

ИСПОЛЬЗОВАНИЕ МЕТОДА ГРУППОВОГО УЧЕТА АРГУМЕНТОВ ДЛЯ ИЗУЧЕНИЯ ВОСПРИЯТИЯ ЛИЧНОЙ БЕЗОПАСНОСТИ

Бова Андрей Андреевич – кандидат социологических наук, начальник научно-исследовательского отдела Государственного научно-исследовательского института МВД Украины (м. Київ)

В статье рассматриваются различные программные реализации метода группового учета аргументов (МГУА) для изучения зависимости восприятия личной безопасности от уровня убийств, Индекса человеческого развития, доверия к правительству и людям. МГУА как современный индуктивный метод машинного обучения, основанный на принципах самоорганизации, позволяет генерировать полиномиальные регрессионные модели с учетом нелинейного взаимодействия предикторов, искусственные нейронные сети оптимальной сложности и их ансамбли, включать в анализ как очень малые, так и большие выборки, а также обрабатывать большое количество независимых переменных (от нескольких десятков до нескольких сотен). Благодаря использованию полиномиальной нейронной сети удалось повысить долю объясненной дисперсии зависимой переменной с 44 % до 60 %. Эмпирической базой исследования служат обобщенные результаты Всемирного опроса Гэллапа 2011 и материалы Доклада о человеческом развитии 2013 ООН по 115 странам мира. Точность моделей МГУА дополнительно проверена на данных, опубликованных в Докладе о человеческом развитии 2014 года.

Ключевые слова: машинное обучение, метод группового учета аргументов, восприятие личной безопасности, кросс-национальные исследования.

У статті розглядаються різноманітні програмні реалізації методу групового урахування аргументів (МГУА) для вивчення залежності сприйняття особистої безпеки від рівня вбивств, Індексу людського розвитку, довіри до уряду і людей. МГУА як сучасний індуктивний метод машинного навчання, що ґрунтується на принципах самоорганізації, дає змогу генерувати поліноміальні регресійні моделі з урахуванням нелінійної взаємодії предикторів, штучні нейронні мережі оптимальної складності та їхні ансамблі, включати до аналізу як дуже малі, так і великі вибірки, обробляти велику кількість незалежних змінних (від кількох десятків до кількох сотень). Завдяки використанню поліноміальної нейронної мережі вдалося збільшити частку поясненої дисперсії залежної змінної з 44% до 60%. Емпіричною базою дослідження слугують узагальнені результати Всесвітнього опитування Геллапа 2011 і матеріали Доповіді про людський розвиток 2013 ООН по 115 країнам світу. Точність моделей МГУА додатково перевірена на даних, оприлюднених у Доповіді про людський розвиток 2014 року.

Ключові слова: машинне навчання, метод групового урахування аргументів, сприйняття особистої безпеки, крос-національні дослідження, регресійний аналіз.

The article considers various program realizations of the method of Group Method of Data Handling (GMDH) for studying of dependency of perception of personal safety from level of homicides, Human Development Index, trust to the government and people. As empirical base of research, the generalized results of Gallup World Poll 2011 and materials of United Nations Human Development Report 2013 on 115 world countries are used. Machine learning techniques allow to build a more accurate and stable model, but substantive interpretation of these solutions, with some exceptions is limited. GMDH is a modern inductive method of adaptive machine learning can be regarded as advanced regression analyses, which formalizes relationship between the variables in the form of a polynomial. Being based on the principles of self-organization it can generate the polynomial regression models, taking into account the nonlinear interaction of the predictors, artificial neural networks of optimal complexity and its ensembles, to handle a large number of independent variables (from few tens to few hundreds), both very small and large samples. The advantage of the method is to obtain sustainable results on training and examination sample. Correlation analysis reveals a statistically significant positive relationship between the proportion of people who feel safety, who trust the government and other people, the Human Development Index and a negative relationship with level of homicides. Thanks to polynomial neural network it was possible to raise accuracy of the explained variation of dependent variable from 44 % to 60 %. Predictive value of the share of respondents who feel safe walking alone at night in the city or area in Ukraine was 50% with 48 % real. Accuracy of the GMDH models was also tested on data published in the Human Development Report 2014. Modern software implementation of GMDH algorithms as GMDH Shell 3, KnowledgeMiner 3, NeuroShell 2 and Varireg has various types of optimization parameters.

