Библиографические ссылки

1. Официальный сайт компании Maxgen Promo [Электронный ресурс]. – Режим доступа : https://maxgenpromo.com/ Дата доступа : 31.09.2019.
2. Официальный сайт Яндекс.Директ [Электронный ресурс]. – Режим доступа : https://direct.yandex.ru/ Дата доступа : 31.09.2019.
3. Рындина, С. В. Электронный бизнес: создание, развитие и продвижение цифровых продуктов : учеб. пособие / С. В. Рындина. – Пенза : Изд-во ПГУ, 2019. – 88 с.
4. McClure, D. Startup Metrics for Pirates: AARRR! [Electronic resource]] – Access mode : https://www.youtube.com/watch?v=irjgfW0BIrw Date of access : 31.09.2019.

УДК 330.4
JEL – C6

О РАЗДЕЛЕНИИ СОВОКУПНОЙ ВЫБОРКИ ДАННЫХ НА ТРЕНИРОВОЧНУЮ, КОНТРОЛЬНУЮ И ПРОВЕРОЧНУЮ ВЫБОРКИ

В. О. Сувалов[1)], И. А. Карачун[2)]

[1)] *Аспирант экономического факультета Белорусского государственного университета, г. Минск*
[2)] *Кандидат экономических наук, доцент, заведующий кафедрой цифровой экономики Белорусского государственного университета, г. Минск*

Работа посвящена современным подходам к разделению массива данных на тренировочную, контрольную и проверочную выборки, применяемые в ходе машинного обучения для целей прогнозирования. Рассматриваются методы определения оптимального соотношения данного разделения.

Ключевые слова: машинное обучение; большие данные; анализ данных; тренировочная выборка; контрольная выборка; проверочная выборка.

ON THE DIVISION OF AN AGGREGATE DATA SAMPLE FOR THE TRAINING, CONTROL AND VALIDATION SAMPLES

V. Suvalov[1)], I. Karachun [2)]

[1)] *Postgraduate Student of the Faculty of Economics, Belarusian State University, Minsk*
[2)] *PhD in Economics, Associate Professor, Head of Digital Economy Department Belarusian State University, Minsk*

This paper is devoted to modern approaches to dividing the data array into training, control and verification samples used in machine learning for forecasting purposes. The methods of determining the optimal ratio of this separation are considered.

Key words: machine learning; big data; data analysis; training sample; control sample; validation sample.

В рамках углубления цифровой трансформации и с учетом достижений последних лет в сборе, обработке и хранении данных происходит накопление огромных массивов данных, получивших название больших данных (big data). С учетом того, что такие объемы информации практически невозможно проанализировать вручную, в профессиональной среде анализа данных все чаще обращаются к методам машинного обучения.

Данные методы позволяют увеличить скорость обработки и анализа информации, но в тоже время ставят перед исследователями новые вопросы. Среди таких актуальных вопросов следует указать проблему выбора оптимального разделения всей имеющейся совокупности данных на тренировочную, контрольную и проверочную выборки. Первая используется для обучения моделей, а вторая применяется для оценки прогнозных качеств полученных моделей. Таким образом производится калибровка параметров применяемых моделей. Проверочная выборка необходима для определения лучшей из всех построенных по обучающей выборки модели прогнозирования.

Разделение всей совокупности данных необходимо для того, чтобы получить независимые выборки для разных этапов построения моделей. В противном случае оценки качества построенных моделей и оценка выбранной оптимальной модели становятся смещенными.

Вопрос выбора оптимального соотношения разделения на текущий момент является открытым. На практике исследователи прибегают к процентному разделению выборки, например в соотношении 50 %–25 %–25 %, или 50 %–30 %–20 %, или 33,(3) %–33,(3) %–33,(3) %. В конечном итоге выбор конкретного варианта связан с тем, на сколько велика имеющаяся совокупность данных, и зависит от профессионального мнения исследователя.

В тоже время существуют отдельные исследования, направленные на поиск подхода к оптимальному разделению всей исследуемой совокупности не необходимые части. Так в исследовании Kevin K Dobbin и Richard M Simon [1], основанном на анализе влияния пропорции разделения данных на среднеквадратичную ошибку (MSE) оценки точности прогноза, утверждается, что для набора данных с более чем 100 наблюдениями оптимальным остается отделение 2/3 общей выборки в качестве тренировочной. Кроме того исследователями было установлено, что оптимальная пропорция зависит от размера полного набора данных и точности классификации (или прогноза) – с требованием к более высокой точности и меньшим набором наблюдений, требуется больше данных для включения в тренировочную выборку.

С другой стороны в исследовании Georgios Afendras, Marianthi Markatou [2] приводят теоретические и эмпирические обоснования оптимального разбиения выборки данных с применением метода перекрестной проверки (cross-validation).. Авторы, решая задачу оптимизации для определения оптимального размера выборки, приходят к выводу, что для широкого класса функций оптимальный размер обучающей выборки равен половине общего размера выборки, независимо от распределения данных. Таким образом, стремясь установить правила, позволяющие «оптимально» выбирать размер тренировочной выборки для фиксированного набора данных размера n,.авторы подтверждают мнение своих коллег-практиков о необходимости отделения лишь половины данных для целей обучения.

Таким образом следует заключить, что в зависимости от применяемого критерия исследователи приходят к различным выводам об оптимальной пропорции разделения выборки для целей обучения. Это подтверждает необходимость дальнейшего изучения для поиска единого подхода.

Библиографические ссылки

1. Dobbin K. K. Optimally splitting cases for training and testing high dimensional classifiers / K. K. Dobbin, R. M. Simon – BMC Med Genomics – 2011. – Vol. 4 (31).

2. Afendras G. Optimality of training/test size and resampling effectiveness in cross-validation / G. Afendras, M. Markatou. – Journal of Statistical Planning and Inference. – 2019. – Vol. 199. – P. 286–301.

УДК 316.334.3

ВЛИЯНИЕ ВИРТУАЛЬНЫХ ЦИФРОВЫХ СОЦИАЛЬНЫХ ПРОСТРАНСТВ НА ПРОТЕСТНЫЕ НАСТРОЕНИЯ МОЛОДЕЖИ ПРОВИНЦИАЛЬНОГО ГОРОДА

М. А. Танина[1)], И. А. Юрасов[2)], В. А. Юдина[3)]

[1)] *Кандидат экономических наук, доцент, доцент кафедры менеджмента, информатики и общегуманитарных наук Пензенского филиала ФГОБУ ВО «Финансовый университет при Правительстве РФ», г. Пенза (Россия)*

[2)] *Доктор социологических наук, доцент, профессор кафедры менеджмента, информатики и общегуманитарных наук Пензенского филиала ФГОБУ ВО «Финансовый университет при Правительстве РФ», г. Пенза (Россия)*

[3)] *Кандидат экономических наук, доцент, доцент кафедры менеджмента, информатики и общегуманитарных наук Пензенского филиала ФГОБУ ВО «Финансовый университет при Правительстве РФ», г. Пенза (Россия)*