УДК 330.4

С. Математические и количественные методы
C. Mathematical and Quantitative Methods

К ВОПРОСУ ОБ ОПТИМАЛЬНОМ РАЗДЕЛЕНИИ СОВОКУПНОЙ ВЫБОРКИ ПРИ МАШИННОМ ОБУЧЕНИИ

В. О. СУВАЛОВ[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Изучаются современные подходы к разделению массива данных на тренировочную, контрольную и проверочную выборки, применяемые в ходе машинного обучения для целей прогнозирования. Рассматривается актуальный вопрос выбора подходящего разделения всей имеющейся совокупности данных на названные выборки. Анализируются результаты работы программного алгоритма, разработанного для поиска оптимального разделения массива данных на отдельные выборки в целях машинного обучения прогнозирующих моделей. Дается рекомендация отделять 80 % совокупной выборки, чтобы минимизировать ошибки прогноза разрабатываемых моделей.

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

Благодарность. Автор выражает благодарность специалисту первой категории управления регулирования ликвидности Национального банка Республики Беларусь Е. С. Боголюбской-Синяковой и заместителю начальника управления регулирования ликвидности Национального банка Республики Беларусь А. А. Казакевичу за их комментарии и предложения, которые помогли улучшить статью.

TO THE QUESTION OF OPTIMAL DIVISION OF THE FULL SAMPLE DURING MACHINE LEARNING

V. O. SUVALOV[a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

The article is devoted to modern approaches to dividing of a data set into training, control and test samples, used in the process of machine learning for forecasting purposes. The actual issue of choosing the optimal division of the entire available set of data into named above samples is considered. The author analyses the results of the operation of a software algorithm developed to find the optimal division of a data set into isolated samples for the purposes of machine learning of predictive models. A recommendation to separate 80 % of the total sample to minimise the forecast error of the developed models is given.

Keywords: machine learning; big data; data analysis; econometric tools; training sample; control sample; test sample.

Acknowledgements. The author is grateful to the 1st category specialist of the liquidity regulation department of the National Bank of the Republic of Belarus K. S. Bogolyubskaya-Sinyakova and deputy head of the liquidity regulation department of the National Bank of the Republic of Belarus A. A. Kazakevich for their valuable comments and suggestions, which helped to improve the article.

Образец цитирования:
Сувалов ВО. К вопросу об оптимальном разделении совокупной выборки при машинном обучении. *Журнал Белорусского государственного университета. Экономика*. 2021;1:37–45.

For citation:
Suvalov VO. To the question of optimal division of the full sample during machine learning. *Journal of the Belarusian State University. Economics.* 2021;1:37–45. Russian.

Автор:
Валентин Олегович Сувалов – аспирант кафедры цифровой экономики экономического факультета. Научный руководитель – кандидат экономических наук, доцент И. А. Карачун.

Author:
Valentin O. Suvalov, postgraduate student at the department of digital economy, faculty of economics.
https://orcid.org/0000-0001-6748-5805
valentin.suvalau@gmail.com

Введение

В условиях стремительной 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 их массивы. Этот феномен получил название «большие данные» (*big data*). Обработка таких больших объемов информации вручную для поиска имеющихся в ней закономерностей становится слишком трудозатратной и относительно медленной. В связи с этим 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 задачи. Среди них следует отметить выбор оптимального разделения всей имеющейся совокупности данных на тренировочную, контрольную и проверочную выборки.

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

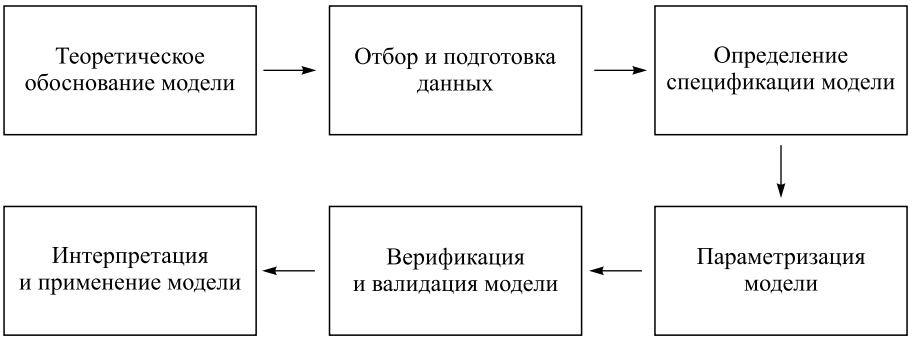

Рис. 1. Порядок построения модели
Fig. 1. Model-building order

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

Разделение совокупности данных необходимо при получении независимых выборок для разных этапов построения моделей. В противном случае оценки качества построенных моделей и выбранной оптимальной модели смещаются, что не позволяет использовать их в дальнейшем для целей исследователя.

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

В то же время существуют отдельные исследования, направленные на поиск подхода к оптимальному разделению всей изучаемой совокупности на необходимые части. Так, в работе К. К. Доббина и Р. М. Саймона [1] анализируется влияние пропорции разделения данных на среднеквадратическую 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, исследователи подтвердили, 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 то же время следует отметить, что исследователи не ставили целью найти оптимальное соотношение из всех возможных. В рамках данной работы рассматривались лишь два варианта стратегий разделения: в качестве тренировочной выборки использовали или 1/2, или 2/3 имеющейся совокупности. В то же время отметим, что вопрос определения контрольной выборки в данной работе не рассматривался.

