УДК 336

МАГОМЕДТАГИРОВ МУРАД МУСАЕВИЧ
к.э.н., доцент кафедры социальных и информационных технологий
ФГБОУ ВО «Дагестанский государственный университет»,
e-mail: m-tag@yandex.ru

АЙГУБОВ САЙДАРХАН ЗАНКУЕВИЧ
к.ф-м.н., доцент, заведующий кафедрой социальных и информационных технологий
ФГБОУ ВО «Дагестанский государственный университет»,
e-mail: aygubov.s62@mail.ru

ОМАРОВА САБИНА РАМАЗАНОВНА
магистрант кафедры социальных и информационных технологий
ФГБОУ ВО «Дагестанский государственный университет»,
e-mail: m-tag@yandex.ru

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ В СОЦИОЛОГИИ

Аннотация. Цель работы. *В статье рассматриваются тенденции применения машинного обучения в социологии и смежных науках.* ***Метод или методология проведения работы.*** *Проведено исследование фактов применения методов контролируемого и неконтролируемого машинного обучения в прогнозировании и моделировании.* ***Результаты.*** *Машинное обучение – это область знаний на стыке статистики и информатики, которая использует алгоритмы для извлечения информации и знаний из данных. Его все чаще применяют в экономике, политологии и социологии. Мы предлагаем краткий экскурс в этот обширный инструментарий, где будет проиллюстрировано его нынешнее использование в социальных науках, включая извлечение информации из новых источников данных, таких, как текст и изображения, характеризующие неоднородность населения; улучшение причинно-следственных выводов и выработка прогнозов для оказания помощи в развитии теории. В дополнение к обеспечению подобного использования в социологии мы утверждаем, что инструменты машинного обучения могут применяться в моделировании.* ***Область применения результатов.*** *Результаты проведенного исследования могут быть использованы при анализе и прогнозировании изменений в социальной сфере.* ***Выводы.*** *Делается вывод, что дальнейшее развитие социологии и смежных наук тесно связано с применением методов контролируемого и неконтролируемого машинного обучения, а инструменты машинного обучения дополняют, а не заменяют существующие методы в социологии*

Ключевые слова: *контролируемое обучение, неконтролируемое обучение, причинно-следственный вывод, прогнозирование, неоднородность, обнаружение.*

MAGOMEDTAGIROV MURAD MUSAEVICH
Ph. D., associate Professor of the Department of social and information technologies
Dagestan state University,
e-mail: m-tag@yandex.ru

AIGUBOV SAYDARKHAN ZANKUEVICH
Ph. D., associate Professor, head of the Department of social and information technologies
Dagestan state University,
e-mail: aygubov.s62@mail.ru

OMAROVA SABINA RAMAZANOVNA
master's student of the Department of social and information technologies
Of the "Dagestan state University",
e-mail: m-tag@yandex.ru

APPLICATION OF MACHINE LEARNING IN SOCIOLOGY

Abstract. Purpose of work. *The article discusses the trends in the use of machine learning in sociology and related Sciences.* ***Method or methodology of the work.*** *A study of the facts of using methods of controlled and uncontrolled machine learning in forecasting and modeling is conducted.* ***Results.*** *Machine learning is a field of knowledge at the intersection of statistics and computer science that uses algorithms to extract information and knowledge from data. It is increasingly used in Economics, political science, and sociology. We offer a brief introduction to this extensive Toolkit, which will illustrate its current use in the social Sciences, including extracting information from new data sources, such as text and images that characterize population heterogeneity; improving causal inferences; and making predictions to help develop the theory. In addition to providing such use in sociology, we claim that machine learning tools can be applied in modeling.* ***Scope of the results.*** *The results of the research can be used in the analysis and forecasting of changes in the social sphere.* ***Conclusions.*** *It is concluded that the further development of sociology and related Sciences is closely related to the use of methods of controlled and uncontrolled machine learning, and machine learning tools complement, rather than replace, existing methods in sociology*
Keywords: *controlled learning, uncontrolled learning, causal inference, prediction, heterogeneity, detection.*

Введение. Машинное обучение (МО) позволяет автоматизировать исследование данных, что является прорывом в области компьютерных наук, где интеллектуальные системы обычно задействовали фиксированные алгоритмы (логические наборы инструкций), которые кодируют желаемый результат для всех возможных ситуаций. В настоящее время интеллектуальные системы «учатся» на основе данных и оценивают сложные функции, которые обнаруживают представления некоторых входных данных (X) или связывают входные данные с выводами (Y), чтобы делать прогнозы по новым данным. МО можно рассматривать как ответвление непараметрической статистики [13].