Keywords: machine learning, group method of data handling, perception of personal safety, cross-national research.

Регрессионный анализ является наиболее часто употребляемым методом в эмпирических социальных исследованиях. Практическому применению новых методов регрессионного анализа данных способствуют потребности в объяснении различных социальных явлений, а также рост коммерческого и свободно распространяемого программного обеспечения, который содержит набор различных методов моделирования, отбора информативных признаков, обработки большого числа предикторов и т.д. В ряде случаев нелинейные методы показывают большую прогностическую точность, чем линейная регрессия. Существуют разные подходы к построению нелинейных моделей, например, нелинейные преобразования признаков, выбор лучших аппроксимаций из библиотеки функций, обучение искусственных нейронных сетей, эволюционное моделирование из предварительно заданных грамматик, непараметрическое сглаживание. Исходя из актуальности совершенствования многомерной статистической обработки, автором подготовлен цикл научных статей с результатами тестирования различных методов регрессионного анализа на одних и тех же данных.

Обобщением регрессионных моделей, включающих как взаимодействие предикторов, так и нелинейность, является *метод группового учета аргументов* (разработчик А. Г. Ивахненко) – адаптивная техника машинного обучения на основе автоматической генерации многослойных полиномиальных зависимостей, объединение которых способно воспроизводить оптимальную структуру искусственной нейронной сети с использованием принципов эволюции (наследования и отбора) [1, 2]. Семейство алгоритмов МГУА решает также задачи классификации, прогнозирования временных рядов, выбора потенциальных экзогенных и эндогенных переменных. Самоорганизующиеся модели используются как на выборках небольшого объема (коротких временных рядах), так и на больших наборах данных, при малом или очень большом количестве независимых переменных. Реализация широкого спектра алгоритмов МГУА присутствует в отдельных программах и пакетах анализа данных наряду с другими методами *Data Mining*.

Настройка МГУА включает такие основные этапы.

Преобразование данных – логарифмирование, возведение в степень, взятие квадратного и кубического корня и др., а также масштабирование переменных, что позволяет получить более компактное решение (часто используется и в других топологиях искусственных нейронных сетей и классификации).

Формирование обучающей и экзаменационной выборки. Настройки некоторых алгоритмов позволяют включать в обработку объекты с пропущенными значениями как независимых переменных, так и отклика.

Выбор вида полинома или архитектуры нейронной сети, состоящей из количества слоёв (входной, внутренний, внешний) и функции активации – формулы, по которой производятся расчеты во внутреннем и выходном слое (линейная, полиномиальная, сигмоидная, радиально-базисные функции или их синтез, рациональная функция), максимальное количество нейронов слоя.

Выбор критерия, который необходимо оптимизировать, способа селекции переменных и желаемой сложности модели, определяемой числом параметров.

В результате получается схема нейронной сети с рассчитанными по регрессионному уравнению прогностическими значениями (ответ сети). Для улучшения точности и стабильности моделирования создают композитную модель, в которой прогнозные значения вычисляются как среднее, полученное из нескольких уравнений (ансамбль моделей).

Целью статьи является выявления статистической зависимости субъективного восприятия личной безопасности (*PS*) от Индекса человеческого развития (*HDI*), уровня доверия к национальному правительству (*TNG*) и людям (*TP*), коэффициента убийств (*HR*) с помощью МГУА, сравнение точности результатов некоторых программных реализаций алгоритма между собой и с линейными моделями. Данные результатов Всемирного опроса Гэллага за 2007–2011 г., коэффициента убийств (2004–2011 г.) и интегрального индекса содержатся в Докладе о человеческом развитии 2013 Программы развития ООН, а показатели определялись таким образом [3, с. 114–147, с. 174–177].

