DOI 10.46320/2077-7639-2025-2-135-48-54

Использование модели DHMM для оптимизации анализа распределения ресурсов безсерверной вычислительной платформы

Ван Яо

Исследование сосредоточено на использовании динамической скрытой марковской модели (DHMM) для оптимизации распределения ресурсов в платформе безсерверных вычислений. Безсерверные вычисления как новая парадигма облачных вычислений характеризуются динамичностью и неопределенностью потребности в ресурсах, что создает вызовы для их распределения. В данном исследовании сначала рассматриваются фон и значимость безсерверных вычислений, подчеркивая важность оптимизации распределения ресурсов для снижения операционных затрат и повышения эффективности использования ресурсов. Затем подробно описываются параметры и методы оценки DHMM модели, включая матрицу вероятностей перехода состояний A, матрицу вероятностей наблюдений B, вектор начальных вероятностей состояний π, количество состояний M и количество возможных значений наблюдений N. Анализ переходов состояний позволяет модели DHMM прогнозировать изменения потребности в ресурсах, обеспечивая поддержку принятия решений по управлению ресурсами. Исследование также включает этапы проверки модели, такие как подготовка данных, обучение модели и кросс-валидация, чтобы гарантировать точность и надежность модели. В конечном итоге, исследование показывает, что модель DHMM может эффективно улучшить производительность платформы безсерверных вычислений, повысить ее адаптивность и имеет важное теоретическое и практическое значение.

ДЛЯ ЦИТИРОВАНИЯ ГОСТ 7.1–2003

Ван Яо. Использование модели DHMM для оптимизации анализа распределения ресурсов безсерверной вычислительной платформы // Дискуссия. — 2025. — Вып. 135. — С. 48–54.

КЛЮЧЕВЫЕ СЛОВА

Безсерверные вычисления, динамическая скрытая марковская модель, оптимизация распределения ресурсов, экономия, финансовые ресурсы.

DOI 10.46320/2077-7639-2025-2-135-48-54

Using the DHMM model to optimize resource allocation analysis in a serverless computing platform

Wang Yao

This study focuses on optimizing the resource allocation of the dynamic Hidden Markov model (DHMM). As an emerging cloud computing paradigm, the dynamics and uncertainty of its resource requirements bring challenges to resource allocation. This study first introduces the background and research significance of servers less computing, emphasizing the importance of optimizing resource allocation to reduce operating costs and improve resource utilization. Then, the parameter definition and estimation method of DHMM model are elaborated, including state transition probability matrix A, observation probability matrix B, initial state probability vector π, state number M and possible number of observed values N. Through state transition analysis, the DHMM model is able to predict changes in resource demand and provide decision support for resource management. The study also involved model validation steps, including data preparation, model training, and cross-validation, to ensure the accuracy and reliability of the model. Finally, the study shows that DHMM model can effectively improve the performance of server-less computing platform and enhance its adaptability, which has important theoretical significance and practical value.

FOR CITATION APA

Wang Yao. Using the DHMM model to optimize resource allocation analysis in a serverless computing platform. *Diskussiya [Discussion]*, 135, 48–54.

KEYWORDS

Serverless computing, dynamic hidden Markov model, optimization of resource allocation, savings, financial resources.

