Метод выявления подгрупп индивидов, однородных с точки зрения эффект от воздействия

Алексей Владимирович Бузмаков, с.н.с., НИУ ВШЭ – Пермь.

В маркетинге, медицине и некоторых других областях необходимо уметь оценивать эффект от некоторого воздействия на индивидов. Так, например, при разработке маркетинговой кампании необходимо уметь оценить насколько такая кампания была эффективной. Для этой цели для каждого воздействия разрабатывается дизайн рандомизированного случайного эксперимента. Классические методы сравнения средних значений некоторой целевой переменной позволяют определить насколько в среднем воздействие было эффективным, а также проверить статистическую значимость найденного эффекта.

Однако, во многих ситуациях нет оснований полагать, что индивиды однородны с точки зрения эффекта от воздействия. Соответственно встаёт вопрос, каким именно образом можно найти индивидов имеющих высокий или низкий эффект от воздействия, что позволит повысить эффективность последующих воздействий. Более того, требуется уметь не только перечислить таких индивидов, но и уметь их описать. В дальнейшем такое описание, может быть использовано как для дальнейшего описания и улучшения эффективности рассматриваемого воздействия (в случае индивидов с высоким эффектом), так и для разработки новых эффективных воздействий (в случае индивидов с низким эффектом).

Ответ на этот вопрос могут дать различные подходы. Наиболее простым и часто применяемым подходом является ручная проверка небольшого количества гипотез, которые принимаются во внимание уже при разработке дизайна эксперимента. Несмотря на простоту статистической проверки таких гипотез, многие важные подгруппы могут быть пропущены.

Следующим возможным подходом являются специальные эконометрические методы анализа взаимодействия между переменной воздействия и независимыми переменными (Athey & Imbens, 2017; Chernozhukov et al, 2018). Такие методы позволяют оценить, насколько меняется эффект от воздействия для среднего индивида, при изменение некоторой характеристики этого индивида. Несмотря, на то, что такие методы позволяет понять предпосылки к изменению эффекта, они не позволяют определить границы группы индивидов, имеющие большой эффект.

Другой подход представлен методами оценки эффекта от воздействия (Devriendt et al 2018). Цель данных методов является наиболее точное предсказание эффекта от воздействия на индивидуальном уровне. Такие методы имеют достаточно неплохое качество предсказания, однако не порождают описания групп индивидов. Подгруппы могут быть получены посредством анализа модели как чёрного ящика (Lundberg & Lee, 2017), или посредством анализа распределения независимых переменных среди индивидов с высоким или низким предсказанием. Однако первый метод не позволяет исследовать взаимодействие независимых переменных, в то время как второй зависит от большого числа ситуативных решений, например, выбор порога отсечения индивидов с высоким предсказанием. Также оба метода не дают конкретных значений независимых переменных, описывающих группы.

Таким образом, существующие методы оценки и анализа неоднородных эффектов от воздействия не позволяют описывать группы индивидов с отличающимся эффектом от воздействия, что необходимо для разработки новых более эффективных воздействий.

В данной работе предлагается подход выявления и описанию подгрупп индивидов имеющих высокий (или низкий) эффект от анализируемого воздействия. Предлагаемый подход является адаптацией идей методов класса Subgroup Discovery (SD) (Lavrac et al, 2004; Boley et al, 2017). Основной идеей этих методов является умный обход пространства всех возможных описаний подгрупп индивидов с последующим выбором наилучшей подгруппы с точки зрения некоторой функции качества. Как правило, такая функция качества обладает следующими свойствами. Во-первых, она должна возрастать, при улучшении подгруппы с точки зрения решаемой задачи, так в нашем случае, она должна возрастать при увеличении (или уменьшении) среднего эффекта от воздействий в группе. С другой стороны функция качества должна убывать, при уменьшении количества индивидов в группе.

Также нужно отметить, что полный обход всего пространства поиска не представляется возможным, в силу его большого размера. Соответственно, методы класса SD, помимо некоторой функции качества подгрупп, требуют иметь также оптимистичную оценку (оценку сверху) для функции качества. Это позволяет отсекать неперспективные ветви поиска. Действительно, если оценка функции качества показывает, что уточненее рассматриваемой подгруппы не может привести к более высокому качеству, чем уже найденная подгруппы, то нет необходимости эту подгруппу уточнять.

В данной работе рассматривается функция качества Q(|S|,delta(S)), где |S| – это количество элементов в подгруппе S, a delta(S) – – это размер эффекта для подгруппы S. В частности предлагается использовать в качестве delta расстояние между доверительными интервалами целевой переменной в тестовой и контрольной групп рандомизированного эксперимента.

В предлагаемой работе показывается каким образом можно посчитать оптимистичную оценку для функции Q за квадратичное время от размера подгруппы. Однако, для возможности использования предлагаемого подхода на реальных данных необходимо уменьшить вычислительное время расчёта оптимистичной оценки. Соответственно, для линейной функции Q показывается возможность расчёта её оптимистичной оценки за линейное время.

Результаты данной работы были применены для анализа рандомизированного эксперимента в области маркетинга и позволили описать подгруппы индивидов с экстремальным размером эффекта от воздействия, что подчеркивает важность предлагаемой работы.

Список литературы:

1. Athey S., Imbens G.W. Chapter 3 – The Econometrics of Randomized Experiments // Handbook of Economic Field Experiments / ed. Banerjee A. V., Duflo E. 2017. Vol. 1. P. 73–140.
2. Chernozhukov V. et al. Generic machine learning inference on heterogenous treatment effects in randomized experiments. – National Bureau of Economic Research, 2018. – №. w24678.
3. Devriendt F., Moldovan D., Verbeke W. A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics // Big data. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, 2018. Vol. 6, № 1. P. 13–41.
4. Lundberg S.M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems 30 / ed. Guyon I. et al. Curran Associates, Inc., 2017. P. 4765–4774.
5. Lavrač N. et al. Subgroup Discovery with CN2-SD // J. Mach. Learn. Res. 2004. Vol. 5. P. 153–188.
6. Boley M. et al. Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery // Data Min. Knowl. Discov. Springer, 2017. Vol. 31, № 5. P. 1391–1418.