Восприятие безопасности – процент респондентов, ответивших «да» на вопрос: «Чувствуете ли вы себя в безопасности, когда прогуливаетесь в одиночестве ночью в городе или районе, в котором вы живете?».

Индекс человеческого развития – комбинированный индекс, измеряющий среднюю величину достижений в трех основных измерениях человеческого развития: здоровье и долголетие, знания и достойные условия жизни, в ответивших «да» на вопрос: «Доверяете ли правительству вашей страны?».

Доверие к людям – процент респондентов, ответивших «да» на вопрос: «В целом, считаете ли вы, что большинству людей можно доверять, или вы считаете, что следует проявлять осторожность, имея дело с людьми?».

Коэффициент убийств – число умышленных убийств, т. е. смертей, незаконно причиненных человеку другим человеком, в пересчете на 100 тыс. чел.

Полные данные доступны для 115 стран. При сравнении прогностической точности уравнений будем использовать коэффициент детерминации R^2 , который вычисляется как коэффициент корреляции между наблюдаемыми значениями и предсказанными по модели, возведенный в квадрат.

В Украине доля тех, кто чувствует себя ночью в безопасности на улице составляет 48 % (в среднем по выборке 60 %). Deskriptivные статистики свидетельствуют о значительной неоднородности информации. Так, коэффициент вариации для *HR* составляет 138 %, *TP* – 49 %, *TNG* – 39 %, *HDI* – 27 %, *PS* – 26 %. По показателю *TP* за границы 2σ выходят 5 резко выделяющихся значений, 3σ – 1; по показателю *TNG* за границы 2σ – 5; по показателю *HR* за границы 2σ – 4, 3σ – 3, 4σ – 2, по показателям *HDI* за границы 2σ – 4, по показателю *PS* за границы 2σ – 2.

Корреляционный анализ обнаруживает положительную статистически значимую связь между долей людей, которые чувствуют себя в безопасности в ночное время с долей респондентов, которые доверяют правительству (0,4) и людям (0,43), Индексом человеческого развития (0,2) и отрицательную связь – с коэффициентом убийств (-0,4).

Для линейной регрессионной модели с оценкой параметров по методу наименьших квадратов R^2 составляет 0,44.

$$\hat{PS} = 27,4 - 0,31HR + 0,32TNG + 0,41TP + 14,8HDI$$

Исходя из значений коэффициентов в стандартизованном масштабе, наибольший вклад в модель вносит колебания уровней доверия населения к национальному правительству (0,39) и людям (0,30), коэффициент убийств (-0,29) и, наконец, Индекс человеческого развития (0,17). Интерпретация направления связи коэффициентов регрессии аналогична коэффициентам корреляции.

Общая линейная модель с включением статистически значимых общих эффектов и эффектов взаимодействия увеличивает R^2 до 0,54.

$$\hat{PS} = -29,5 + 1,6HR + 1,3TNG + 0,33TP + 89,1HDI - 1,7HR \cdot HDI - 0,02HR \cdot TNG - 1,2HDI \cdot TNG$$

По величине бета-коэффициентов переменные расположились таким образом: *TNG* (1,6), *HR* (1,4), *HDI*TP* (1,1), *HDI* (1), *HR*HDI* (1), *TNG*HR* (0,8), *TP* (0,2). Заметим, в частности, что уменьшает долю людей, чувствующих себя в безопасности, одновременно высокие показатели убийств и Индекса человеческого развития, убийств и высокого доверия к национальному правительству.