Результаты исследования. Мы можем классифицировать инструменты МО по тому, как они «обучаются» (извлекают информацию из данных). Отдельные группы МО используют различные алгоритмы, которые характеризуются предположениями о принципах, лежащих в основе интеллекта (Domingos, 2015). Также возможно классифицировать инструменты МО по типу опыта, который приобретается ими в процессе обучения [11]. Именно этой классификацией мы воспользуемся для настоящего исследования.

В управляемом машинном обучении (УМО) алгоритм наблюдает выводы (Y) для всех входящих данных (X). Этот вывод дает алгоритму цель для прогнозирования. В неконтролируемом машинном обучении (НМО) алгоритм наблюдает только входные данные (X). Он должен расшифровать данные без информации о правильных выводах.

Управляемое машинное обучение (УМО) включает в себя поиск функции f (X), изучая выводы (Y) с учетом входных данных (X). Можно рассматривать различные классы функций, такие, как линейные модели, деревья решений или нейронные сети.

Существует две разновидности статистического анализа: моделирование данных и алгоритмическое моделирование. Классическая статистика следует за моделированием данных. Главной целью является понимание того, как результат (Y) связан с входными данными (X). Аналитик предлагает модель, которая могла бы генерировать данные и оценивает параметры модели из этих данных. Такое моделирование приводит к простым и интерпретируемым моделям, но часто игнорирует неопределенность модели. Машинное обучение использует прогнозное моделирование. Главной его целью является прогнозирование результата (Y) для будущих входных данных (X). Аналитик рассматривает базовую модель для данных как неизвестную и рассматривает прогнозную точность альтернативных моделей для новых данных. Прогнозное моделирование отдает предпочтение сложным моделям, которые хорошо работают вне выборки, но могут давать мало информации о механизме, связывающем входные данные с выходными.

Методы УМО стремятся достичь идеального баланса между уменьшением ошибки в вы-

борке и ошибки вне выборки. Эта цель помогает избежать двух подводных камней анализа данных: недостаточной подгонки и чрезмерной подгонки.

Одним из способов использования данных УМО является выбор модели по результатам оценки производительности альтернативных моделей (функций, параметров). Этот процесс требует решения оптимизационной задачи.

Важным шагом в УМО является отделение данных, используемых для выбора модели, от данных, используемых для оценки модели. Для этого создается три, а не два отдельных набора данных. Обучающие данные используются для подгонки модели; проверочные данные откладываются для выбора между различными моделями (или для выбора между различными параметрами одной и той же модели), и, наконец, тестовые данные хранятся для вычисления ошибки обобщения выбранной модели. Не существует общего правила для определения идеального разделения данных, но обычно исследователь может зарезервировать половину данных для обучения, а четверть – для проверки и тестирования [12].

УМО использует гибкие функции входных сигналов (X), для того чтобы предсказать выход (Y). Некоторые инструменты УМО вообще не имеют параметров. Другие методы дают оценки параметров, но эти оценки не всегда согласуются (т. е. не сходятся к истинному значению по мере роста количества операций) [14].

Социологи привыкли работать со статистическими моделями, которые дают оценки параметров с определенными свойствами (объективными и непротиворечивыми). Но УМО не предназначен для воспроизводства статистических моделей. Вместо этого УМО хорошо решает более сложные задачи. Социологи выделяют три их класса: прогнозирование развития, поиск причинно-следственных связей и увеличение данных.

