6. Tagiew R., Buder T., Tilly R., Hofmann K., Klotz C. Datensätze für das autonome Fahren als Grundlage für GoA3+ // Eisenbahntechnische Rundschau. 2021.№ 9.

Е.В. Орлова

ВЫЯВЛЕНИЕ ПРИЧИННЫХ СВЯЗЕЙ В РАНДОМИЗИРОВАННЫХ ИСПЫТАНИЯХ НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

(Уфимский государственный авиационный технический университет)

Аннотация. Описан метод статистического эксперимента - рандомизированное испытание, который используется для тестирования эффективности решений и выбора наилучшего решения из множества возможных в условиях неоднородных статистических данных. В этом случае решение выступает в качестве фактора-причины, а эффективность функционирования объекта является фактором-следствием. Обработку результатов таких испытаний можно проводить с помощью методов машинного обучения. Применение методов машинного обучения позволит исключить возможные ложные корреляции факторов, идентифицировать истинные причинные зависимости и установить наиболее эффективное решение.

Ключевые слова: статистический эксперимент, рандомизированное испытание, методы машинного обучения; кластеризация; классификация

Неоднородность статистической выборки характеризуется наличием в ней выбросов (то есть резко выделяющихся значений) влияют на результаты исследования. В соответствии с центральной предельной теоремой (ЦПТ) средние значения, вынутые из многочисленных выборок, будут иметь форму нормального распределения, даже если исходная совокупность элементов не является нормально распределенной. Условием применения ЦПТ являются: большой объем выборок и отклонения данных от нормального распределения не большое. В соответствии с ЦПТ используются формулы аппроксимации нормальным распределением (*t*-распределение), с помощью чего реализуются вычисления доверительных интервалов и осуществляется проверка статистических гипотез.

Параметрические статистические методы не обладают свойством устойчивости (робастности), т.е. результаты моделирования не всегда одинаковы при допустимых отклонениях реальных данных. А методы аппроксимации – метод наименьших квадратов, метод максимального правдоподобия, являются очень чувствительными к неоднородности данных. Для характеристики центра распределения и разброса могут быть использованы показатели медианы и интерквартильный размах.

Проблема неоднородности статистических выборок может разрешаться с помощью применения методов кластеризации, а также на основе методов непараметрической статистики и методов статистического машинного обучения. Статистические методы машинного обучения как часть методов науки о данных отличаются от классических статистических методов тем, что они основаны на данных и не стремятся описывать эти данные с помощью линейной или другой общей функции. Машинное обучение, как правило, уделяет большое внимание разработке эффективных алгоритмов, которые масштабируются для больших объемов данных, чтобы оптимизировать прогностическую модель.

Для тестирования эффективности решений и выбора из них наилучшего используется тип статистического эксперимента – рандомизированное испытание. Особенностью такого испытания является снижение источников систематической ошибки при оценке эффективности проектов, реформ, последствий проводимых изменений, программ. Дизайн рандомизированного испытания следующий:

1. Оцениваемых субъектов (представительную выборку) случайным образом распределяют в две или более группы, которым задают разные условия.

2. В одной группе – экспериментальной (treatment) осуществляют воздействия в соответствии с исследуемой программой, во второй – контрольной (control) – никаких воздействий не осуществляют.

3. Затем сравнивают результаты в этих группах и делают вывод о влиянии воздействий (причин) на эффективность функционирования субъектов (следствий).

Обработку результатов таких испытаний можно проводить с помощью методов машинного обучения. Применение этих методов позволит исключить возможные ложные корреляции факторов и идентифицировать истинные причинные зависимости. Это возможно за счет проведения множества экспериментов, например, на основе алгоритмов бутсрапирования, или применения бэггинга, моделей случайных лесов, позволяющих надежно установить зависимости факторов и выявить наиболее эффективное решений.

Ниже приводится краткое описание методов машинного обучения, которые могут использоваться для выявления причинных зависимостей.

Деревья решений (decision trees). Модели дерева решений представляет собой модель классификации и прогнозного моделирования. Модель дерева решений основана на рекурсивном разбиении, то есть многократном разделении данных на разделы и подразделы с целью создания однородных классов в каждом сводном подразделе. Древовидная модель представляет собой набор импликационных правил «если-то-иначе». Деревья способны обнаруживать скрытые закономерности, соответствующие сложным взаимодействиям данных. Модель может быть выражена в терминах отношений между предикторами, которые хорошо интерпретируются. Алгоритм рекурсивного разбиения для построения дерева решений довольно интуитивно понятен. Данные разделяются несколько раз с использованием значений предикторов, которые разбивают данные на относительно однородные сегменты. Существуют различные индук-

торы деревьев решений сверху вниз, такие как ID3, C4.5, CART. Некоторые состоят из двух концептуальных фаз: выращивания и обрезки (C4.5 и CART). Другие индукторы выполняют только фазу роста. Подробные алгоритмы реализации деревьев решений можно найти в [1].

Случайный лес (random forest) представляет собой тип бутстрап-агрегированной оценки на основе деревьев решений. Основан на применении бэггинга к деревьям решений с одним расширением – алгоритм случайного леса предполагает не только отбор объектов, но и отбор их переменных.

Метод опорных векторов (*support vector machines*). Это набор контролируемых методов обучения, используемых для классификации и регрессионного анализа. Основная идея метода заключается в построении гиперплоскости, которая оптимальным образом разделяет объекты выборки. Алгоритм работает в предположении, что чем больше расстояние (зазор) между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше средняя ошибка классификатора [2].