Эмпирические данные могут описываться разнообразными нелинейными регрессионными уравнениями. Нейронная сеть производит объединение этих уравнений, моделируя сложную поверхность отклика. В программе *Varireg* построена двухслойная нейронная сеть ($R^2 = 0,52$) с шаговым отбором переменных и Байесовским информационным критерием (*Bayesian Information Criterion*), которая представлена в едином полиномиальном уравнении.

$$\hat{PS} = 49,94 + 0,003TNG^2 + 0,008HR^2 + 0,43TPHDI - 1,31HRHDI$$

Точность одной из возможных мультилинейных моделей МГУА программы *KnowledgeMiner (yX) for Excel*, при среднем иммунитете к шуму, $R^2 = 0,54$. Алгебраическое выражение трехслойной сети с полиномиальным нейроном приведено ниже.

$$\begin{aligned}\hat{PS} &= 9,26z_{11} + 14z_{22} + 59,7 \\ z_{11} &= 0,04TP - 0,88 \\ z_{22} &= 0,82z_{11} + 1,16z_{12} + 1,42z_{11}z_{12} \\ z_{11} &= 0,06TNG + 4,71HDI - 0,06HDI \cdot TNG - 4,5 \\ z_{12} &= -0,05HRHDI + 0,3\end{aligned}$$

Исходя из зависимостей исходных переменных с переменными промежуточного слоя, увеличение в обществе межличностного доверия, доверия к государственным органам, длительности жизни, улучшение образования и материального уровня в целом способствует росту субъективного уровня безопасности.

С целью повышения надежности результатов регрессионной модели в программе *GMDH Shell 3* общая выборка 115 стран была разделена на обучающую (80% или 92 объекта) и экзаменационную (20% или 23 объекта) и использован алгоритм МГУА с пошаговым смешанным отбором переменных, который показало соответственно такую точность на указанных выборках – $R^2 = 0,54$ и $R^2 = 0,51$.

$$\hat{PS} = 48,96 + 0,12TNG \sqrt[3]{TP} + 0,1HR \sqrt[3]{TP} + 27HDI^2 - 18,34HDI \sqrt[3]{HR}$$

Значимость переменных определялась по степени влияния на среднеквадратическое отклонение модели: $\sqrt[3]{HR}$ (84%), *TNG* (35,7%), *HR* (12,8%), *HDI* (11,4%), $\sqrt[3]{TP}$ (6%).

Для построения полиномиальной модели МГУА в программе *NeuroShell 2* данные были масштабированы в интервале от -1 до 1 по легенде: $TP_s=2$ ($TP-7$)/53-1; $TNG_s=2$ ($TNG-11$)/84-1; $HR_s=2$

$(HR-0,2)/91,4-1$; $HDI_s = 2 (HDI-0,3)/0,63-1$; $PS_s = 2 (PS-29)/63-1$. С учетом этого получена такая статистическая модель.

$$\begin{aligned} \wedge \\ PS_s = -0,55 + 0,23TP_s - 0,38HR_s - 0,96HDI_s + 0,37HR_s^2 + 0,8HDI_s^3 - \\ - 0,65HR_sTNG_s - 0,44HDI_sTNG_s - 0,62HR_sHDI_s - 0,18TP_sHDI_s \end{aligned}$$

Точность полиномиальной сети из семи слоев, при заданной сильной нелинейности, низкой сложности, среднем разнообразии и «хитрой» оптимизации, рассчитанная по исходным и прогнозным значениям, составляет $R^2 = 0,60$. Прогнозное точечное значение доли респондентов, чувствующих себя ночью в безопасности на улице, для Украины составило 50 %. Наиболее существенной переменной аналитического выражения является межличностное доверие и менее существенными – доверие к правительству, коэффициент убийств, Индекс человеческого развития.

В уравнении присутствуют члены, фиксирующие негативное влияние на изменение доли людей, чувствующих себя ночью на улице в безопасности, совместного взаимодействия факторов $HR*TNG$, $HR*HDI$, $HDI*TNG$, $TP*HDI$. Значения коэффициентов корреляции Пирсона нулевого порядка PS с переменными составляют: $HR*TNG - 0,32$, $HR*HDI - 0,43$, $HDI*TNG - 0,45$, $TP*HDI - 0,45$. Изменение знака коэффициентов регрессии по сравнению с коэффициентами парной корреляции происходит вследствие мультиколлинеарности объясняющих переменных.

