.A. Personalized risk prediction in clinical oncology research: Applications and practical issues using survival trees and random forests // Journal of Biopharmaceutical Statistics, 28(2):333-349, 2018.

[11] Ishwaran H., Blackstone E.H., Pothier C.E., and Lauer M.S. Relative risk forests for exercise heart rate recovery as a predictor of mortality // Journal of the American Statistical Association, 99:591-600, 2004.

[12] Ishwaran H. and Kogalur U.B. Random survival forests for R // R News, 7(2):25-31, 2007.

[13] Ishwaran H., Kogalur U.B., Blackstone E.H., and Lauer M.S. Random survival forests // Annals of Applied Statistics, 2:841-860, 2008.

[14] Katzman J.L., Shaham U., Cloninger A., Bates J., Jiang T., and Kluger Y. Deepsurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network // BMC medical research methodology, 18(24):1-12, 2018.

[15] LeBlanc M. and Crowley J. Relative risk trees for censored survival data. Biometrics // 48(2):411-425, 1992.

[16] Lee E.T. and Wang J.W. Statistical Methods for Survival Data Analysis / John Wiley & Sons, New Jersey, 2003.

[17] Nezhad M.Z., Sadati N., Yang K., and Zhu D. A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer // arXiv:1804.03280, April 2018.

[18] Tibshirani R. The lasso method for variable selection in the cox model // Statistics in medicine, 16(4):385-395, 1997.

[19] Utkin L.V. and Ryabinin M.A. A Siamese deep forest // Knowledge-Based Systems, 139:13-22, 2018.

[20] Van Belle V., Pelckmans K., Suykens J.A.K, and Van Huffel S. Support vector machines for survival analysis // Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007), pages 1-8, 2007.

[21] Wang P., Li Y., and Reddy C.K. Machine learning for survival analysis: A survey // arXiv:1708.04649, August 2017.

[22] Wolpert D.H. Stacked generalization // Neural networks, 5(2):241-259, 1992.

[23] Zhou Z.-H. and Feng J. Deep forest: Towards an alternative to deep neural networks // arXiv:1702.08835v2, May 2017.