В альтернативном исследовании Г. Афендрас и М. Маркатоу [2] приводят теоретические и эмпирические обоснования оптимального разделения выборки данных с применением метода перекрестной проверки (*cross-validation*). Авторы, решая задачу оптимизации для определения подходящего размера выборки, приходят к выводу, что для большого класса функций такой размер обучающей выборки равен половине общего ее размера независимо от распределения данных. Таким образом, стремясь установить правила, позволяющие выбирать оптимальный размер тренировочной выборки для фиксированного набора данных размера n, авторы подтверждают мнение своих коллег-практиков о необходимости отделения лишь половины данных для обучения. В то же время в рассматриваемой статье, как и в исследовании К. К. Доббина и Р. М. Саймона, не изучается вопрос определения контрольной выборки.

Таким образом, в зависимости от применяемого критерия ученые приходят к различным выводам об оптимальной пропорции разделения выборки для обучения. Кроме того, требуется дальнейшее исследование проблемы определения оптимальной пропорции контрольной выборки.

Материалы и методы исследования

Для изучения рассматриваемого вопроса был создан программный алгоритм на языке *R*. Этот алгоритм необходим для поиска оптимального соотношения в разделении имеющейся совокупности данных на тренировочную и контрольную выборки на основе информации по автономным факторам ликвидности банковского сектора Республики Беларусь.

Алгоритм основан на следующих условиях и допущениях:

- при анализе используются ежедневные данные по подкреплению банков Республики Беларусь наличностью, инкассации наличности, налоговым доходам, поступлениям таможенных платежей и расходам Правительства Республики Беларусь;
- принципиально не применяются временные ряды, сгенерированные искусственно;
- алгоритм строит эконометрическую модель, известную как сезонная интегрированная модель авторегрессии – скользящего среднего (*seasonal autoregressive integrated moving average* (далее – SARIMA));
- для автоматизации построения модели SARIMA используется метод, предложенный Р. Хайндменом и Я. Кандекар [3], основанный на разработках С. Ван, К. Смит и Р. Хайндмена [4];
- верификация и валидация построенных моделей проходят последовательно;
- основным параметром верификации является Байесовский информационный критерий, или критерий Шварца;
- степень валидности модели определяется ее способностью к прогнозированию будущих значений временного ряда;
- в качестве параметра точности прогнозирования используется средняя абсолютная масштабированная ошибка (*mean absolute scaled error* (далее – MASE));
- алгоритм основан на принципе итерационного поиска оптимального решения;
- шаг прохождения алгоритма подобран так, чтобы для каждого временного ряда разделение совокупности данных происходило 1000 раз.

Рассмотрим причины, повлиявшие на выбор данных параметров алгоритма. Анализируемый набор временных рядов обусловлен проводимым исследованием по улучшению качества прогнозирования автономных факторов ликвидности банковской системы Республики Беларусь. Кроме того, высокая продолжительность, ежедневная частота и предварительный анализ для определения особенностей представленных данных позволили предложить к использованию в проводимом исследовании модель SARIMA.

Вместе с тем искусственно сгенерированные временные ряды не используются в построении алгоритма, поскольку их свойства строго функционально определяются при их генерации. Указанная особенность, по нашему мнению, исказит исследование, направленное на поиск отличительных черт, характерных для реальных данных.

Применяемая в алгоритме модель SARIMA является расширением для базовой интегрированной модели авторегрессии – скользящего среднего (*autoregressive integrated moving average* (далее – ARIMA)). Данное расширение включает в себя дополнительные сезонные переменные для авторегрессии, порядка интегрирования, скользящего среднего, а также вводит дополнительный параметр вида сезонности. Последний предполагает определение времени, которое относительно временного ряда занимает один цикл сезонности. Включение всех дополнительных параметров позволяет учитывать сезонный фактор при моделировании с применением ARIMA [5, с. 242].

Предложенное расширение для ARIMA выбрано в результате предварительного анализа используемых данных, позволившего обнаружить во временных рядах внутренние циклы, схожие с сезонностью.

В свою очередь, относительная простота параметризации и верификации модели ARIMA привела к использованию именно этой спецификации для исследования.

Выбор метода автоматического построения модели SARIMA, разработанного Р. Хайндменом и Я. Кандекар, связан с возможностью подбора параметров модели в соответствии со значением одного информационного критерия из перечня. Решение брать за основу значения информационных критериев при выборе конкретной модели связано в первую очередь с тем, что данные инструменты изначально разрабатывались именно в целях выбора оптимальной по соотношению точности и сложности модели [6]. Такой подход к пониманию информационных критериев позволяет применять их при обучении модели на тренировочной выборке в автоматизированном режиме, т. е. реализовать программный алгоритм поиска оптимальной модели. Решение использовать Байесовский информационный критерий, известный также как критерий Шварца (предложен Г. Шварцем в 1978 г. [7]), было принято, поскольку среди прочих вариантов, таких как классический и скорректированный информационный критерии Акаике [8], он является оптимальным. Кроме того, скорректированный информационный критерий Акаике рекомендуется применять при малом количестве наблюдений (до 100), что неуместно в рассматриваемом случае. В свою очередь, критерий Шварца считается более строгим по сравнению с классическим критерием Акаике, потому что налагает больший штраф на рост количества переменных моделей.

Выбор программного решения Р. Хайндмена и Я. Кандекар связан с тем, что их подход основывается на методе кластеризации анализируемых временных рядов по разработанной С. Ван, К. Смит и Р. Хайндменом методике. Благодаря предварительной кластеризации появляется возможность определять временные ряды с сезонными и циклическими свойствами и строить для них модели SARIMA автоматизированными программными методами.