Модели МГУА, продуцируемые *GMDH Shell* и *NeuroShell*, валидизированы на частично обновленных данных, содержащихся в Докладе о человеческом развитии 2014 ООН [4, p.160–163, p. 204–207, p. 220–223]. Коэффициенты множественной детерминации на выборке из 116 стран составили соответственно $R^2 = 0,53$ и $R^2 = 0,51$.

На основе изложенного можно сделать не только такие частные, по отношению к исследуемым данным, но и общие, по отношению к анализу социологической информации, выводы.

1. Социальные явления не подвержены жесткому детерминизму, а скорее описываются мягкими закономерностями. Неоднородность выборки, погрешность измерения, поисковый характер многих исследований, недостаточная разработка теории, большое количество факторов и их нелинейное взаимное влияние на отклик, невозможность экспериментальной проверки, а также многие другие причины часто не позволяют достичь достоверных статистических результатов и существенно ограничивают точность прогнозов в социальных науках. Линейное регрессионное уравнение не всегда адекватно описывает эмпирические закономерности (особенно в ситуации проверки модели на новых наблюдениях).

2. Методы машинного обучения позволяют строить более точные и стабильные модели, устойчивые в том числе, и к резко выделяющимся наблюдениям, однако содержательная интерпретация таких решений, за некоторым исключением ограничена, поскольку совместное взаимодействие переменных носит достаточно сложный характер. Для искусственных нейронных сетей, например, фиксируется лишь направление влияния входных переменных на нейроны промежуточного слоя. Использование МГУА позволило повысить на обучающей выборке долю объясненной дисперсии зависимой переменной с 44 % до 60 %. Одновременно высокие показатели Индекса человеческого развития и доверия к правительству, а также Индекса человеческого развития и коэффициента убийств в некоторых обществах снижают чувство личной безопасности. Отметим также, что с МГУА могут конкурировать другие методы машинного обучения, позволяющие по новому взглянуть на ранее полученные результаты в области социальных наук. В частности, хорошим дополнением глобальных, релевантных для всей выборки, моделей являются регрессионные уравнения для типологических однородных групп, которые позволяют подробнее изучить найденные взаимодействия [5].

3. Обнаружен некий парадокс между точными моделями, которые не совсем соотносятся с объяснительной схемой из-за мультиколлинеарности, и менее точной, но информативно правильной линейной моделью. Данные описываются одновременно различными аналитическими моделями, а важность переменных зависит от используемого метода или коллектива моделей. Выбор регрессионного уравнения может диктоваться желанием получить наиболее точную модель, понятную содержательную интерпретацию, которая соответствует развитию теории, или же модель, состоящую из переменных, измерение которых не несет ощутимых затрат. Своеобразный баланс между простотой и точностью модели может быть достигнут путем подачи в научной публикации, с одной стороны, матрицы корреляций факторов с откликом и, с другой – основных параметров и результатов работы сложных алгоритмов машинного обучения, в том числе, ансамбля моделей.

ЛИТЕРАТУРА: 1. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем. – Киев: Наукова думка, 1981. – 296 с. 2. Anastasakis L., Mort N. The Development of Self-organization Techniques in Modeling: A Review of the Group Method of Data Handling (GMDH). Research Report No. 813. Department of Automatic Control & Systems Engineering, The University of Sheffield, United Kingdom. – 2001. – October. – 38 p. 3. Доклад о человеческом развитии 2013. Возвышение Юга: человеческий прогресс в многообразном мире / [директор и основной автор Х. Малик]. – М.: Издательство «Весь Мир», 2013. – 202 с. 4. Human Development Report 2014. «Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience» / Director and lead author Khalid Malik – Publisher: United Nations Development Programme, 2014 – 236 p. 5. Бова А.А. Страх перед злочинністю, кримінальна віктимізація та довіра до соціальних інститутів: результати порівняльного дослідження // Методологія, теорія та практика соціологічного аналізу сучасного суспільства. – 2013. – Випуск 19. – С. 346–252.