

УДК 004.415.2

Нуриев Марат Гумерович, доцент кафедры автоматизированных систем обработки информации и управления, Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, г. Казань

Шкиндерова Инна Николаевна, преподаватель, Межрегиональный центр компетенций – Казанский техникум информационных технологий и связи, г. Казань

Шкиндеров Максим Сергеевич, старший преподаватель кафедры автоматизированных систем обработки информации и управления, Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, г. Казань

Юнусов Азат Робертович, обучающийся бакалавриата, Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, г. Казань

Сафиуллина Камиля Маратовна, учащаяся, Муниципальное автономное общеобразовательное учреждение «Гимназия №139 – Центр образования», г. Казани

ОПТИМИЗАЦИЯ ВЫБОРА СТАТИСТИЧЕСКИХ ПАРАМЕТРОВ ДЛЯ ОЦЕНКИ ПРОИЗВОДИТЕЛЬНОСТИ АЛГОРИТМОВ

Аннотация. В статье рассматриваются ключевые методы статистического анализа, применяемые при бенчмаркинге для оценки производительности систем и алгоритмов. Авторы анализируют такие параметры, как среднее арифметическое, медиана, размах, квартили и процентиля, подчеркивая их важность для получения достоверных результатов. Особое внимание уделяется выбору размера выборки, влиянию выбросов на интерпретацию данных, а также практическим рекомендациям по комбинированию статистических методов. Статья демонстрирует, как использование пятичисловой сводки и высоких процентилей позволяет более

точно оценивать производительность в сложных распределениях. Результаты исследования могут быть полезны при оптимизации алгоритмов и проектировании высоконагруженных систем.

Annotation

The article discusses key statistical analysis methods used in benchmarking to evaluate the performance of systems and algorithms. The authors analyze such parameters as arithmetic mean, median, spread, quartiles and percentiles, emphasizing their importance for obtaining reliable results. Particular attention is paid to the choice of sample size, the impact of outliers on data interpretation, and practical guidelines for combining statistical methods. The article demonstrates how the use of five-number summaries and high percentiles allows for more accurate estimation of performance in complex distributions. The results of the study may be useful in optimizing algorithms and designing highly loaded systems.

Ключевые слова: бенчмаркинг, статистический анализ, производительность, среднее арифметическое, медиана, размах, квартили, процентиля, пятичисловая сводка, выбросы, оптимизация, распределение данных.

Keywords: benchmarking, statistical analysis, performance, arithmetic mean, median, spread, quartiles, percentiles, five-number summary, outliers, optimization, data distribution.

При проведении бенчмаркинга для оценки производительности систем или алгоритмов критически важно правильно выбрать методы статистического анализа. Это позволяет получить достоверные результаты, минимизировать ошибки и сделать обоснованные выводы. В данной статье рассматриваются ключевые статистические характеристики, их применение и практические рекомендации для эффективного анализа производительности [1,2].

Одним из первых шагов в бенчмаркинге является определение размера выборки. Минимально возможный размер выборки, обеспечивающий необходимую точность и повторяемость, позволяет сократить время

тестирования без потери качества результатов. Для начальных запусков рекомендуется использовать выборку от 15 до 30 итераций. При таком объеме ключевые статистические характеристики начинают демонстрировать достоверные значения, а большинство тестов – показывать правдоподобные результаты. Однако эти значения являются ориентировочными: в некоторых случаях достаточно 10 итераций, тогда как в других даже 40 могут оказаться недостаточными [3,4]. Для выявления значительных изменений, таких как пятикратное ускорение, иногда хватает одной итерации. Тем не менее, для сложных распределений может потребоваться сотни или тысячи запусков. Окончательные выводы следует делать только после многократных проверок.

Простейшими характеристиками распределения являются минимум и максимум, которые вместе образуют размах. Эти значения отражают наилучший и наихудший показатели производительности, позволяя оценить диапазон возможных результатов. Например, в стерильном окружении размах может быть небольшим, как (15,181 мс, 15,226 мс), что упрощает сравнение распределений без дополнительных расчетов. Однако при значительном размахе требуется более глубокий анализ с привлечением других статистических параметров [5,6].

Среднее арифметическое – один из самых распространенных методов агрегации данных. Оно вычисляется как сумма всех значений, деленная на их количество, и обозначается как \bar{x} или μ . Этот параметр часто используется для сравнения производительности алгоритмов. Например, средние значения {99, 104, 105, 108, 114} и {503, 765, 653, 741, 593} составляют 106 и 651 соответственно, что явно указывает на преимущество первого метода. Однако среднее арифметическое может вводить в заблуждение при наличии выбросов или сложных распределений [7,8]. Например, для выборки {95, 101, 304, 97, 295, 314} среднее равно 201, что не отражает типичные значения. Поэтому использование только среднего арифметического без учета формы распределения может привести к ошибочным выводам.

Медиана представляет собой альтернативный способ описания центральной тенденции выборки. Для ее вычисления значения сортируются, и выбирается средний элемент (или среднее арифметическое двух центральных элементов для четных выборок). Например, медиана для $\{1, 4, 7, 15, 20\}$ равна 7. Медиана менее чувствительна к выбросам, чем среднее арифметическое, что делает ее предпочтительной для анализа распределений с аномальными значениями. Это особенно важно в бенчмаркинге, где отдельные выбросы могут исказить общую картину [9,10].

