УДК 004.85

ПОДХОДЫ К ИНТЕРПРЕТАЦИИ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ПРИ ЦЕНЗУРИРОВАННЫХ ДАННЫХ

Л.В. Уткин
Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, Санкт-Петербург, Политехническая ул., 29
E-mail: lev.utkin@gmail.com

А.В. Константинов
Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, Санкт-Петербург, Политехническая ул., 29
E-mail: andrue.konst@gmail.com

Д.Ю. Еременко
Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, Санкт-Петербург, Политехническая ул., 29
E-mail: danilaeremenko@mail.ru

В.С. Заборовский
Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, Санкт-Петербург, Политехническая ул., 29
E-mail: vlad2tu@yandex.ru

Ключевые слова: машинное обучение, модель выживаемости, объяснительный интеллект, цензурированные данные, модель Кокса, оценка Берана.

Аннотация: Методы интерпретации получили широкое распространение для объяснения, почему модель машинного обучения типа «черный ящик» выдает то или иное предсказание. В работе рассматриваются основные методы интерпретации моделей выживаемости, которые оперируют с цензурированными данными и определяют характеристики времени до определенного события. Особенностью таких моделей является то, что их предсказания представляются в виде функции выживаемости или функции риска. Это требует разработки специальных методов интерпретации. Рассмотрены наиболее известные методы SurvLIME, SurvLIME-KS, SurvNAM и SurvBeX, которые основаны на использовании известного метода интерпретации предсказаний LIME, модели Кокса и ее модификации, а также оценки Берана.

1. Введение

Модели машинного обучения, обученные на данных, характеризующих времена до наступления определенных событий некоторых объектов в зависимости от структуры этих объектов, приобретают все большее распространение [1]. Это обусловлено их использованием в самых различных областях, например, в надежности систем, когда рассматриваются события отказа системы, в медицине, когда событием является выздоровление или смерть пациента. Особенностью многих моделей является то, что событие для ряда объектов может и не наблюдаться, а фиксируется только последний момент наблюдения в предположении, что событие когда-либо произойдет в будущем, но мы не знаем точно когда. Такие данные называются цензурированными, и они

содержат существенно меньшую информация об объекте, чем нецензурированные данные, для которых известны времена наступления события. Тем не менее, цензурированные данные также могут быть использованы в моделях машинного обучения, которые будем называть моделями выживаемости (survival models).

Одной из известных моделей выживаемости является модель пропорционального риска Кокса [2], особенность которой заключается в том, что атрибуты (признаки) объекта связаны линейной зависимостью. Это является ограничением модели, так как в ряде случаев взаимосвязь может быть существенно нелинейной. В последнее время разработано достаточно большое количество моделей выживаемости, которые учитывают различные особенности данных, например, случайные леса выживаемости, глубокие нейронные сети выживаемости, модификации метода опорных векторов и другие. В то же время, каждая из этих моделей представляет собой черный ящик, то есть структуру для которой известен только вход и выход (предсказание), но не известно, почему получен такой выход, какие атрибуты объекта повлияли на предсказание модели, например, показатели каких датчиков двигателя наибольшим образом повлияли на предсказание, в соответствии с которым среднее время до его отказ 10 часов. Эта задача относится к классу задач интерпретации или объяснения предсказаний модели машинного обучения, решению которых посвящено большое количество публикаций [3,4]. Однако особенностью моделей выживаемости является то, что предсказания большинства из них представлены в виде функции выживаемости либо кумулятивной функции риска. Это существенно затрудняет решение задачи объяснения и требует специальных подходов к ее решению.

В работе рассмотрен один из наиболее популярных методов интерпретации моделей LIME [5] и представлены основные методы интерпретации предсказаний моделей выживаемости, разработанных на основе этого метода с использованием модели Кокса и оценке Берана [6].

2. Элементы анализа выживаемости

Рассмотрим обучающее множество D, состоящее из n троек $(\boldsymbol{x}_i, T_i, \delta_i)$, $i = 1, \dots, n$, где каждая тройка характеризует объект, $\boldsymbol{x}_i \in \mathbf{R}^m$ – вектор атрибутов или признаков; T_i – время до события i-го объекта, δ_i – индикатор события, $\delta_i = 1$, если событие наблюдалось (нецензурированное наблюдение), $\delta_i = 0$, если событие не наблюдалось (цензурированное наблюдение). Цель – оценить время до события T на основе D для нового объекта, имеющего вектор атрибутов $\boldsymbol{x}$.