Несмотря на то что этапы верификации и валидации являются близкими, они не могут выполняться одновременно. Кроме того, суть данных процессов различна. Верификация определяет соответствие формальным признакам, т. е. подтверждает, что построенная модель является адекватной, а полученные в результате ее использования оценки – несмещенными, состоятельными и эффективными (или максимально приближенными к этим требованиям). Таким образом, мы удостоверяемся, что построили именно задуманную модель.

На этапе валидации выясняется, выполняет ли модель свои функции и насколько хорошо она это делает. В случае применения описываемого алгоритма целью модели было прогнозирование будущих значений временного ряда с максимальной точностью. С учетом этого показателя для оценки точности прогноза была выбрана MASE. В отличие от часто применяемого на практике квадратного корня MSE, показатель MASE не зависит от масштаба данных. Благодаря этому с его помощью можно сравнивать прогностические качества моделей на основании прогнозов на разные периоды, т. е. уровень ошибки не зависит от прогностического масштаба моделей. Данный показатель был предложен в работе [9].

Итерационный поиск решения для определения оптимального разделения совокупной выборки в целях повышения точности прогноза выбран с учетом большого объема анализируемых данных. При этом вся совокупность разделяется на 1000 вариантов и проверяется последовательно. Разделение на большее количество вариантов невозможно при использовании текущих данных ввиду ограниченности сопоставимых наблюдений. Однако полагаем рациональным увеличить степень разделения совокупной выборки не меньше, чем на порядок, при наличии достаточного объема наблюдений.

Результаты и их обсуждение

В результате работы алгоритма были получены ряды MASE в зависимости от разделения совокупной выборки данных каждого ряда на тренировочную, контрольную и проверочную выборки. Для дальнейшего анализа были проигнорированы первый и последний децили диапазонов возможных разделений совокупной выборки, поскольку они с большей вероятностью будут смещены в сторону завышения и занижения ошибки прогнозирования соответственно (при этом на графиках данные диапазоны не исключены для наглядной демонстрации их специфичности). В случае заведомо малой тренировочной выборки модель получает недостаточно данных для корректного обучения. В свою очередь, при применении для обучения почти всей совокупности данных весьма вероятно получение «переобученной» модели, или же модели, которая максимально точно предскажет ближайшие периоды, но будет бесполезна для реального применения.

В результате проведенного анализа были обнаружены некоторые особенности. Изменение MASE для первого ряда в зависимости от доли, отделяемой на тренировку и верификацию модели, продемонстрировано на рис. 2.

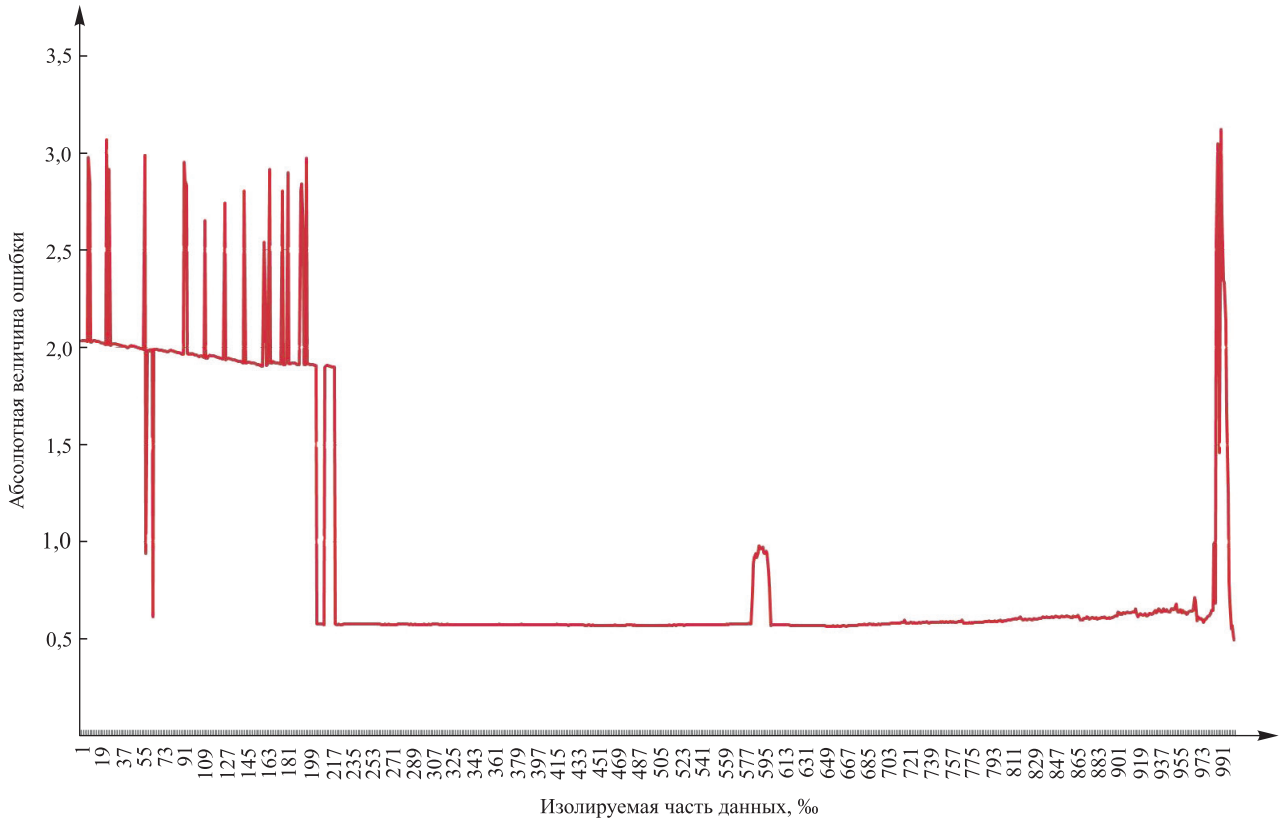

Рис. 2. Изменение ошибки прогнозирования для 1-го тестируемого ряда данных

Fig. 2. Forecasting error changes of 1st analysed series

Можно обратить внимание на то, что после отделения чуть более 25 % данных для тренировки и верификации модели дальнейшее увеличение доли данных, отведенной на тренировку, не приводит к значительному повышению точности прогноза. Если ремасштабировать и проранжировать ряд ошибок от наибольшего значения, то можно обнаружить на графике плато относительно стабильных уровней ошибок (рис. 3).