При анализе производительности рекомендуется начинать с простых характеристик, таких как размах и среднее арифметическое, особенно если размах невелик. Однако для более сложных случаев необходимо дополнять анализ медианой и другими параметрами, такими как стандартное отклонение или квантили. Важно избегать исключительного использования среднего арифметического без понимания формы распределения. Комбинация нескольких статистических методов позволяет получить более точные и надежные результаты, что особенно критично при сравнении алгоритмов или оптимизации систем.

Таким образом, правильный выбор статистических методов и их комбинация в зависимости от особенностей данных являются залогом успешного анализа производительности в бенчмаркинге.

Для более детального анализа данных используются квантили, которые делят выборку на равные интервалы. Например, 2-квантиль, или медиана, разделяет данные пополам. Широко применяются также 4-квантили, известные как квартили. Они включают три значения: Q_1 (первый квартиль), Q_2 (медиана) и Q_3 (третий квартиль), которые делят выборку на четыре равные части. Например, для набора $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ квартили равны 3, 5 и 7. Эти значения, вместе с минимумом и максимумом, формируют пятичисловую сводку, которая предоставляет краткое, но информативное описание распределения данных. Например, для упомянутого набора пятичисловая сводка будет $\{1, 3, 5, 7, 9\}$.

ежквартильный диапазон (IQR) – это разница между верхним и нижним квартилями Q_3-Q_1 . Этот показатель полезен для оценки разброса данных, особенно в присутствии выбросов. Процентили, являющиеся 100-квантилями, позволяют еще более точно анализировать данные. Например, 25-й процентиль соответствует Q_1 , 50-й — медиане, а 75-й — Q_3 . Для большей точности иногда используют промилле (1000-квантили), хотя на практике чаще применяются процентили. Например, 99,9-й процентиль соответствует 999-му промилле.

В анализе производительности веб-приложений часто используются высокие процентили, такие как p_{95} , p_{99} или $p_{99,9}$. Эти значения помогают оценить редкие, но критически важные случаи. Например, если веб-страница отправляет 300 запросов к дополнительным ресурсам, вероятность того, что хотя бы один из них превысит 99-й процентиль, составляет около 26 %. Это демонстрирует, насколько важно учитывать высокие процентили при проектировании систем.

Пятичисловая сводка и процентили являются мощными инструментами для описания распределения данных. Если разброс значений велик, процентили обеспечивают необходимую точность. Именно поэтому они широко применяются для анализа длительности веб-запросов и других метрик производительности. Понимание этих показателей позволяет более эффективно интерпретировать данные и принимать обоснованные решения.

Список литературы

1. Гибадуллин Р.Ф. Стегостойкость и вычислительная стойкость ассоциативной стеганографии / Р.Ф. Гибадуллин, И.С. Вершинин, В.А. Райхлин // Методы моделирования VII : Труды Республиканского научного семинара «Методы моделирования», Казань, 01 сентября 2016 года – 30 2019 года / Под редакцией В.А. Райхлина. Том Вып. 7. Казань:

- Издательство "Фэн" Академии наук Республики Татарстан, 2019. С. 23-38.
2. Ямалеева Г.Н., Перухин М.Ю., Гибадуллин Р.Ф. Оптимизация исполнения SQL-запросов к базам данных под управлением MySQL // Информационные технологии и математическое моделирование (ИТММ-2017): Материалы XVI Международной конференции имени А.Ф. Терпугова, Казань, 29 сентября – 03 октября 2017 года. Том Часть 2. Казань: Издательство научно-технической литературы, 2017. С. 239-241.
 3. Райхлин В.А. Элементы содержательной теории ассоциативной стеганографии / В.А. Райхлин, И.С. Вершинин, Р.Ф. Гибадуллин // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. 2019. № 1. С. 41-47.
 4. Викторов И.В., Гибадуллин Р.Ф. Принципы построения сервиса преобразования последовательного программного кода в параллельный на базе машинного обучения // Международный форум KAZAN DIGITAL WEEK - 2023: Сборник материалов, Казань, 20–22 сентября 2023 года. Казань: ГБУ «НЦБЖД», 2023. С. 743-748.
 5. Конструктивное моделирование систем информатики / В.А. Райхлин, И.С. Вершинин, Р.Ш. Минязев, Р.Ф. Гибадуллин ; Под ред. В.А. Райхлина. Казань: Издательство "Фэн" Академии наук Республики Татарстан, 2016. 312 с.
 6. Гибадуллин Р.Ф., Вершинин И.С. Состояние и перспективы развития ассоциативной стеганографии // Математические методы в технологиях и технике. 2024. № 11. С. 107-111.
 7. Двумерно-ассоциативная защита информации в картографических системах / В.А. Райхлин, И.С. Вершинин, Р.Ф. Гибадуллин, С.В. Пыстогов // Проблемы и перспективы развития наукоемкого машиностроения : Международная научно-техническая конференция, Казань, 19–21 ноября 2013 года. Казань: Казанский государственный университет, 2013. С. 48-50.