Ключевыми понятиями в анализе выживаемости являются функции выживаемости и риска [1]. Функция выживаемости $S(t|\boldsymbol{x})$ есть вероятность того, что событие, связанное с объектом $\boldsymbol{x}$, не произошло до момента времени t. Важным понятием является кумулятивная функция риска $H(t|\boldsymbol{x})$, которая выражается через функцию выживаемости $H(t|\boldsymbol{x}) = -\ln(S(t|\boldsymbol{x}))$.

Одной из фундаментальных моделей выживаемости является модель Кокса [2], для которой кумулятивная функция риска определяется как

$$H(t|\boldsymbol{x}, \boldsymbol{b}) = H_0(t) \cdot \exp(\boldsymbol{b}^T \boldsymbol{x}),$$

где $H_0(t)$ – базовая кумулятивная функция риска, определяемая, например, как оценка Каплана-Мейера или Нельсона-Аалена, $\boldsymbol{b}$ – вектор параметров модели.

Главная особенность модели Кокса заключается в том, что функция риска или функция выживаемости зависят от линейной комбинации атрибутов, что позволит в дальнейшем использовать эту модель для объяснения предсказаний моделей выживаемости. Однако модель Кокса не учитывает взаимных расстояний между

точками в обучающей выборке. Эту проблему решает оценка Берана [6], в соответствии с которой функция выживаемости определяется следующим образом:

$$S_B(t|\boldsymbol{x}) = \prod_{T_i \le t} \left\{ 1 - \frac{W(\boldsymbol{x}, \boldsymbol{x}_i)}{1 - \sum_{j=1}^{i-1} W(\boldsymbol{x}, \boldsymbol{x}_j)} \right\}^{\boldsymbol{\delta}_i}.$$

Здесь вес $W(\boldsymbol{x}, \boldsymbol{x}_i)$ характеризует близость векторов $\boldsymbol{x}, \boldsymbol{x}_i$. Например, вес может определяться через операцию softmax: $W(\boldsymbol{x}, \boldsymbol{x}_i) = \text{softmax}(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2/\tau)$, где τ – параметр оценки. Интересно отметить, что модель Каплана-Мейера является частным случаем оценки Берана, когда все веса одинаковы и равны $1/n$.

3. Формальная постановка и решение задачи объяснения

Формально задача объяснения решается при помощи обучения мета-модели или модели объяснения, которая аппроксимирует основную модель «черного ящика» в окрестности объясняемого примера и принадлежит множеству «простых» моделей, которые являются интерпретируемыми (линейные модели, деревья решений). Основная модель реализует функцию $f: \mathbf{R}^m \to \mathbf{R}^d$, например, в классификации $f(\boldsymbol{x})$ – вероятность (или индикатор) того, что $\boldsymbol{x}$ принадлежит определенному классу. Мета-модель – это модель $g \in G$, где G - класс интерпретируемых моделей, которая является решением задачи оптимизации:

$$\min_{g \in G} \{L(f, g, \theta) + \Omega(g)\},$$

где $L(f, g, \theta)$ – мера того, насколько g плохо аппроксимирует f; θ – вектор параметров; $\Omega(g)$ – регуляризатор.

Одна из наиболее популярных интерпретируемых функций – линейная функция. Это связано с тем, что коэффициенты линейной функции как раз характеризуют влияние каждого атрибута на значение функции. Фактически локальная задача объяснения (объяснение одного примера или объекта) сводится к аппроксимации функции $f(\boldsymbol{x})$ основной модели линейной функцией $g(\boldsymbol{x})$ в точке $\boldsymbol{x}$. На этой идее как раз основан известный метод LIME [5] и его модификации. В LIME для построения аппроксимирующей функции $g(\boldsymbol{x})$ заданное количество N точек (векторов $\boldsymbol{z}_i \in \mathbf{R}^m$) генерируются в окрестности точки $\boldsymbol{x}$. Используя основную модель, находится предсказание $y_i = f(\boldsymbol{z}_i)$ для каждой сгенерированной точки $\boldsymbol{z}_i$ и образуется новый датасет из N точек $(\boldsymbol{z}_i, y_i)$. Полученный датасет используется для построения линейной функции $g(\boldsymbol{x})$. Метод LIME применяется для объяснения моделей классификации и регрессии. Однако его применение для моделей выживаемости представляет определенные сложности, так как, во-первых, модели выживаемости имеют дело с цензурированными данными, для которых построение регрессионной модели отличается от стандартных моделей. Во-вторых, выходом модели выживаемости, то есть y, является функция выживаемости, а не точечное значение, что также усложняет задачу объяснения, так как необходимо интерпретировать функцию выживаемости.