УМО является полезным инструментом для прогнозирования политики, если исследователь заинтересован не в понимании взаимосвязи между X и Y, а, скорее, в использовании X для прогнозирования Y в новых случаях. Политические прогнозы навязывают четкую цель и позволяют создать «общую структуру задач», где разные команды могут соревноваться по одному и тому же вопросу [9].

Экономист Эд Глаэсар и его коллеги использовали эту идею для организации конкурса на разработку прогнозных алгоритмов для городских властей. Социолог Мэтт Салганик и его коллеги начали работу по прогнозированию образовательных (и других) результатов в данных о хрупких семьях. Организационная группа оценивала представленные 150 междисциплинарными группами материалы по точности прогнозирования на основе тестовых (удерживающих) данных. На продолжающемся втором этапе команда планирует провести углубленное изучение несовпадающих случаев в модели выигрыша (например, студенты, которые «побили шансы»), и, таким образом, рассматривает прогнозы как первый шаг к созданию новых идей и теории, а не как конечную цель [10].

Ученые применяют УМО к различным вопросам в экономике, политологии и криминологии. Д. Клейнберг использовал модель, чтобы предсказать, какие пациенты получат наибольшую пользу от операции по замене суставов. Отдельные исследователи обсуждают, как УМО может предсказать и предотвратить смертельный конфликт, используют нейронные сети для прогнозирования милитаризованных международных споров, в частности, палестино-израильских конфликтов, насильственных эпизодов в Африке. Р. Берк использует УМО для прогнозирования криминального риска. Этот ученый использует его предсказания в качестве отправной точки для расследования рассматриваемого процесса и для развития существующей теории.

Д. Клейнберг, например, иллюстрируют, как машинные прогнозы могут помочь нам понять процесс, лежащий в основе судебных решений. Авторы сначала обучают регрессионную модель для прогнозирования решений судей о залоге или освобождении в Нью-Йорке, а затем используют назначение судей по делам для объяснения источников несоответствия между прогнозами модели и фактическими решениями. Их выводы показывают, что судьи оценивают текущее обвинение, оправдывая обвиняемых при сильной доказательной базе, если их нынешнее обвинение является незначительным, и осуждая обвиняемых при относительно слабой доказательной базе, если нынешнее обвинение серьезно [13]. Эти результаты раскрывают важные идеи о человеческом принятии решений и несут в себе потенциал для разработки новой

теории. Важная дискуссия в литературе касается того, как УМО-инструменты должны взвешивать различные виды «ошибок» прогнозирования. Р. Берк, например, применил модель для прогнозирования повторных правонарушений в случаях бытового насилия. В консультации с заинтересованными сторонами авторы взвешивают ложные отрицательные значения (где модель предсказывает отсутствие повторного преступления, когда оно есть) в 10 раз сильнее, чем ложные положительные значения (где модель предсказывает повторное преступление, когда его нет). Их модель, следовательно, дает очень точные предсказания случаев отсутствия правонарушений (которые требуют очень сильных доказательств), но менее точные прогнозы повторных преступлений [4].

Есть законные опасения, что прогнозы УМО (и данные, на которых они основаны) могут усилить социальное неравенство. Что делать, если «прогнозируемые» правонарушители непропорционально сильно набираются из отдельных социальных групп? В настоящее время ученые признают неизбежный компромисс между предсказательной точностью и справедливостью алгоритма [4, 13]. Открытым остается вопрос, как определить понятие «справедливость». Хотя большинство определений касаются обращения с «защищенными группами», можно определить справедливость многими различными способами.

Различные определения справедливости приводят к различным результатам, очень трудно реализовать несколько определений одновременно. Обращение к справедливости алгоритма – это не просто технический вопрос в ОД; он требует от нас – как от общества – рассмотрения трудных компромиссов.

Социологи часто заинтересованы в выявлении причинно-следственных связей между входными данными (X) и выводами (Y). Инструменты УМО могут помочь в определенных процедурах поиска причин, которые включают задачи прогнозирования.

