МЕТОДЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ В ПСИХОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Мелешкевич Д.В., Александрович А.Ф., Ситник М.Ю.
Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь
Осипович Т.А. – кандидат экононических наук, доц.

Методы больших данных, часто называемые машинным обучением, статистическим обучением и извлечением данных, представляют собой совокупность статистических методов, способных находить сложные сигналы в больших объемах данных. Пользуясь доступностью данных из различных источников, таких как приложения для мобильных телефонов, биосенсоры и социальные сети, исследователи стремятся извлечь структуру и смысл из огромных объемов данных, чтобы выявить закономерности и сделать прогнозы. Учитывая, что большая часть этих данных является поведенческой, психологи должны играть главную роль в анализе этих данных.

Методы интеллектуального анализа данных можно условно разделить на два основных класса: методы обучения под наблюдением и методы обучения без присмотра. В контролируемом обучении есть интересный результат - цель состоит в том, чтобы разработать модель прогнозирования на основе набора переменных. Большинство контролируемых методов обучения ориентированы на выбор переменных, нелинейность и интерактивные эффекты и, таким образом, предлагают много преимуществ по сравнению со стандартными регрессионными моделями. Модели регрессии с большим количеством переменных могут быть нестабильными, особенно если существует высокая степень корреляции между переменными предиктора. Кроме того, когда количество переменных велико, может быть почти невозможно вручную найти, какие взаимодействия могут присутствовать. Целью контролируемых методов обучения является выявление важных переменных, нелинейные формы переменных или их интерактивные эффекты. Эти подходы часто дают модель, которая является более простой и более понятной, поскольку важные эффекты могут быть изолированы. Кроме того, полученная модель с большей вероятностью будет воспроизводиться в новом образце.

В обучении без наблюдения нет никакой переменной результата, если поставить цель: сгруппировать переменные или участников по степени их сходства или ковариации, понимаем, что в отличие от контролируемых методов обучения, неконтролируемое обучение обычно используется в психологических исследованиях. Например, методы сокращения данных, такие как анализ основных компонентов и анализ поисковых факторов, довольно распространены в психологии, как и методы группировки участников [1].

Методы обучения под наблюдением редко используются в психологии. Однако эти методы должны и будут играть большую роль в психологических исследованиях в будущем. Как уже отмечалось, одна из причин, по которой эти методы могут не закрепиться в психологии, заключается в том, что исследователи могут подумать, что методы требуют огромных объемов данных - множество участников и множество переменных. Стоит отметить, что многие методы интеллектуального анализа данных хорошо работают при небольших настройках данных [2].

Хотя алгоритмы интеллектуального анализа данных могут применяться с небольшими выборками, исследователи должны быть осторожны с их использованием. Чем меньше наборы данных, тем выше склонность к объяснению шума или уникальных особенностей данных. Чтобы преодолеть эту проблему, абсолютно необходимо использовать различные формы перекрестной проверки в сочетании с этими методами. Хотя это не новая концепция в психологии, перекрестная проверка редко используется в психологических исследованиях. Перекрестная проверка обычно влечет за собой разделение набора данных на две части: обучающий набор данных и тестовый набор данных. С набором обучающих данных можно исследовать практически все, но обычно используется форма внутренней перекрестной проверки, чтобы предотвратить переобучение в наборе обучающих данных. После того, как исследуем, небольшое количество моделей (от 1 до 3) выбираются так, как это целесообразно - исследуем прогнозирующую природу этих моделей в тестовом наборе данных. Стоит понимать, что это не означает, что мы переоцениваем модель на тестовом наборе данных. Вместо этого возьмем модель, созданную на основе обучающего набора данных, и создадим прогнозы на основе тестовых данных. Это дает нам более реалистичную оценку того, насколько хорошо будет работать модель, если будут собраны данные из новой выборки [3].

Как уже отмечалось, неконтролируемые методы обучения довольно распространены в психологии. Анализы основных компонентов и поисковых факторов являются общими методами сокращения данных, поэтому поиск факторов часто является первым шагом в понимании размерности данных. Во многих случаях эта модель применяется к половине набора данных, а затем модель оставляющего фактора оценивается на оставшейся половине данных как способ отделить исследовательские и подтверждающие аспекты анализа данных. Этот подход похож на перекрестную проверку, но в психологии исследователи часто не проверяют точную модель. Как правило, модель переоценивается, а коэффициенты нагрузки, которые были незначительными, фиксируются на 0.