4. Методы объяснения предсказаний моделей выживаемости

Основная идея ряда методов объяснения, называемых SurvLIME, SurvLIME-KS, SurvLIME-Inf, заключается в применении модели Кокса для аппроксимации «черного ящика» в локальной окрестности вокруг объясняемого объекта. В модели Кокса используется линейная комбинация $\boldsymbol{b}^T\boldsymbol{x}$ атрибутов объекта. При этом важно, что атрибуты, как и их комбинация, не зависят от времени. Следовательно, коэффициенты $\boldsymbol{b}$ можно рассматривать как меру влияния соответствующих атрибутов на

предсказание. Однако мы аппроксимируем не точечное предсказание, а функции, например, кумулятивную функцию риска. В соответствии с методом SurvLIME [7] вокруг объяснимого примера случайным образом генерируются синтетические точки (векторы) $\boldsymbol{z}_i$, и для каждого синтетического вектора рассчитывается функция риска с помощью основной модели или «черного ящика». Так как в модели Кокса функция риска (выживаемости) есть функция неизвестных коэффициентов $\boldsymbol{b}$, то задача оптимизации для вычисления коэффициентов определяется весовым средним расстоянием между функциями выживаемости «черного ящика» и модели Кокса по всем сгенерированным точкам $\boldsymbol{z}_i, i = 1, \dots, N$, с учетом весов, которые определяются расстояниями между точкой $\boldsymbol{x}$ и каждой точкой $\boldsymbol{z}_i$. Общая схема метода показана на рис. 1. Если расстояние между функциями риска вычисляется на основе квадратичной нормы L_2, то задача оптимизации сводится к квадратичному программированию, что позволяет найти решение (вектор $\boldsymbol{b}$) достаточно просто. Для норм L_1 и L_∞ (SurvLIME-Inf [8]) показано, что задача оптимизации сводится к линейной.

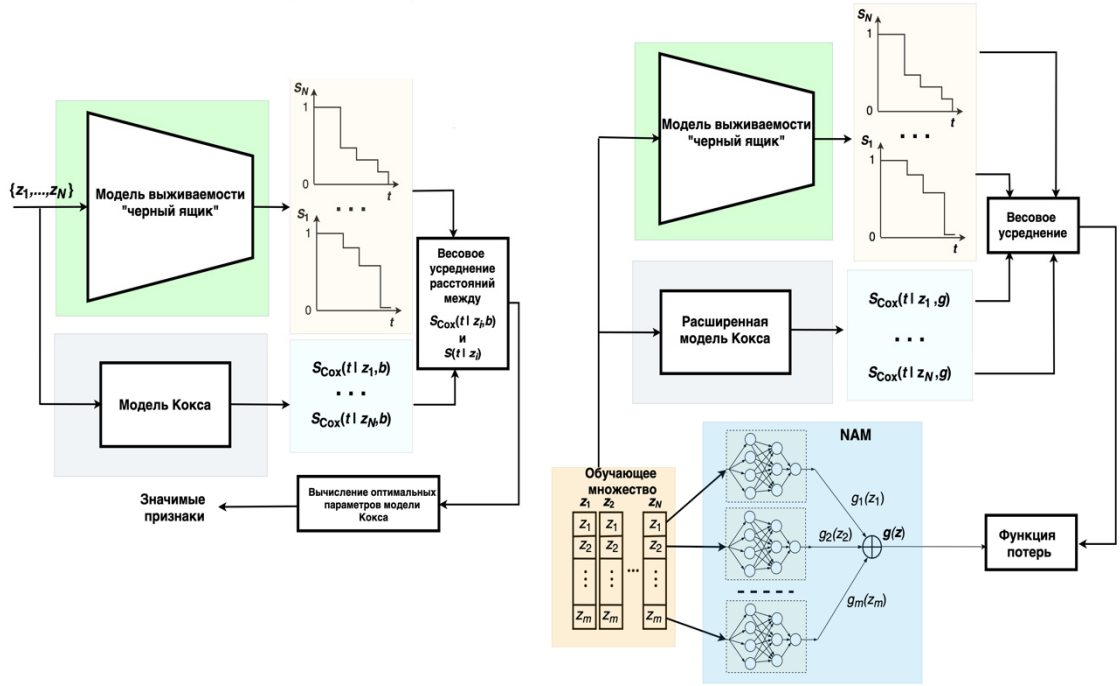

Рис. 1. Структура SurvLIME (слева) и SurvNAM (справа).