Неконтролируемое машинное обучение (НМО) ищет представление входных данных (X), которое является более полезным, чем выводы. Некоторые инструменты НМО уменьшают размерность данных (например, факторный анализ, тематическое моделирование). Другие методы разделяют данные на группы (например, кластерный анализ, анализ латентных классов, анализ последовательности)

В связи с отсутствием вывода (Y), показывающего алгоритм, к чему должна стремиться модель, нет непосредственной меры успеха. Исследователи используют эвристические инструменты для оценки полученных результатов.

Социологи могут использовать НМО для измерения и обнаружения. Выходные данные из НМО обычно становятся входными данными, которые позволяют последующий анализ или теоретизирование. В отсутствие «основной истины» исследователи должны уделять особое внимание проверке моделей и подтверждать свои результаты с помощью статистических, предметных или внешних критериев.

НМО может создавать единицы измерения из данных, которые будут использоваться в последующем статистическом анализе. Социологи уже давно используют основные компоненты и факторный анализ, чтобы свести многие исходные данные к меньшему набору. Социологи теперь используют НМО для обработки новых видов данных (изображений или текста). Экономисты, например, классифицируют спутниковые изображения с помощью НМО для выработки мер (обезлесение, загрязнение, ночное освещение и т. д.), которые относятся к экономическим результатам [8].

Следуя сложившейся традиции, социологи также используют НМО для группировки данных социальных сетей.

Некоторые методы НМО

Анализ главных компонент – обнаруживает небольшое число линейных комбинаций входных данных (X), которые не коррелируют друг с другом и захватывают большую часть изменчивости в данных. Эти линейные комбинации («главные компоненты») могут использоваться в качестве входных данных в последующем анализе (например, в регрессии для прогнозирования некоторого выходного сигнала, Y).

Факторный анализ – обнаруживает скрытые (ненаблюдаемые) факторы, которые учитывают корреляцию во входных данных (X); возвращает «факторные нагрузки» для каждого входного сигнала, который может быть использован для интерпретации факторов.

Кластерный анализ – группирует наблюдения в заданное число «кластеров» таким образом, что наблюдения в кластере более похожи друг на друга, чем на наблюдения в других кластерах; возвращает принадлежность кластера для каждого наблюдения.

Латентный анализ классов – обнаруживает скрытые классы наблюдений, которые могут учитывать корреляции в наблюдаемых данных; возвращает вероятность принадлежности к классу для каждого наблюдения.

Анализ последовательности – сравнивает последовательности (упорядоченные элементы или события) с «оптимальным соответствием», чтобы обнаружить группы наблюдений с аналогичными особенностями (обычно с помощью кластерного анализа).

Обнаружение сообщества – определяет «сообщества» в сетях (графах) на основе структурного положения узлов.

НМО может помочь охарактеризовать неоднородность популяции. Например, С. Бейл применяет «нечеткий» кластерный анализ (который позволяет случаям принадлежать к нескольким группам), чтобы обнаружить три конфигурации символических границ между иммигрантами и коренными жителями в Европе. Б. Бониковский используют латентный классовый анализ для характеристики четырех типов популярного национализма в Соединенных Штатах.

Голдберг разрабатывает «реляционный анализ классов», который рассматривает ассоциации между ответами на опрос отдельных лиц (а не сами ответы), чтобы обнаружить три отдельные логики культурного различия вокруг музыкальных вкусов. Также он применяет один и тот же инструмент для выявления трех конфигураций политических убеждений среди американцев.

Эти примеры используют различные методы, но имеют общую цель. Они ищут скрытую структуру в популяции, которая была бы предположительно однородной в соответствии с традиционным статистическим подходом. Этот подход часто дает новые гипотезы, которые появляются из данных.

В отличие от задач прогнозирования, в НМО часто отсутствует «основная истина», поэтому проверка модели является важным шагом. Исследователи используют методы статистической проверки, которые включают некоторые эвристические меры, чтобы захватить, например, «кластеры».

Исследователи также прибегают к внешней проверке, которая приносит новые данные для оценки того, подтверждают ли выявленные закономерности ожидания. Например, С. Бейл показывает, что три типа символических границ, возникающих из данных об отношении, связаны с миграционными моделями на страновом уровне и интеграционными философиями в Европе. Б. Бониковский обнаружил, что четыре разновидности национализма в США коррелируют с социальными и политическими установками, которые не были использованы при идентификации типологии.