Дополнительно можно заметить, что даже в наилучшем случае разделения совокупности ошибка прогнозирования не будет значительно ниже уровня ближайшего к нему плато. Данная особенность характерна для всех исследуемых рядов, что продемонстрировано на рис. 4.

Следует полагать, что это связано с возможностью модели SARIMA к прогнозированию исследуемых временных рядов. Данную особенность определим как предельную способность применяемой модели к прогнозированию временного ряда. Под ней предлагаем понимать минимальную ошибку прогноза, которую возможно получить при идеальных условиях обучения модели. К основным причинам появления такого ограничения отнесем невозможность повышения точности прогнозирования ввиду отсутствия возможности учета дополнительных значимых факторов в применяемой для прогнозирования модели.

Для 2-го проанализированного ряда минимальные значения ошибки прогнозирования получаются в случае отделения для обучения и верификации не более 45–60 % совокупной выборки данных. При этом стоит отметить, что дальнейшее увеличение тренировочной выборки не только не уменьшало величину ошибки прогнозирования, но и приводило к ее росту (рис. 5).

Несколько другая картина наблюдается при анализе 3-го временного ряда (рис. 6). В данном случае минимальные ошибки прогнозирования наблюдаются при отделении 40–55 % совокупной выборки для обучения, однако вблизи этого диапазона присутствуют также локальные и глобальные пики уровня ошибки, в связи с чем рекомендуется отделять 65–85 % совокупной выборки. В указанном диапазоне ошибка несколько выше минимальной, но ее волатильность незначительна.

Для 4-го временного ряда характерно постепенное уменьшение ошибки прогнозирования и снижение уровня ее волатильности. Для прогнозирования данного ряда оптимальное значение ошибки достигается при отделении 70–85 % данных для тренировки и верификации модели (рис. 7).

Схожий диапазон является оптимальным и для 5-го исследуемого ряда, что может свидетельствовать о том, что они также близки и по свойствам.

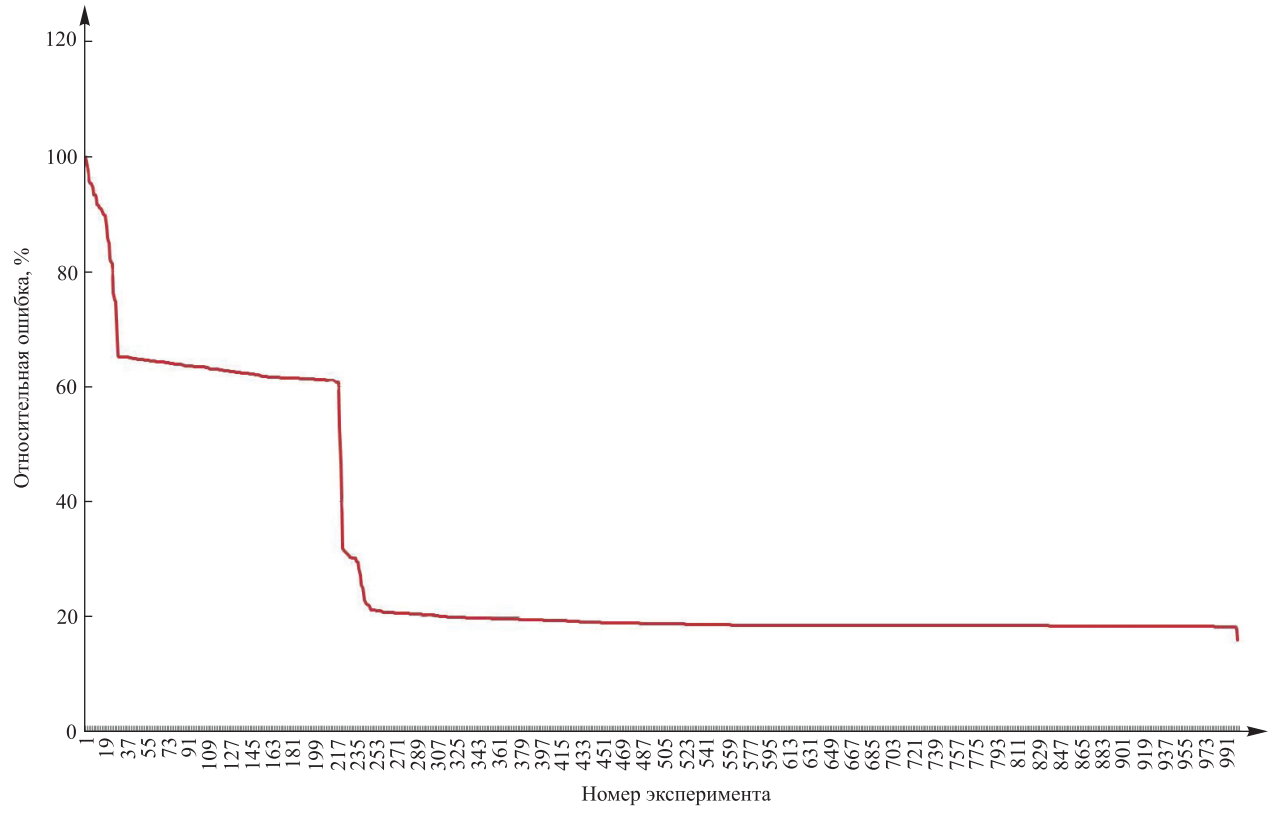

Рис. 3. Относительная ошибка 1-го тестируемого ряда данных
Fig. 3. Relative error of the 1st analysed series

