4. Степанов Е.А., Гамова А.Н. Алгоритм декодирования по максимуму апостериорной вероятности. // Ломоносовские чтения на Алтае: фундаментальные проблемы науки и образования: избранные труды международной конференции, Барнаул, 11–14 ноября 2017 г. – Барнаул: Изд-во Алтайского университета, 2017. С. 818-828.

УДК 51-74, 004.6

Наука о данных в контексте статистики

Т.О. Сундукова, Г.В. Ваныкина
Тульский государственный педагогический университет им. Л.Н. Толстого, г. Тула

Наука о данных как научная дисциплина находится под влиянием информатики, математики, исследования операций и статистики, а также прикладных наук. В 1996 году термин Data Science впервые был включен в название статистической конференции (Международная федерация классификационных обществ (IFCS) «Data Science, классификация и смежные методы») [1]. Несмотря на то, что этот термин был основан статистиками в общедоступном имидже Data Science, важность компьютерных наук и бизнес-приложений часто гораздо больше подчеркивается, особенно в эпоху больших данных. В 1970-х годах идеи J.W. Tukey [2] изменили точку зрения статистики с чисто математической установки, например, статистического тестирования, на вывод гипотез из данных (исследовательская установка), то есть первично понять данные, прежде чем выдвигать гипотезы. Другим этимологическим корнем Data Science является «Обнаружение знаний в базах данных» (Knowledge Discovery in Databases – KDD) [3] с его подтемой Data Mining. KDD уже объединяет множество различных подходов к обнаружению знаний, включая индуктивное обучение, (байесовскую) статистику, оптимизацию запросов, экспертные системы, теорию информации и нечеткие множества. Таким образом, KDD является большим строительным блоком для стимулирования взаимодействия между различными областями для общей цели выявления знаний в данных.

В настоящее время эти идеи объединены в понятии Data Science, что приводит к различным определениям. Одно из наиболее полных определений науки о данных было недавно дано L. Cao в виде формулы [4, с. 28]:

Наука о данных = (статистика + информатика + вычисления + коммуникация + социология + управление) / (данные + среда + мышление).

В этой формуле социология выступает за социальные аспекты и знаменатель (данные + среда + мышление) означает, что все упомянутые науки действуют на основе данных, среды и так называемого мышления «данные-знания-мудрость».

Обзор Data Science, представленный D. Donoho в 2015 году [5], посвящен эволюции Data Science из статистики. Действительно, еще в 1997 году существовал еще более радикальный взгляд, предлагающий переименовать статистику в Data Science [6]. В 2015 году лидеры ASA (American Society of Appraisers – Американское общество оценщиков) [7] опубликовали заявление о роли статистики в науке о данных, заявив, что статистика и машинное обучение играют центральную роль в науке о данных. На наш взгляд, статистические методы имеют решающее значение на самых фундаментальных этапах науки о данных. Следовательно, сформулируем предпосылку: статистика является одной из наиболее важных дисциплин, предоставляющих инструменты и методы для поиска структуры и более глубокого понимания данных, а также наиболее важной дисциплиной для анализа и количественной оценки неопределенности. Рассмотрим влияние статистики на наиболее важные этапы в науке о данных.

Одним из предшественников Data Science со структурной точки зрения является известный CRISP-DM (Cross-Industry Standard Process for Data Mining – Межотраслевой стандартный процесс для интеллектуального анализа данных), который состоит из шести основных этапов: понимание бизнеса, понимание данных, подготовка данных, моделирование, оценка и развертывание [8]. Такие идеи, как CRISP-DM, теперь являются фундаментальными для прикладной статистики.

На наш взгляд, основные шаги в науке о данных были основаны и развивались на CRISP-DM, что привело, к вариативности определений науки о данных как последовательности следующих этапов: первый вариант – сбор и обогащение данных, хранение и доступ к данным, результирующие данные; второй вариант – исследование, анализ и моделирование данных, оптимизация алгоритмов, проверка и отбор моделей, представление и отчетность по результатам, а также бизнес-развертывание результатов.

