Сравнение методов дискретизации интервальных переменных для получения эффектов взаимодействия

Жучкова Светлана
11.12.19

До сих пор во всех представляемых исследованиях мы в основном рассматривали эффекты взаимодействия, в которых основной переменной, формирующей взаимодействие, становилась номинальная переменная, а точнее набор дамми-переменных из этой номинальной переменной. Это объяснимо, поскольку эффект взаимодействия, формируемый одной дихотомической и любой другой переменной, интерпретируется гораздо проще, чем, например, эффект взаимодействия, в котором перемножаются две интервальные переменные. Однако если в фокусе исследования находится именно интервальная переменная и, согласно теоретическим предположениям, она должна образовывать взаимодействие, необходимо найти баланс между выполнением этой задачи и простотой интерпретации будущего взаимодействия. Один из возможных способов решения этой задачи – дискретизация интервальной переменной, то есть разбиение её на меньшее количество категорий-интервалов, которые затем в модель вводятся так же, как и номинальные переменные.

Как осуществить такую дискретизацию, то есть как выбрать количество и ширину интервалов для группировки категорий исходной интервальной переменной, чтобы получить максимально эффективные для задач регрессии признаки (будущие эффекты взаимодействия)? Этим докладом мы начинаем серию сравнений подходов к осуществлению такой дискретизации.

Нашей задачей становится поиск наиболее универсального подхода к осуществлению дискретизации – такого, который можно было бы использовать без дополнительных экспертных знаний о предметной области очередного исследования. В связи с этим мы не будем включать в сравнение подход, при котором категории переменной сгруппированы по какому-либо заранее выделенному содержательному основанию. Как было замечено по опыту, применение различных правил дискретизации, распространённых в компьютерных науках (например, использование деревьев решений), автоматически приводит к получению содержательно осмысленных групп категорий.

Для пробного сравнения были выбраны следующие подходы: 1) разбиение переменной на равные интервалы, при этом количество таких интервалов определяется как $\sqrt{k}$, где k – общее количество уникальных категорий (такой подход оказывается наиболее распространённым при построении гистограмм интервальных и непрерывных переменных), и 2) разбиение переменной на интервалы, выделенные в ходе построения дерева решений, в котором участвуют две переменные: текущая интервальная переменная

в роли независимой и будущая зависимая переменная из регрессии в роли зависимой для дерева. В качестве метода построения дерева был выбран алгоритм ChAID – Chi-Square Automatic Interaction Detection, адаптированный для работы с интервальной зависимой переменной (использующий F-статистику в качестве критерия формирования узлов). Был выбран именно ChAID, поскольку это один из самых распространённых методов построения деревьев решений в социальных науках, а также поскольку этот метод позволяет получать небинарные деревья (а значит, и количество получаемых категорий в интервальной переменной с большой вероятностью будет отлично от двух).

Сама идея использования деревьев решения в дискретизации переменной для получения наиболее «сильных» эффектов взаимодействия исходит из следующей логики: чтобы выбрать такие интервалы, которые продемонстрируют наивысшую прогностическую силу в будущей регрессионной модели, необходимо уже на этапе дискретизации переменной максимизировать связь этой переменной с будущим откликом. Именно эту задачу помогают решить деревья, в частности ChAID: категории независимой переменной объединяются в узлы на основании того, как они связаны с зависимой переменной.

В качестве эмпирического примера для проведения сравнения были использованы данные о российских фильмах, собранные с помощью веб-скрапинга с портала Кинопоиск. Содержательная задача состояла в предсказании «популярности» фильма, выраженном в количестве оценок, поставленных этому фильму пользователю, на основе некоторого набора предикторов: жанра фильма, формата фильма, года выхода фильма, продолжительности фильма, количества рецензий пользователей, процента положительных рецензий российских и зарубежных критиков, а также «качества» фильма, выраженного в его рейтинге на портале. В качестве переменной, впоследствии формирующей взаимодействия, был выбран год выхода фильма: в базе эта переменная принимает значения от 1918 до 2019. Исходная регрессионная модель, не содержащая взаимодействий (далее – модель с главными эффектами), имеет коэффициент детерминации 0.399.

Согласно первому подходу, исходная интервальная переменная о годе выхода фильма была разбита на $\sqrt{100}$ = 10 равных интервалов – таким образом полученная переменная отражает *десятилетие* выхода фильма. Согласно второму подходу, были автоматически выделены интервалы переменной на основе дерева решений. Полученные интервалы для обоих подходов представлены в таблице 1.