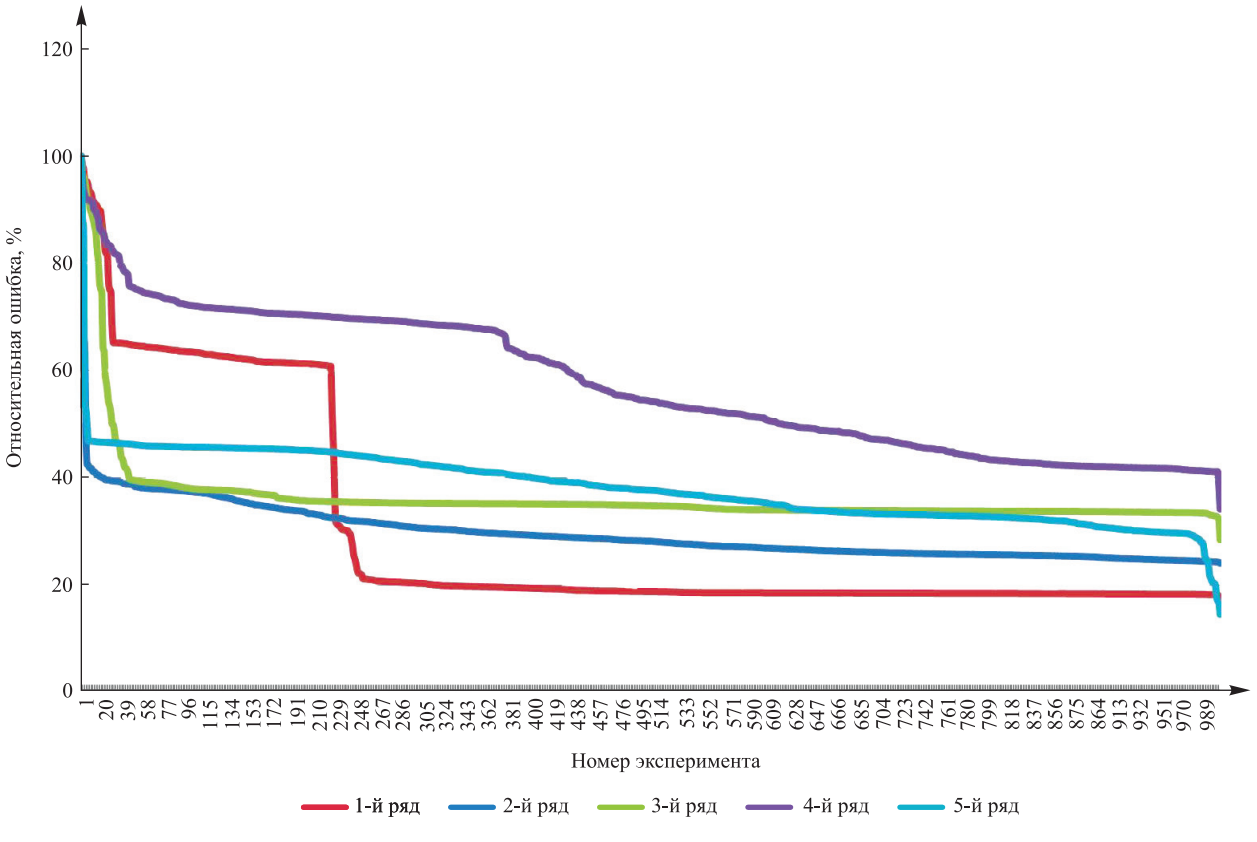

Рис. 4. Относительные ошибки по всем рядам данных
Fig. 4. Relative errors for all analysed series