Для обеспечения робастности модели объяснения в работе было предложено в работе [9] использовать границы Колмогорова-Смирнова для функции выживаемости. Предлагаемый метод, называемый SurvLIME-KS, является расширением метода SurvLIME и использует результаты этого метода, но в предположении, что вместо одной функции выживаемости используется множество функций. В результате получаем максиминную задачу оптимизации для вычисления $\boldsymbol{b}$, где максимум определяется по всем функциям выживаемости в пределах границ Колмогорова-Смирнова. Несмотря на внешнюю сложность задачи, она сводится к конечному множеству задач квадратичной или линейной оптимизации.

Еще одним интересным методом объяснения является обобщение метода Neural Additive Model (NAM) [10] на случай цензурированных данных, называемое SurvNAM [11]. Идея метода аналогична методу SurvLIME, но в отличие от линейная комбинация $\boldsymbol{b}^T\boldsymbol{x}$ атрибутов, принятой в SurvLIME в соответствии с моделью Кокса, в SurvNAM эта комбинация заменена набором нейронных сетей, которые вычисляют функции атрибутов $g_i(z_i)$. Обучение сетей осуществляется в соответствии с функцией потерь,

определяемой средним расстоянием между функциями выживаемости «черного ящика» и модели Кокса по всем сгенерированным точкам. Скорость изменения каждой функции характеризует влияние соответствующего признака на предсказание.

Модель Кокса даже с учетом функций, реализуемых нейронной сетью, требует вычисления базовой функции риска или выживаемости, которая не зависит от атрибутов. Более мощная модель – оценка Берана. Поэтому был разработан метод, называемый SurvBeX [12] и использующий эту оценку вместо модели Кокса. Основная идея метода заключается в том, что в оценке Берана вес определяется как

$$W(\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{b}) = \text{softmax}(\|\boldsymbol{b} \odot (\boldsymbol{x} - \boldsymbol{x}_i)\|^2 / \tau),$$

где вектор $\boldsymbol{b}$ характеризует влияние признаков на предсказание, $\odot$ – скалярное произведение.

В целом SurvBeX использует алгоритм метода SurvLIME с генерацией точек вблизи объясняемого объекта, но при отличии, что вместо модели Кокса используется модель Берана. В этом случае получаем более сложную задачу оптимизации. Однако многочисленные числовые эксперименты показывают, то она достаточно просто решается существующими методами и обеспечивают существенно лучшие результаты интерпретации.

Список литературы

1. Wang P., Li Y., Reddy C.K. Machine learning for survival analysis: A survey // ACM Computing Surveys. 2019. Vol. 51. P. 1-36.
2. Cox D.R. Regression models and life-tables // Journal of the Royal Statistical Society, Series B (Methodological). 1972. Vol. 34. P. 187-220.
3. Burkart N., Huber M.F. A survey on the explainability of supervised machine learning // Journal of Artificial Intelligence Research. 2021. Vol. 70, P. 245-317.
4. Sahakyan M., Aung Z., Rahwan T. Explainable artificial intelligence for tabular data: A survey // IEEE Access. 2021, Vol. 9. P. 135392-135422.
5. Ribeiro M.T., Singh S., Guestrin C. Why should I trust you? Explaining the predictions of any classifier // arXiv:1602.04938. 2016.
6. Beran R. Nonparametric regression with randomly censored survival data. Technical report. University of California, Berkeley, 1981.
7. Kovalev M.S., Utkin L.V., Kasimov E.M. SurvLIME: A method for explaining machine learning survival models // Knowledge-Based Systems. 2020, Vol. 203. P. 106164.
8. Utkin L.V., Kovalev M.S., Kasimov E.M. An explanation method for black-box machine learning survival models using the Chebyshev distance // Artificial Intelligence and Natural Language. AINL 2020. Cham: Springer, 2020. Communications in Computer and Information Science. Vol. 1292. P. 62-74.
9. Kovalev M.S., Utkin L.V. A robust algorithm for explaining unreliable machine learning survival models using the Kolmogorov-Smirnov bounds // Neural Networks. 2020, Vol. 132. P. 1-18.
10. Agarwal R., Frosst N., Zhang X., Caruana R., Hinton G.E. Neural Additive Models: Interpretable Machine Learning with Neural Nets // arXiv:2004.13912. 2020.
11. Utkin L.V., Satyukov E.D., Konstantinov A.V. SurvNAM: The machine learning survival model explanation // Neural Networks. 2022. Vol. 147. P. 81-102.
12. Utkin L.V., Eremenko D.Y., Konstantinov A.V. SurvBeX: An explanation method of the machine learning survival models based on the Beran estimator // arXiv:2308.03730, 2023.