Таблица 1

Частоты для сгруппированных значений переменной «год выпуска фильма»

Первый подход (bins)		Второй подход (ChAID)	
Интервал	**Частота**	**Интервал**	**Частота**
1918-1926	381	1915-1933	1197
1927-1936	1015	1934-1946	785
1936-1947	631	1947-1960	1213
1947-1957	694	1961-1973	2762
1957-1968	2000	1974-1978	1381
1968-1979	2934	1979	317
1979-1989	3251	1980-1983	1232
1989-2000	2377	1984	314
2000-2011	6148	1985-1989	1705
2011-2022	9007	1990-1991	831
		1992-1994	669
		1995-2012	8100
		2013-2018	7035
		2019-2022	897

Примечательно, что «одиночные» модели, построенные с использованием взаимодействий на основе полученных с помощью каждого подхода переменных не отличались по прогностической силе друг от друга (R2 = 0.58). Также примечательно, что включение только этих взаимодействий уже почти в два раза повысило прогностическую силу модели по сравнению с моделью на главных эффектах.

Затем для проведения сравнения на статистическом уровне была осуществлена процедура бутстрепа: из исходной выборки были сгенерированы 200 подвыборок «с повторением» того же объёма, на каждой из выборки была оценена та же модель, её прогностическая сила (скорректированный R2) и количество значимых предикторов. На основе полученных значений были построены доверительный интервалы для скорректированного R2, позволяющие выявить статистически значимые различия между этими показателями. По итогам эксперимента статистически значимых различий не обнаружилось. На рис. 1 представлены распределения показателей R2 для модели с главными эффектами и моделей с эффектами взаимодействия, интервальные переменные для которых дискретизировались двумя разными подходами. Следует отметить, что для модели, в которой использовалась переменная после ChAID, доверительный интервал

оказался намного шире, чем для первой модели. Это говорит о том, что такой способ дискретизации приводит к более неусточивым результатам, более переобученным моделям.

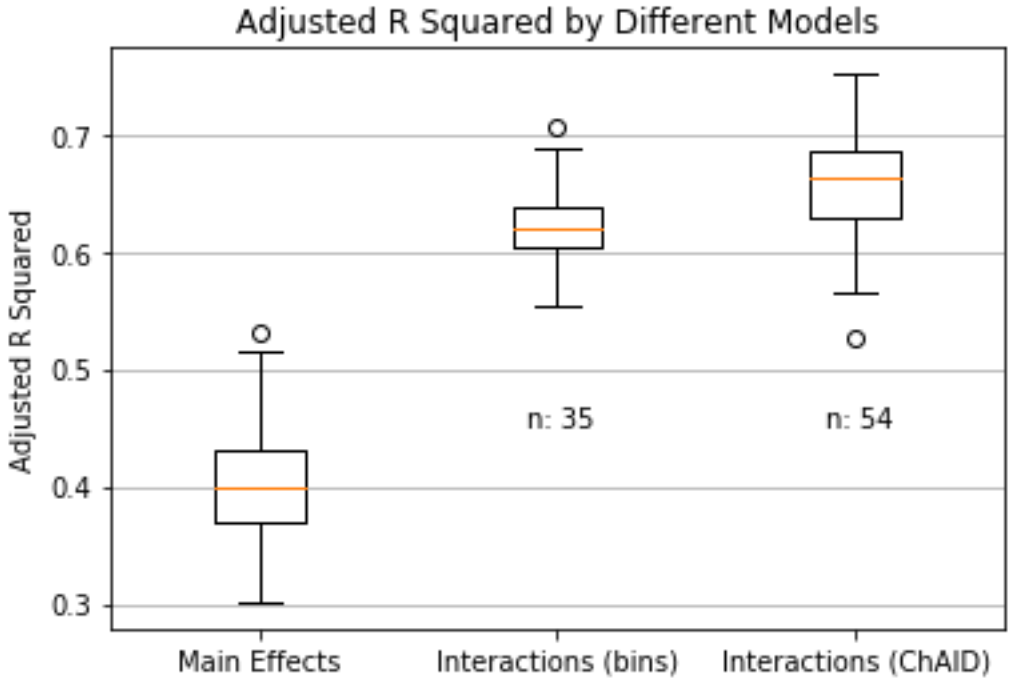

Рис. 1. Распределения показателей R2 для трёх моделей