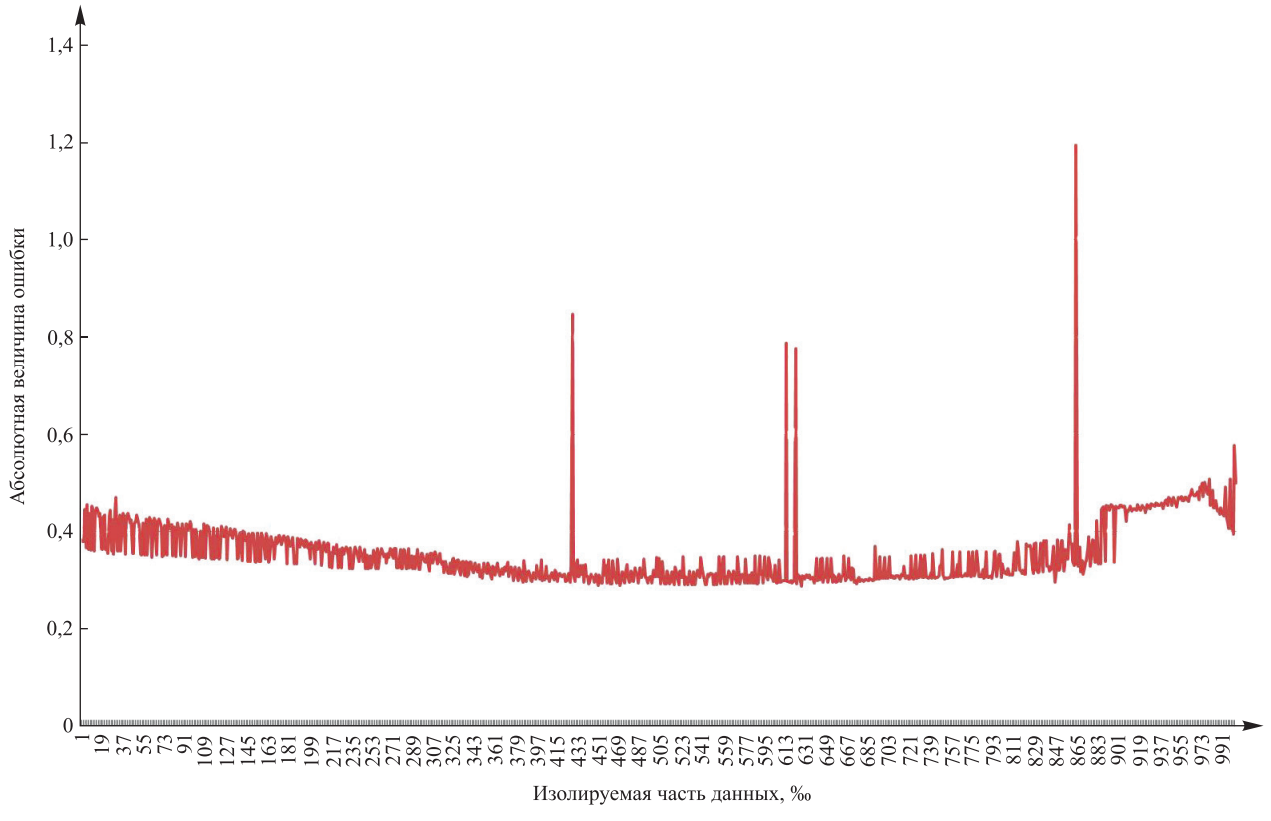

Рис. 5. Изменение ошибки прогнозирования для 2-го тестируемого ряда данных
Fig. 5. Forecasting error changes of the 2nd analysed series

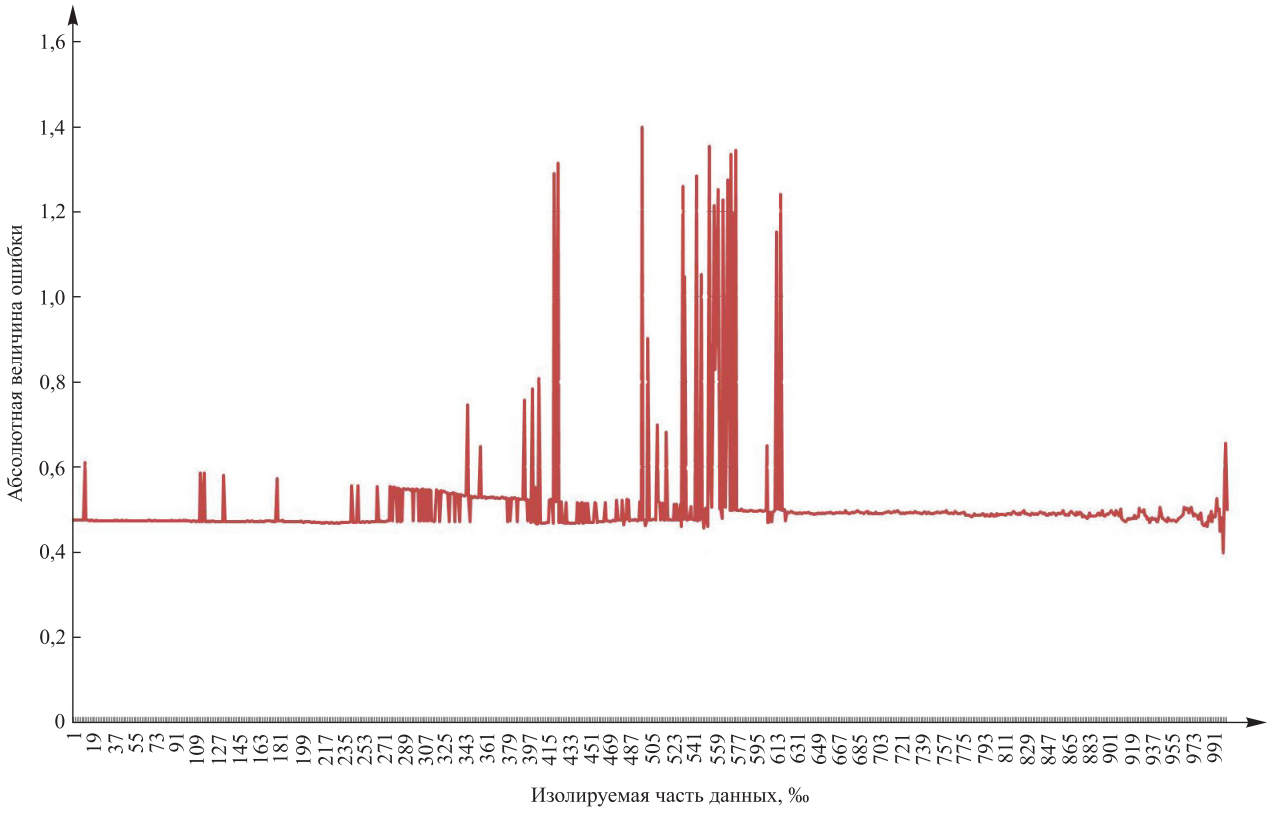

Рис. 6. Изменение ошибки прогнозирования для 3-го тестируемого ряда
Fig. 6. Forecasting error changes of the 3rd analysed series