ВВЕДЕНИЕ

Современные безсерверные вычислительные платформы предоставляют пользователям возможность выполнять программный код и сами вычисления без необходимости директивного управления инфраструктурой. Подобный подход позволяет значительно сократить затраты на поддержку серверов, упростить масштабирование процессов и повысить эффективность использования, как финансовых, так и материальных ресурсов. Динамическая природа таких систем, где рабочие нагрузки (производственная мощность) могут меняться в зависимости от времени суток, дня недели или сезонности, создает прикладные сложности для оптимального распределения ресурсов. В условиях высокой вариативности спроса, ограниченности вычислительных мощностей возникает необходимость в разработке методов анализа и прогнозирования распределения ресурсов мощностей, которые позволят минимизировать издержки, максимизировать производительность, увеличить экономическую эффективность в плане экономии финансовых ресурсов и затрат рабочего времени. Одним из перспективных подходов к решению этой задачи является использование скрытых марковских моделей (англ. Hidden Markov Models, HMM), в частности их дискретной версии – модели DHMM (Discrete Hidden Markov Model). DHMM представляет собой статистическую модель, которая позволяет анализировать последовательности наблюдений и выявлять скрытые состояния системы, влияющие на эти наблюдения – эффективный контур обратной связи с подкреплением. На безсерверных платформах подобными наблюдениями могут быть метрики использования ресурсов (например, частота вызова функций, время выполнения запросов, объем потребляемой памяти), а скрытыми ресурсными состояниями – различные режимы работы системы, такие как периоды высокой нагрузки, низкой активности или переходные состояния. Экономический аспект данной проблематики состоит в том, что неоптимальное распределение ресурсов может привести к значительным финансовым потерям: избыточное выделение ресурсов ведет к увеличению финансовых затрат на облачную инфраструктуру, тогда как их недостаток может вызвать задержки в обработке запросов и снижение качества обслуживания пользователей. То есть оптимизация распределения ресурсов становится ключевым фактором повышения экономической эффективности безсерверных платформ на принципах= Парето-оптимальности. Теоретическая значимость данного исследования состоит в изучение возможностей применения модели DHMM для анализа и прогнозирования распределения ресурсов в безсерверных вычислительных системах. Автор выдвигает теоретическую гипотезу, что использование DHMM позволит более точно идентифицировать скрытые режимы работы платформы, что, в свою очередь, обеспечит более эффективное управление ресурсами, снижение операционных затрат, позволяя достичь ресурсного оптимального состояния всей системы. В рамках этого исследования автор рассмотрел теоретические основы модели DHMM, проанализировал ее применимость для решения задачи оптимизации распределения ресурсов, а также рассмотрел экономические выгоды от внедрения такой модели.

ОСНОВНАЯ ЧАСТЬ

В эпоху цифровой трансформации безсерверные вычисления становятся ключевой технологией, позволяющей разработчикам сосредоточиться на создании кода без необходимости управления инфраструктурой. Однако динамичная и неопределенная природа распределения ресурсов в безсерверных вычислительных платформах делает традиционные методы малопригодными для таких задач. В этом исследовании предлагается использовать динамическую скрытую модель Маркова (DHMM) для прогнозирования изменений в ресурсных потребностях и оптимизации их распределения. Теоретическая значимость данного подхода заключается в том, что DHMM вводит новый способ анализа изменений в потребностях ресурсов, обогащая теоретическую базу безсерверных вычислений. Модель анализирует динамические процессы, помогая глубже понять переходы состояний и разработать более эффективные алгоритмы управления ресурсами. Практическая значимость этого подхода проявляется в возможности оптимизации распределения ресурсов, что позволяет сократить эксплуатационные расходы, повысить скорость реакции системы и улучшить пользовательский опыт. Использование DHMM помогает предсказывать колебания спроса, снижая узкие места в производительности и затраты, что делает этот подход полезным как для научного сообщества, так и для облачных провайдеров, стремящихся к гибкой и эффективной архитектуре. Модель DHMM сочетает цепи Маркова и анализ временных рядов для описания и прогнозирования изменений состояния системы. В контексте безсерверных вычислений эта модель позволяет отражать

динамику спроса на ресурсы, прогнозировать будущие потребности на основе исторических данных и преобразовывать задачу распределения ресурсов в задачу идентификации последовательностей состояний, анализируя вероятности их переходов и наблюдений. Модель DHMM обеспечивает точные прогнозы изменений в нагрузках, что делает её мощным инструментом для гибкого и экономичного управления ресурсами в безсерверной среде.

В модели DHMM параметры играют важную роль в описании динамики изменения потребности в ресурсах на безсерверной вычислительной платформе. Параметр A представляет собой матрицу вероятностей перехода состояний, где элементы aija_{ij} показывают вероятность перехода из состояния ii в состояние jj. Эта матрица описывает возможные переходы между различными уровнями потребности в ресурсах, при этом 0⩽aij⩽10 \leq a_{ij} \leq 1 и для каждого состояния ∑jaij=1\sum_{j}a_{ij} = 1. Параметр B описывает вероятность наблюдения определенного значения потребности в ресурсах kk в состоянии jj, что позволяет моделировать распределение спроса на ресурсы в зависимости от состояния. Параметр π – это вектор вероятности начального состояния, где πi\pi_i указывает вероятность того, что система начнет с состояния ii, и сумма всех πi=1\pi_i = 1. Параметр M определяет количество состояний, которые описывают уровни потребности в ресурсах, например, низкий, средний и высокий, а параметр N указывает на количество возможных значений наблюдаемой потребности в ресурсах, например, для загрузки процессора это может быть количество интервалов.