8. Гибадуллин Р.Ф., Новиков А.А., Курочкин Г.В. и др. Параллельное управление защищенными картографическими базами данных в среде PostgreSQL // Геодезия, картография и маркшейдерия: Материалы Всероссийской научной Интернет-конференции с международным участием, Казань, 05 июня 2014 года / Сервис виртуальных конференций Рах Grid; составитель Д.Н. Синяев. Казань: Индивидуальный предприниматель Синяев Дмитрий Николаевич, 2014. С. 7-10.
9. Гибадуллин Р.Ф. Сервер и клиент систем управления картографическими базами данных / Р.Ф. Гибадуллин, С.В. Пыстогов, И.С. Вершинин // Инновационные технологии XXI века : материалы Международной научно-практической конференции, Казань, 17 апреля 2015 года / Казанский государственный технический университет им. А.Н. Туполева. Казань: Казанский государственный технический университет им. А.Н. Туполева, 2015. С. 5-8.
10. Гибадуллин Р.Ф., Гарипов Р.М., Диаров М.М. Параллельные модули импорт и экспорта защищенной картографической базы данных // Поиск эффективных решений в процессе создания и реализации научных разработок в российской авиационной и ракетно-космической промышленности: Международная научно-практическая конференция, Казань, 05–08 августа 2014 года. Том II. Казань: Издательство Казанского государственного технического университета, 2014. С. 418-421.

References

1. Gibadullin R.F. Stegostability and computational stability of associative steganography / R.F. Gibadullin, I.S. Vershinin, V.A. Raikhlin // Methods of modeling VII : Proceedings of the Republican scientific seminar "Methods of modeling", Kazan, September 01, 2016 – 30 2019 / Edited by V.A. Raikhlin. Vol. 7. Kazan: Publishing house "Fen" of the Academy of Sciences of the Republic of Tatarstan, 2019. P. 23-38.

2. Yamaleeva G.N., Perukhin M.Yu., Gibadullin R.F. Optimization of SQL query execution in MySQL databases // Information Technologies and Mathematical Modeling (ITMM-2017): Proceedings of the XVI International Conference named after A.F. Terpugov, Kazan, September 29 – October 03, 2017. Part 2. Kazan: Publishing House of Scientific and Technical Literature, 2017. P. 239-241.
3. Raikhlin V.A. Elements of the meaningful theory of associative steganography / V.A. Raikhlin, I.S. Vershinin, R.F. Gibadullin // Moscow University Bulletin. Series 15: Computational Mathematics and Cybernetics. 2019. No. 1. P. 41-47.
4. Viktorov I.V., Gibadullin R.F. Principles of building a service for converting sequential program code into parallel based on machine learning // International Forum KAZAN DIGITAL WEEK - 2023: Proceedings, Kazan, September 20–22, 2023. Kazan: NCBJD, 2023. P. 743-748.
5. Constructive modeling of informatics systems / V.A. Raikhlin, I.S. Vershinin, R.Sh. Minyazev, R.F. Gibadullin ; Ed. by V.A. Raikhlin. Kazan: Publishing house "Fen" of the Academy of Sciences of the Republic of Tatarstan, 2016. 312 p.
6. Gibadullin R.F., Vershinin I.S. State and prospects for the development of associative steganography // Mathematical Methods in Technologies and Engineering. 2024. No. 11. P. 107-111.
7. Two-dimensional associative information protection in cartographic systems / V.A. Raikhlin, I.S. Vershinin, R.F. Gibadullin, S.V. Pystogov // Problems and prospects of science-intensive engineering development : International scientific and technical conference, Kazan, November 19–21, 2013. Kazan: Kazan State University, 2013. P. 48-50.
8. Gibadullin R.F., Novikov A.A., Kurochkin G.V., et al. Parallel management of protected cartographic databases in a PostgreSQL environment // Geodesy, Cartography and Mine Surveying: Proceedings of the All-Russian Scientific Internet Conference with International Participation, Kazan, June 05, 2014 /

Pax Grid Virtual Conference Service; compiled by D.N. Sinyaev. Kazan: Individual Entrepreneur Sinyaev Dmitry Nikolaevich, 2014. P. 7-10.

9. Gibadullin R.F. Server and client of cartographic database management systems / R.F. Gibadullin, S.V. Pystogov, I.S. Vershinin // Innovative technologies of the XXI century : materials of the International scientific and practical conference, Kazan, April 17, 2015 / Kazan State Technical University named after A.N. Tupolev. Kazan: Kazan State Technical University named after A.N. Tupolev, 2015. P. 5-8.
10. Gibadullin R.F., Garipov R.M., Diarov M.M. Parallel import and export modules of a protected cartographic database // Search for effective solutions in the process of creating and implementing scientific developments in Russian aviation and rocket-space industry: International Scientific and Practical Conference, Kazan, August 05–08, 2014. Volume II. Kazan: Publishing House of Kazan State Technical University, 2014. P. 418-421.