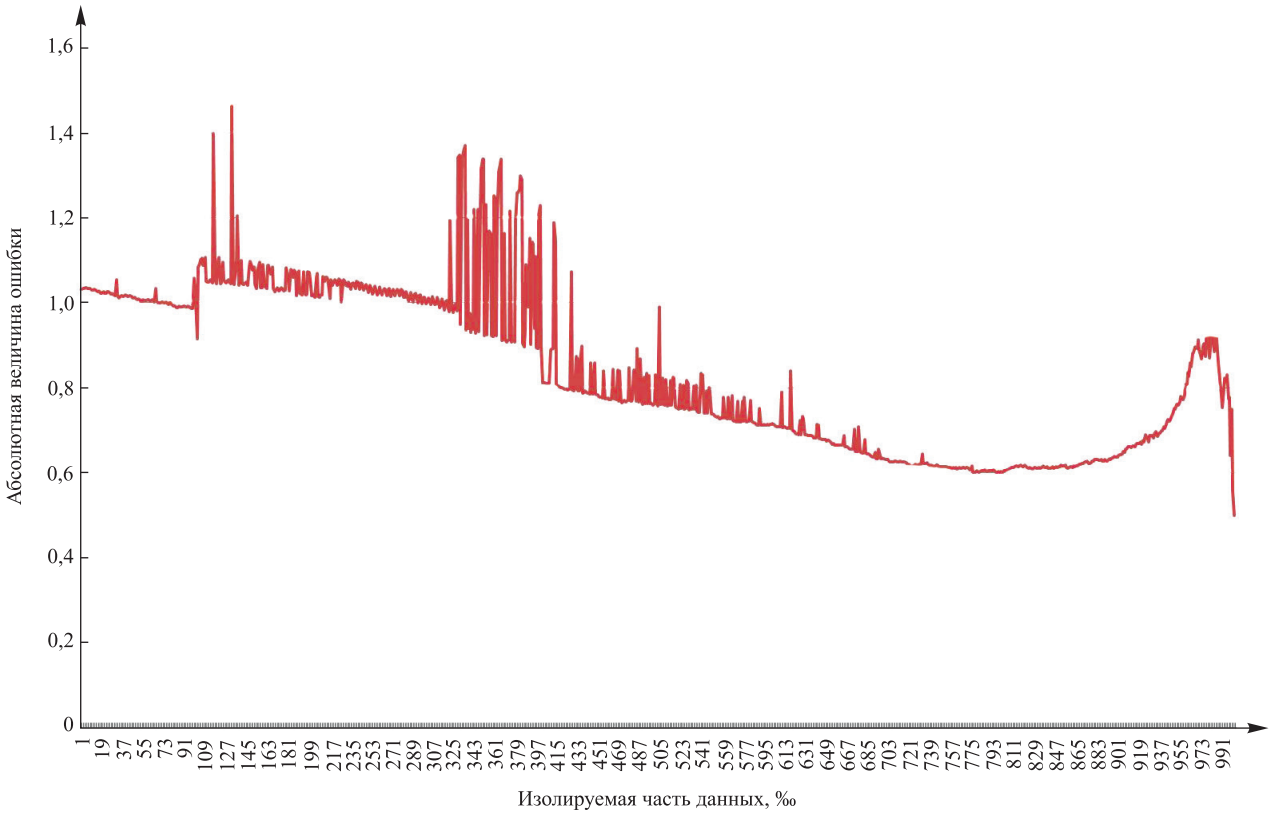

Рис. 7. Изменение ошибки прогнозирования для 4-го тестируемого ряда данных

Fig. 7. Forecasting error changes of the 4th analysed series

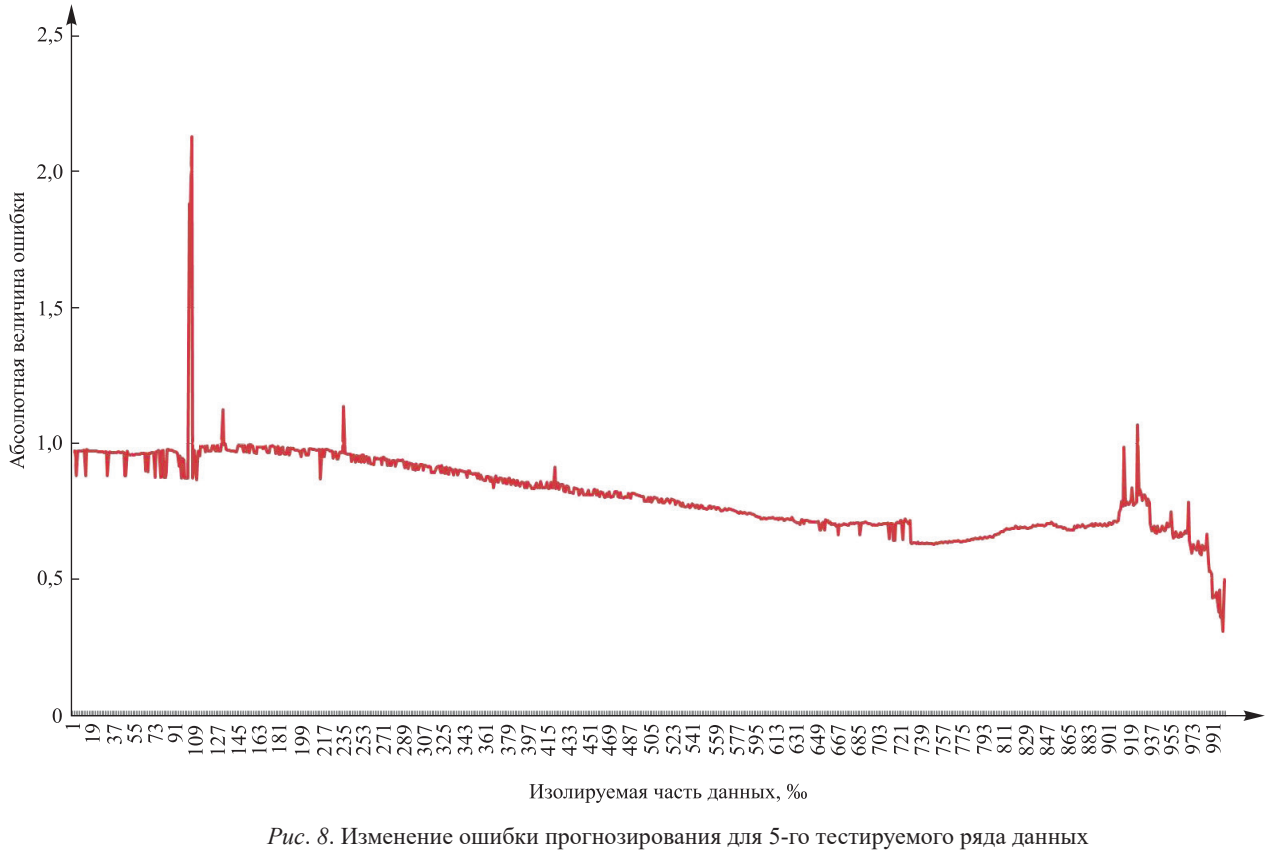

Рис. 8. Изменение ошибки прогнозирования для 5-го тестируемого ряда данных

Fig. 8. Forecasting error changes of the 5th analysed series

Заключение

Проведенное исследование позволяет сделать вывод о том, что для повышения качества прогнозов, получаемых с помощью моделей, необходимо разделять совокупную выборку данных на отдельные наборы для разных этапов построения модели. Этот вопрос не теряет актуальности в рамках увеличения объема данных, доступного для анализа, и перехода к методу машинного обучения при построении моделей. При этом следует отметить, что широкоприменяемые информационные критерии могут выступать в качестве ориентира при верификации моделей, построенных программными алгоритмами на тренировочных выборках. На этапе валидации моделей, предназначенных для прогнозирования, рационально использовать подход, основанный на поиске минимальной ошибки прогнозирования, используя при этом те показатели, что не зависят от прогнозного масштаба моделей. Примером такого показателя является MASE.

Для большинства проанализированных нами временных рядов оптимальным является отделение 70–80 % всей имеющейся совокупности данных для обучения и валидации модели, что несколько больше рекомендованного предшественниками и практикующими коллегами (оптимальным считалось отделение в пределах 1/2–2/3 всех данных для обучения). В то же время было обнаружено, что для некоторых временных рядов отделение в целях обучения и верификации не более 45–60 % совокупной выборки данных позволяет минимизировать ошибку. Другие ряды не демонстрируют значительного повышения точности прогноза при увеличении обучающих выборок данных уже после отделения 25 % совокупной выборки.

Кроме того, было обнаружено явление, определенное как предельная способность модели к прогнозированию временного ряда. Данный феномен связан с ограничениями конкретных применяемых моделей и их неспособностью учитывать все возможные факторы, оказывающие влияние на будущие значения прогнозируемого временного ряда. Указанное свойство представляет интерес для дальнейшего изучения.