Одна из проблем, с которой в настоящее время используется моделирование в психологии, заключается в том, что перекрестная проверка редко используется для оценки жизнеспособности модели. Однако в последнее время перекрестной проверке уделялось больше внимания при моделировании [4,5].

Хотя контролируемые методы обучения не часто используются в психологии, большая часть этого может объясняться отсутствием внимания, которое эти методы получают от методологов в психологических науках. Медленно, но верно это меняется, поскольку все больше и больше методов интеллектуального анализа данных адаптируются к нюансам и сложностям психологических данных и методов [6]. В частности, необходимо сосредоточиться на том, чтобы объединить многие из этих методов больших данных с моделями скрытых переменных, которые распространены в психологии.

Латентные переменные модели (например, модели подтверждающих факторов, модели структурных уравнений широко распространены в психологии, учитывая многомерные измерения и довольно распространенные продольные конструкции. Комбинация алгоритмов интеллектуального анализа данных с моделями скрытых переменных является необходимым шагом для расширения использования среди психологов, и есть несколько недавних примеров этой интеграции. Например, объединили SEM с алгоритмами дерева классификации и регрессии для разработки деревьев SEM. В деревьях SEM ряд переменных предикторов используется для разделения данных, а пользовательский SEM подходит для каждого раздела данных. Цель состоит в том, чтобы найти предикторы с точками среза, которые максимизируют соответствие модели. По существу, это автоматический способ поиска групп участников, в которых члены одной группы однородны по отношению к SEM, а члены разных групп неоднородны по отношению к SEM. Например, деревья SEM могут использоваться для поиска групп с разными траекториями во времени или групп, в которых присутствуют разные модели измерения.

Аналогичным образом в 2016 году объединили регуляризацию, метод, распространенный в многомерной регрессии, с SEM для создания регуляризованной SEM, что позволяет исследователям штрафовать конкретные параметры в SEM. Это приводит к более простым и более воспроизводимым SEM. Также были аналогичные разработки в рамках многоуровневого моделирования, где объединили модели смешанных эффектов и деревья регрессии для создания деревьев регрессии смешанных эффектов. Эти подходы могут эффективно выполнять поиск многомерных иерархически структурированных данных для нелинейных и интерактивных эффектов [7].

Выделяем проблему, которой уделяется меньше внимания - неполные данные. Проще говоря, многие алгоритмы интеллектуального анализа данных требуют полных данных. Кроме того, разные программы по-разному обрабатывают неполные данные. Учитывая, что неполные данные являются общими в психологических исследованиях и часто не пропускаются полностью случайным образом, модели могут давать смещенные результаты или, по меньшей мере, результаты будут зависеть от метода, используемого для обработки неполных данных. Таким образом, одним из направлений будущих исследований, которое значительно повысит полезность многих из этих методов в психологических исследованиях, является включение современных методов недостающих данных, таких как множественное вменение или полная оценка информации, в программы интеллектуального анализа данных.

Исследователи часто стремятся проверить гипотезы, основанные на теории, с помощью своих статистических моделей, но в то же время исследователи готовы учиться на своих данных путем исследования. Озабоченность этим исследованием заключается в том, что исследователи проводят свои исследования уникальными способами, без необходимых мер предосторожности для предотвращения случайных результатов, и стремятся адаптировать модели к имеющимся данным. Методы извлечения данных, по большей части, являются строго исследовательскими процедурами, способными эффективно искать в данных ассоциации и нелинейные эффекты, и имеют меры предосторожности для предотвращения переобучения. По этим причинам кажется объективным призвать психологических исследователей рассмотреть и оценить использование алгоритмов интеллектуального анализа данных в своих исследованиях.

Список использованных источников:

1. Breiman, L., Friedman, J., Stone, C.J., & Olshen, R.A. (1984). Classification and regression trees. Boca Raton, Florida: CRC press.

2. Hayes, T., Usami, S., Jacobucci, R., & McArdle, J.J. (2015). Using Classification and Regression Trees (CART) and random forests to analyze attrition: Results from two simulations. Psychology and aging, 30, 911-929.

3. Browne, M.W. (2000). Cross-validation methods. Journal of Mathematical Psychology, 44, 108-132.

4. Grimm, K.J., Mazza, G., & Davoudzadeh, P. Model selection in finite mixture models: A k-fold cross-validation approach. Structural Equation Modeling: A Multidisciplinary Journal.

5. Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), Oxford library of psychology. The Oxford handbook of quantitative methods: Statistical analysis (p. 551–611). Oxford University Press.

6. McNeish, D.M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. Multivariate Behavioral Research, 50, 471-484.