Существует две широкие категории машинного обучения (МО) – это ответвление информатики и статистики. Управляемое машинное обучение (УМО) строит модель входных данных (X) для прогнозирования выходных данных (Y) в новых данных. Неконтролируемое машинное обучение (НМО) обнаруживает шаблоны во входных данных (X) без целевого объекта (Y) для прогнозирования. Хотя многие из инструментов МО являются довольно новыми для социологии, проблемы, которые они решают, таковыми не являются.

В количественной социологии мы часто следуем классическому статистическому подходу: предполагаем распределение данных, выбираем несколько входных данных и задаем параметрическую (обычно линейную) модель, чтобы связать входные данные с выходными. Мы склонны отдавать предпочтение моделям, которые согласуются со здравым смыслом, рассматриваем некоторые альтернативные спецификации (например, вложенные модели, которые постепенно вводят элементы управления), но не исчерпываем все возможности и полностью учитываем неопределенность модели.

УМО позволяет нам включать множество входных данных (включая термины и взаимодействия более высокого порядка) и сложные функции, которые соединяют входные данные (X) с выводами (Y).

Социологи могут идентифицировать чистое предсказание проблемы, в которых различные исследовательские группы потенциально могут конкурировать в рамках «общей задачи» [9].

Экономисты, например, уже используют УМО для составления политических прогнозов [13].

Социологи также могут использовать предсказания в качестве отправной точки для понимания глубинных социальных процессов и развития теории.

Еще одним направлением для социологов является использование УМО для совершенствования классических статистических методов. Экономисты теперь применяют УМО к задачам прогнозирования в рамках причинно-следственных связей,

Количественная социология часто использует дедуктивный подход, когда исследователь выводит гипотезы из теории для проверки на данных. Чтобы вписать нашу работу в форму проверки гипотез, мы разделяем социальные теории на несколько переменных и оцениваем средний эффект каждой переменной в некоторой заданной популяции. Мы также противопоставляем друг другу несколько теорий, чтобы эмпирически определить, какая из них лучше всего подходит. При этом игнорируется возможность того, что различные механизмы могут воздействовать одновременно, а также проблемы неопределенности.

МО предлагает новые инструменты для характеристики неоднородности населения. Социологи используют НМО для выявления подгрупп в популяциях, а затем связывают возникновение каждой подгруппы с различными внешними факторами [3].

Расширяя свой инструментарий для включения МО, социологи могут лучше рассмотреть неоднородность.

Результаты. В количественной социологии мы в основном занимаемся исследовательской работой, но только на языке «проверки гипотез». Мы часто используем гибкие конструкции исследования и статистические модели, пока не узнаем что-то новое и интересное. МО дает нам широкий спектр инструментов для изучения и изучения данных, но, для того чтобы эти инструменты были полезны в социологии, сначала нужно отличить исследовательскую работу от подтверждающих исследований.

Выводы. МО предоставляет широкий набор инструментов, которые могут информировать о различных вопросах. В социологии мы в значительной степени опираемся на систему «проверки гипотез» и классический статистический подход. Большинство ученых обычно приспосабливает свои вопросы к этой настройке и использует данные для оценки влияния некоторых входных данных (X) на выводы (Y). МО не только помогает улучшить части этой стратегии, но и дает инструменты, которые могут вдохновить на новые вопросы. Насколько хорошо набор входных данных (X), например, предсказывает выходные данные (Y)? Как эти прогнозы отклоняются от наблюдаемых результатов и почему? Какова базовая структура некоторых входных данных (X)? Как эта структура связана с внешними факторами (Z)? Ответы на эти вопросы могут помочь совершенствовать теорию или генерировать новые гипотезы. МО обеспечивает не конечную цель, а отправную точку для дальнейшего анализа. Таким образом, инструменты МО дополняют, а не заменяют существующие методы в социологии.