Таким образом, при невозможности провести анализ конкретного временного ряда в целях определения наилучшего соотношения отделяемого объема данных для тренировки модели следует отделять 80 % имеющейся совокупности. При этом в отдельных случаях оптимальный результат достигается при иных пропорциях разделения: для обучения и верификации отбирается не более 45–60 % или не менее 25 %. Наличие данных ориентиров поможет существенно ускорить процесс поиска наилучшей пропорции разделения совокупной выборки для обучения разрабатываемых на практике моделей.

References

1. Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers [Internet; cited 2020 March 19]. Available from: https://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-4-31.

2. Afendras G, Markatou M. Optimality of training/test size and resampling effectiveness of cross-validation estimators of the generalization error. *Journal of Statistical Planning and Inference*. 2019;199:286–301.

3. Hyndman RJ, Khandakar Y. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*. 2008;27(3):1–22. DOI: 10.18637/jss.v027.i03.

4. Xiaozhe Wang, Smith K, Hyndman R. Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*. 2006;13(3):335–364. DOI: 10.1007/s10618-005-0039-x.

5. Hyndman RJ, Athanasopoulos G. *Forecasting: principles and practice*. Melbourne: OTexts; 2013. 291 p.

6. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control.* 1974;19(6):716–723. DOI: 10.1109/TAC.1974.1100705.

7. Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978;6(2):461–464. DOI: 10.1214/aos/1176344136.

8. Sugiura N. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics.* 1978;7(1):13–26. DOI: 10.1080/03610927808827599.

9. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *International Journal of Forecasting*. 2006;22(4): 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

Статья поступила в редколлегию 31.01.2021.
Received by editorial board 31.01.2021.