Преимущества метода заключаются в следующем:

- метод эффективен для данных большой размерности;
- эффективен, если количество измерений превышает количество выборок;
- использует подмножество обучающего набора в функции принятия решения (называемом опорными векторами), поэтому также эффективно использует память;
- для функции принятия решения могут быть указаны различные функции ядра.

К недостаткам метода относятся следующие:

1) если количество факторов намного больше, чем количество выборок, может возникнуть проблема переобучения при выборе признаков ядра;

2) не дает прямых оценок вероятностей.

Метод k-средних (k-means method) является одним из широко используемых методов кластеризации [3]. Алгоритм разбивает множество элементов векторного пространства на заданные k кластеров. Он делит данные на k кластеров путем минимизации суммы квадратов расстояний каждого объекта до среднего значения назначенного ему кластера. Основная идея заключается в том, что на каждой итерации пересчитывается центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы снова разбиваются на кластеры, в соответствии с которыми новые центры ближе по выбранной метрике. Алгоритм завершается, когда на некоторой итерации внутрикластерное расстояние не меняется за конечное число итераций, так как число возможных разбиений конечного множества конечно, и на каждом шаге общее стандартное отклонение уменьшается, поэтому зацикливание невозможно.

Стохастический градиентный бустинг (Stochastic Gradient Boosting). Метрические классификаторы просты в применении, в качестве методов обучения используют анализ сходств объектов в выборке, но не обладает гибкостью, неустойчивы к шумам и выбросам в исходных данных. Линейные классифика-

торы являются гибкими алгоритмами, однако они ограничены тем, что относят объекты к одному из двух классов, то есть используются для бинарной классификации. Бустинг позволяет объединить слабые классификаторы в один сильный и на основе такого объединения позволяет устранить недостатки каждого алгоритма [4-8]. Он основан на композиций разных алгоритмов для компенсации проблем каждого из них.

Самоорганизующиеся карты (self-organising maps) представляют собой разновидность нейросетевых алгоритмов [9]. Основное отличие этой технологии от нейронных сетей, обученных по алгоритму обратного распространения, заключается в том, что метод обучения является неконтролируемым, то есть результат обучения зависит только от структуры входных данных. Алгоритм функционирования самоорганизующихся карт является одним из вариантов кластеризации многомерных векторов. Примером таких алгоритмов является алгоритм *k*-средних. Важным отличием алгоритма является то, что в нем все нейроны (узлы, центры классов) упорядочены в некоторую структуру (обычно двумерную сетку). При обучении модифицируется не только нейрон-победитель, но и его соседи, но в меньшей степени. Метод можно рассматривать как один из способов проецирования многомерного пространства в пространство с меньшей размерностью. При использовании этого алгоритма векторы, сходные в исходном пространстве, на результирующей карте оказываются рядом.

Литература

1. Rokach L.; Maimon O. Decision Trees, 2005. – DOI: 10.1007/0-387-25465-X_9.

2. Awad M.; Khanna R. Support Vector Machines for Classification, 2015. – DOI: 10.1007/978-1-4302-5990-9 3.

3. Adam C.; Andrew Y. Ng. Learning Feature Representations with K-means. – Stanford University, USA, 2012, 20 p.

4. Orlova E.V. Methodology and Models for Individuals' Creditworthiness Management Using Digital Footprint Data and Machine Learning Methods // Mathematics. – 2021. –Vol. 9. – No. 15. – 28 p. https://doi.org/10.3390/math9151820

5. Orlova E.V. Decision-Making Techniques for Credit Resource Management Using Machine Learning and Optimization // Information. – 2020. – Vol. 11. – № 3. – 17 p. – DOI: https://doi.org/10.3390/info11030144

6. Orlova E.V. Innovation in Company Labor Productivity Management: Data Science Methods Application // Applied Sysem Innovation. – 2021. – Vol. 4. – No 3: 68. – 18 p. https://doi.org/10.3390/ asi4030068

7. Орлова Е.В. Модели и механизмы согласованного управления производственно-экономической системой: дис. на соискание уч. степени д-ра техн. наук. – Уфа, 2018. – 340 с.

8. Орлова Е.В. Предиктивная аналитика кредитных рисков на основе данных цифровых следов заемщиков и методов статистического машинного обучения // Программная инженерия. – 2021. – Том 12. – № 7. – С. 358-372. – DOI: 10.17587/prin.12.358-372

9. Kohonen, T. Self-Organizing Maps (Third Extended Edition). – New York, 2001, 501 p.

А.С. Подгорный

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА РАСПОЗНАВАНИЯ РЕЧИ В РОБОТЕ АССИСТЕНТЕ

(Уфимский государственный авиационный технический университет)

Создание ассистента, способного распознавать человеческий голос и отвечать на него всегда было сложной задачей, по большей части из-за несовершенства технологий распознавания речи. Но с тех пор интеллектуальные технологии сильно продвинулись, благодаря чему создание собственного ассистента стало под силу даже студенту, причём абсолютно бесплатно.

Размещение ассистента на базе робота позволяет сделать его мобильным, а использование полноценного микрокомпьютера позволяет значительно расширить набор датчиков и устройств, которыми можно дополнять функционал робота.

Робот перемещается на гусеничном шасси (рис.1). В его основе стоит платформа *Nvidia Jetson*[2], содержащий в себе средство для ускорения работы нейросетей, что не обязательно для работы распознавания голоса, зато позволяет тратить меньше энергии на обработку и выполнять более сложные задачи в будущем.

Рис. 1. Внешний вид робота