

НЕКОТОРЫЕ АСПЕКТЫ РАЗДЕЛЕНИЯ СОВОКУПНОЙ ВЫБОРКИ ПРИ МАШИННОМ ОБУЧЕНИИ

В. О. Сувалов

Белорусский государственный университет, г. Минск;

valentin.suvalau@gmail.com;

науч. рук. – И.А. Карачун, канд. экон. наук, доц.

Работа посвящена современным подходам к разделению массива данных на тренировочную, контрольную и проверочную выборки, применяемых 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.

В условиях стремительного углубления цифровой трансформации современные исследователи сталкиваются с изменением стоящих перед ними вызовов. Достижения последних лет в сборе, обработке и хранении данных привели к возможности накапливать и обрабатывать огромные массивы данных. Данный феномен получил название «большие данные» («big data»). Обработка таких больших объемов информации, с целью поиска имеющихся в ней закономерностей, вручную становится слишком трудозатратной и относительно медленной. В этой связи все большую актуальность приобретают методы программной обработки данных и, в частности, методы машинного обучения.

Методы машинного обучения позволяют увеличить скорость обработки и анализа информации, но в тоже время ставят перед исследователями новые вызовы. Среди таких актуальных вопросов следует указать проблему выбора оптимального разделения всей имеющейся совокупности данных на тренировочную, контрольную и проверочную выборки.

В общем виде классический процесс построения модели проиллюстрирован на рисунке 1. Дополнительно следует отметить, что в рамках машинного обучения возможен отход от представленной схемы.

Каждая из упомянутых выборок применяется на своем этапе построения модели. Так, тренировочная выборка применяется на этапах спецификации и параметризации модели, т.е. для так называемого обучения моделей. Контрольная выборка применяется для оценки прогнозных или описательных качеств полученных моделей, что соответствует этапу верификации модели. В свою очередь проверочная выборка необходима для определения лучшей модели из всей совокупности моделей, построенных на предыдущих этапах, что можно интерпретировать как этап валидации.



Рис. 1. Порядок построения модели

Разделение всей совокупности данных необходимо для получения независимых выборок для разных этапов построения моделей. В противном случае оценки качества построенных моделей и оценка выбранной оптимальной модели становятся смещенными, что не позволит использовать их в дальнейшем для целей исследователя.

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

В тоже время следует отметить существование отдельных исследований, направленных на поиск подхода к оптимальному разделению всей исследуемой совокупности на необходимые части. Так в исследовании Kevin K. Dobbin и Richard M. Simon [1], производится анализ влияния пропорции разделения данных на среднеквадратичную ошибку (MSE) оценки точности прогноза. Авторы утверждают, что для набора данных с более чем 100 наблюдениями, оптимальным является отделение $2/3$ общей выборки в качестве тренировочной. Кроме того, исследователями было подтверждено, что оптимальная пропорция разделения коррелирует с размером полного набора данных и требуемой точностью модели. Так, чем меньше общая совокупность данных и выше требование к точности прогноза, тем больше данных необходимо включить в тренировочную выборку. В тоже время следует отметить, что исследователи не ставили целью нахождение оптимального соотношения из всех возможных. В рамках данной работы рассматривались лишь два варианта стратегий разделения: в качестве тренировочной выборки использовали или $1/2$, или $2/3$ все имеющейся совокупности. В тоже время отметим, что

вопрос определения контрольной выборки в данной работе не рассматривался.

В альтернативном исследовании Georgios Afendras и 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. В тоже время, как и в исследовании Dobbin и Simon не рассматривается вопрос определения контрольной выборки.

Таким образом следует заключить, что в зависимости от применяемого критерия исследователи приходят к различным выводам об оптимальной пропорции разделения выборки для целей обучения. Кроме того, требуется дальнейшее исследование, касающееся вопроса определения оптимальной пропорции контрольной выборки. Для решения данного вопроса на текущем этапе проходит процесс построения программы для поиска оптимального соотношения в разделении всей имеющейся совокупности данных на тренировочную, проверочную и контрольную выборки на основе данных по ликвидности банковского сектора Республики Беларусь.

Библиографические ссылки

1. Dobbin K. K., Simon R. M. Optimally splitting cases for training and testing high dimensional classifiers // BMC Med Genomics. 2011. Vol. 4 (31). DOI <https://doi.org/10.1186/1755-8794-4-31>.
2. Afendras G., Markatou M. Optimality of training/test size and resampling effectiveness in cross-validation // Journal of Statistical Planning and Inference. 2019. Vol. 199. P. 286–301. arXiv:1511.02980.