Регрессионные модели в алгоритме построения деревьев решений CTree

А.А. Белых, А.В. Аталян

Широкое применение алгоритмов, основанных на деревьях принятия решений, обусловлено их умением работать с признаками, измеренными в различных шкалах. Данные не нуждаются в предварительной обработке, такой как нормализация или восстановление пропущенных значений. Наибольшее распространение и популярность получили алгоритмы CART, ID3, C4.5. Но не так давно представленный CTree, основанный на использовании регрессионных моделей, обладает рядом преимуществ при решении ряда практических задач, например, для выявления однородных выборок [1,3].

CTree (Conditional inference trees)

В 2006 году Т. Хотхорном с соавторами был предложен метод дерева условного вывода (CTree) [2]. CTree разделяет процесс разбиения на два отдельных этапа. Первым шагом является определение предикторного атрибута, на основе которого будет строится разбиение, путем определения меры связи между каждым предиктором и интересующим исходом. CTree следует формальным процедурам статистического вывода на каждом этапе разбиения. Связь между каждым предиктором и результатом определяется количественно с использованием коэффициента в регрессионной модели. Линейная регрессия используется для исходов, выраженных непрерывными данными. Логистическая регрессия применяется, когда зависимая переменная (исход) является бинарной или порядковой. Множественная регрессия используется для измерения линейной связи между несколькими независимыми переменными (предикторами) и зависимой переменной (исходом). Цензурированная регрессия применяется с исходом, наблюдаемым с ограничением возможных значений. В цензурированной модели наблюдается не сама зависимая переменная, а её значения в пределах цензурирования. Узел выбирается для разбиения только при наличии достаточных доказательств для отклонения нулевой гипотезы, т. е. гипотезы о том, что ни один из предикторов не имеет взаимосвязи с результатом. Если нулевая гипотеза отклоняется, то в качестве кандидата на разбиение выбирается предиктор, который отражает наиболее сильную связь с интересующим результатом. Если минимальное значение p больше порогового значения значимости, скорректированного с учетом множественности, то разбиение прекращается, и узел объявляется терминальным. Выполняется правило остановки. Затем, после определения атрибута разделения, вычисляется лучшая точка разделения для этой предиктора.

Несмотря на свое название, дерево условного вывода (CTree) основано на одномерных регрессионных моделях, «условный» относится к тому факту, что после первоначального разбиения, последующий вывод происходит в подгруппах.

Алгоритм CTree имеет меньшую точность прогнозирования, по сравнению с другим более известным алгоритмом CART, но в то же время обеспечивает более простой процесс построения дерева, поскольку в CTree общий коэффициент частоты ошибок I типа (α) контролирует размер дерева и устраняет необходимость его обрезания. Значение α может быть установлено независимо от того, в каких

шкалах измерен результат. Используя формальные логические методы, включающие корректировки множественности для выбора разбиений, CTree предоставляет статистические гарантии и действительные значения p при каждом разбиении.

Заключение

При возрастающей популярности метода деревьев решений для решения предметных задач в различных областях знаний, становится актуальным использование различных алгоритмов построения дерева решений, зарекомендовавших себя наилучшим образом. Подтвержденная целесообразность практического применения CTree для решения задачи выделения однородных подгрупп в эпидемиологических медицинских исследованиях дает перспективу практического применения в дальнейшем исследовании популяции Восточной Сибири при решении задачи определения групп риска определенных заболеваний.

Литература

1. Аталян А.В., Белых А.А., Кузьмин О.В. Алгоритмы построения дерева решений CART и CTree: выявление однородных подмножеств в прикладных задачах. Прикладные вопросы дискретного анализа. 2020; 6: 7-12.
2. Hothorn T., Hornik K., Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics. 2006;15(3):651–674. doi: 10.1198/106186006X133933.
3. Venkatasubramaniam A., Wolfson J., Mitchell N., Barnes T., JaKa M., French S. Decision trees in epidemiological research. Emerging themes in epidemiology. 2017;14:11. doi: 10.1186/s12982-017-0064-4.