Для оценки параметров модели DHMM обычно используется метод максимального правдоподобия. Для матрицы AA это означает подсчет частоты переходов между состояниями на основе исторических данных. Оценка матрицы BB также производится через частоту наблюдения значений потребности в ресурсах в каждом состоянии. Вектор начальных состояний π\pi оценивается на основе статистики начальной потребности в ресурсах. Определение параметров MM и NN требует учета характеристик платформы и может включать методы, такие как кластерный анализ, для более точного определения состояний и интервалов значений.

Анализ переходов между состояниями является ключевым элементом модели DHMM, с экономической точки зрения, этот процесс играет важную роль в управлении затратами и повышении эффективности использования ресурсов безсерверных платформ. Так как включает в себя определение состояний потребности в ресурсах, оценку вероятностей переходов между этими состояниями на основе исторических данных и анализ факторов, оказывающих влияние на эти переходы. Такой подход позволяет выявить закономерности спроса, что дает возможность провайдерам минимизировать издержки на поддержание избыточных мощностей одновременно обеспечивая высокую доступность ресурсов для пользователей. Также строится матрица переходов, которая описывает вероятности перехода из одного состояния в другое – экономическая ценность матрицы переходов заключается в том, что она позволяет рассчитать ожидаемые затраты на обслуживание ресурсов в различных сценариях, помогая компаниям принимать обоснованные решения о распределении бюджета. Использование матрицы переходов способствует более точному прогнозированию пиковых нагрузок, что особенно важно для предотвращения экономических потерь, связанных с недостатком ресурсов. То есть подобный процессный инструмент помогает предсказать будущие изменения в потребности в ресурсах и оптимизировать распределение ресурсов, а снижение количества ошибок при прогнозировании приводит к уменьшению штрафов за невыполнение SLA (соглашений об уровне обслуживания), которые могут быть весьма значительными для облачных провайдеров. В условиях высокой конкуренции на рынке облачных услуг, экономическая эффективность становится одним из ключевых факторов успеха, именно анализ переходов между состояниями позволяет провайдерам не только реагировать на текущие изменения спроса, но и заранее планировать распределение ресурсов, что снижает операционные расходы. Например, если модель показывает высокую вероятность перехода системы из состояния низкой активности в состояние повышенной нагрузки, провайдер может заранее выделить дополнительные ресурсы, избежав необоснованных задержек (потерь времени), как и потери клиентов.

Дополнительно, экономический эффект проявляется в возможности масштабирования ресурсов в зависимости от прогнозируемых потребностей, что снижает затраты на хранение и обслуживание неиспользуемых мощностей. Проведение анализа переходов также позволяет выявить повторяющиеся паттерны спроса, такие как суточные или сезонные колебания, что дает возможность провайдерам предлагать гибкие та-

рифные планы, адаптированные под конкретные нужды клиентов. Такие тарифы могут включать скидки на использование ресурсов в периоды низкой активности, что повышает привлекательность сервиса для пользователей и увеличивает доход компании.

Кроме того, внедрение экономически обоснованных решений на основе анализа переходов между состояниями способствует улучшению общего качества обслуживания. Это, в свою очередь, положительно влияет на лояльность клиентов и снижает уровень их оттока, что напрямую связано с увеличением долгосрочной прибыли компании. Таким образом, анализ переходов между состояниями в модели DHMM представляет собой не только технический, но и мощный экономический инструмент, который помогает оптимизировать затраты, улучшить качество услуг и повысить конкурентоспособность облачных провайдеров. Например, если прогнозируется, что система перейдет из состояния низкой потребности в состояние высокой, это позволяет заранее скорректировать выделение ресурсов, чтобы избежать проблем с производительностью. Таким образом, подробный анализ переходов позволяет более точно управлять ресурсами на безсерверных вычислительных платформах и повышать эффективность их использования.