Обычно эти этапы не просто выполняются один раз, а повторяются в циклическом цикле. Кроме того, принято чередовать два или более шагов. Это особенно верно для этапов сбора и обогащения данных, исследования данных и статистического анализа данных, а также для статистического анализа данных и моделирования, а также проверки и выбора моделей.

Таблица 1 сравнивает различные определения этапов в Data Science. Взаимосвязь терминов обозначена горизонтальными блоками. Отсутствующий шаг Сбор и обогащение данных в CRISP-DM указывает, что эта схема имеет дело только с данными наблюдений. В данном контексте этапы Хранение и доступ к данным и Оптимизация алгоритмов добавлены в CRISP-DM, где статистика задействована меньше.

Список этапов для Data Science расширен L. Cao [4, с. 34]: прикладные задачи и проблемы в конкретных областях, хранение данных и управление ими, повышение качества данных, моделирование и представление данных, глубокая аналитика, обучение и обнаружение, моделирование и проектирование экспериментов, высокая производительность, обработка и аналитика, коммуникации, данные к решению и действия. В дальнейшем выделим роль статистики, обсуждающую все этапы, в которых она активно задействована.

Таблица 1 – Шаги в Data Science: сравнение CRISP-DM, определение L. Cao и обобщение

CRISP-DM	Определение L. Cao	Обобщение и выводы
Понимание бизнес-процессов	Данные, приложения и задачи	Сбор и обогащение данных
	Хранение данных и управление	Хранение данных и доступ
Понимание данных, подготовка данных	Улучшение качества данных	Исследование данных
Моделирование	Моделирование и представление данных, глубокая аналитика, обучение и открытие	Анализ данных и моделирование
	Высокопроизводительная обработка и аналитика	Оптимизация алгоритмов
Оценка	Симуляция и дизайн эксперимента	Проверка и выбор модели
Внедрение	Сеть, коммуникации	Представление и отчетность о результатах
	Данные к решению и действия	Развертывание результатов бизнес-процессов

Роль статистики в науке о данных недооценивается, например, по сравнению с информатикой. Это дает, в частности, области сбора и обогащения данных, а также расширенное моделирование, необходимое для прогнозирования. Только дополнение и / или объединение математических методов и вычислительных алгоритмов со статисти-

ческим обоснованием, особенно для больших данных, приведет к научным результатам, основанным на подходящих подходах. В конечном счете, только сбалансированное взаимодействие всех вовлеченных наук приведет к успешным решениям в науке о данных.

Библиографический список

1. Press, G.: A Very Short History of Data Science // Forbes. May 28, 2013. – URL: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#5ec53e6055cf

2. Tukey J. W. Exploratory Data Analysis. Addison – Wesley Publishing Company Reading, Mass. — Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney, 1977. – 688 с.

3. Piateski G., Frawley W. Knowledge Discovery in Databases. – MIT Press, Cambridge, – 1991. – 540 с.

4. Cao L. Data science: a comprehensive overview //ACM Computing Surveys (CSUR). – 2017. – Т. 50. – №. 3. – С. 1-42.

5. Donoho D. 50 years of data science //Journal of Computational and Graphical Statistics. – 2017. – Т. 26. – №. 4. – С. 745-766.

6. Wu J. Statistics = data science? – 1997. – URL: http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

7. Van Dyk D., Fuentes M., Jordan M. I., Newton M., Ray B. K., Lang D. T., Wickham H. ASA Statement on the Role of Statistics in Data Science //Amstat news. – 2015. – Т. 460. – №. 9. – 24 с.

8. Brown M. S. Data mining for dummies. – John Wiley & Sons, London. – 2014. – 410 с.

УДК 519.24; 004.67

Сравнительный анализ методов оценки причинного эффекта: оценка вклада элементов интенсификации технологии в урожайность яровой пшеницы

К.О. Тарасов, Е.В. Понькина

АлтГУ, г. Барнаул

Аннотация. Задача оценки причинных эффектов представляет собой сравнение состояния объекта с учетом и без учета вмешательства и оценки ожидаемой величины полученных различий целевого признака. В работе рассматривается сравнительный анализ методов оценки причинных эффектов, включая тесты попарных сравнений, линейные регрессионные модели и метод псевдорандомизации