Литература

1. Айвазян, С. А., Енюков, И. С., Мешалкин, Л. Д. Прикладная статистика : основы моделирования и первичная обработка данных. – М. : Финансы и статистика, 1983. С. 2.
2. Флах, П. Машинное обучение. – М. : ДМК Пресс, 2015.
3.Bail, CA. The cultural environment : measuring culture with big data // Theor. Soc. 2014 . No. 43. P. 465–482.
4. Berk, R. Criminal Justice Forecasts of Risk. – New York : Springer, 2012.
5. Bonikowski, B. DiMaggio, P. Varieties of American Popular Nationalism // Am. Soc. Rev. 2016. No. 81. P. 949–980.
6. Cederman, L. E., Weidmann, N. B. Predicting armed conflict : Time to adjust our expectations? // Science. 2017. No. 355. P. 474–476.
7. DiMaggio, P., Nag, M., Blei, D. Exploiting affinities between topic modeling and the sociological perspective on culture : Application to newspaper coverage of U.S. government arts funding // Poetics. 2013. No. 41. P. 570–606.
8. Domingos, P. The master algorithm : How the quest for the ultimate learning machine will remake our world. – New York: Basic Books, 2015.
9. Donoho, D. 50 Years of Data Science // J. Comput. Graph. Stat. 2017. No. 26. P. 745–766.
10. Glaeser, E. L., Hillis, A., Kominers, S. D., Luca, M. Crowdsourcing City Government : Using Tournaments to Improve Inspection Accuracy // Am. Econ. Rev. 2016. No. 106. P. 114–118.
11. Goodfellow, I., Bengio, Y., Courville, A. Deep learning. – Cambridge: MIT Press, 2016.
12. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. – 2nd ed. – New York : Springer, 2009. Kleinberg, J., Liang, A., Mullainathan, S. The Theory is

Predictive, but is it Complete? // An Application to Human Perception of Randomness. – URL : https:// arxiv.org/abs/1706. 06974.
13. Knight, K., Fu, W. Asymptotics for lasso-type estimators // Ann. Stat. 2000. No. 28. P. 1356–1378.

References:

1. Ajvazyan, S. A., Enyukov, I. S., Meshalkin, L. D. Prikladnaya statistika : osnovy modelirovaniya i pervichnaya obrabotka dannyh. – M. : Finansy i statistika, 1983. S. 2.
2. Flah, P. Mashinnoe obuchenie. – M. : DMK Press, 2015.
3.Bail, CA. The cultural environment : measuring culture with big data // Theor. Soc. 2014 . No. 43. P. 465–482.
4. Berk, R. Criminal Justice Forecasts of Risk. – New York : Springer, 2012.
5. Bonikowski, B. DiMaggio, P. Varieties of American Popular Nationalism // Am. Soc. Rev. 2016. No. 81. P. 949–980.
6. Cederman, L. E., Weidmann, N. B. Predicting armed conflict : Time to adjust our expectations? // Science. 2017. No. 355. P. 474–476.
7. DiMaggio, P., Nag, M., Blei, D. Exploiting affinities between topic modeling and the sociological perspective on culture : Application to newspaper coverage of U.S. government arts funding // Poetics. 2013. No. 41. P. 570–606.
8. Domingos, P. The master algorithm : How the quest for the ultimate learning machine will remake our world. – New York: Basic Books, 2015.
9. Donoho, D. 50 Years of Data Science // J. Comput. Graph. Stat. 2017. No. 26. P. 745–766.
10. Glaeser, E. L., Hillis, A., Kominers, S. D., Luca, M. Crowdsourcing City Government : Using Tournaments to Improve Inspection Accuracy // Am. Econ. Rev. 2016. No. 106. P. 114–118.
11. Goodfellow, I., Bengio, Y., Courville, A. Deep learning. – Cambridge: MIT Press, 2016.
12. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. – 2nd ed. – New York : Springer, 2009. Kleinberg, J., Liang, A., Mullainathan, S. The Theory is Predictive, but is it Complete? // An Application to Human Perception of Randomness. – URL : https:// arxiv.org/abs/1706. 06974.
13. Knight, K., Fu, W. Asymptotics for lasso-type estimators // Ann. Stat. 2000. No. 28. P. 1356–1378.