Этапы проверки модели DHMM включают подготовку данных, обучение модели, перекрестную проверку, выбор показателей эффективности, оценку и анализ результатов, а также оптимизацию модели. В процессе подготовки данных собираются исторические данные о потребности в ресурсах безсерверной вычислительной платформы, которые используются для обучения модели. На этапе обучения оцениваются параметры модели, такие как матрица вероятности перехода состояний, матрица вероятности наблюдения и вектор вероятности начального состояния. Далее проводится перекрестная проверка с разделением данных на обучающую и тестовую выборки, чтобы оценить способность модели к обобщению.

Для оценки эффективности модели применяются различные метрики, включая точность, полноту и оценку F1. Точность измеряет насколько верно модель прогнозирует состояние спроса на ресурсы, полнота – это способность модели выявлять изменения в спросе, а оценка F1 отражает сбалансированную эффективность модели, учитывая как точность, так и полноту. Алгоритм Баума-Уэлча используется для оптимизации параметров модели, выполняя итерационные вычисления до сходимости. После оптимизации модель тестируется в реальных условиях на безсерверной платформе, что позволяет дополнительно настроить модель в соответствии с рабочими нагрузками.

Показатели эффективности модели DHMM, такие как точность, отзыв, оценка F1 и использование ресурсов, помогают всесторонне оценить ее способность к прогнозированию и оптимизации распределения ресурсов, что снижает операционные расходы и повышает эффективность использования ресурсов.

ЗАКЛЮЧЕНИЕ

В данном исследовании был проведен анализ и оптимизация проблемы распределения ресурсов безсерверных вычислительных платформ с использованием динамической скрытой марковской модели (DHMM). С экономической точки зрения эффективное управление материальными и финансовыми ресурсами в таких системах напрямую влияет на эффективность управления затратами облачных провайдеров и конечных пользователей. Результаты показали, что модель DHMM эффективно прогнозирует изменения спроса на ресурсы, точно оценивая вероятность перехода состояний и вероятность наблюдения. Экономическая ценность такого прогнозирования заключается в возможности минимизировать издержки на поддержку избыточных вычислительных мощностей, одновременно обеспечивая высокий уровень обслуживания – это позволяет платформам заранее реагировать на колебания спроса, оптимизируя распределение ресурсов и снижая эксплуатационные расходы. Оптимизация расходов становится особенно важной в условиях растущей конкуренции между облачными провайдерами, где снижение затрат может стать ключевым конкурентным преимуществом. На прикладном уровне применение модели улучшает производительность, адаптируемость и гибкость платформ перед изменяющимися рабочими нагрузками. Кроме того, повышение производительности системы способствует увеличению чувства удовлетворенности клиентов, что, в свою очередь, может привести к росту дополнительных доходов за счет привлечения новых пользователей и снижения оттока существующих. Оценка параметров модели и анализ переходов состояний предлагают научную методологию для прогнозирования динамических процессов в безсерверных вычислениях, способствуя разработке соответствующих алгоритмов. Экономический эффект от внедрения таких алгоритмов может быть

значительным, так как они позволяют более точно планировать использование ресурсов и избегать простоев или перегрузок системы. Перекрестная проверка и практическое тестирование подтвердили эффективность и надежность модели DHMM. Исследование обогатило теоретическую базу в области безсерверных вычислений и предоставило облачным провайдерам новый инструмент для управления ресурсами, что имеет значительное теоретическое и практическое значение.

Список литературы

1. *Карамзаде, А. М., Сенди, А. Ш.* Сокращение задержки холодного запуска при бессерверных вычислениях с помощью облегченных виртуальных машин // Журнал сетевых и компьютерных приложений. – 2024. – № 232. – С. 104030-104030.
2. *Мауро, Ф., Реали, Д.* Применение проксимальной оптимизации политик для управления ресурсами в бессерверных периферийных вычислениях // Компьютеры. – 2024. – 13 (9). – С. 224-224.
3. *Горбани, М., Арани, М. Г.* Обзор подходов к использованию задержки при запуске в бессерверных вычислениях: перспективы, основанные на оптимизации // Computing. – 2024, (предварительная публикация). – С. 1-55.
4. *Ли Вэй, Ли Гуанхуэй, Чжао Циньлинь, Дай Чэнлун, Чэнь Сы.* Эластичный алгоритм масштабирования для безсерверных вычислений на основе вероятностного распределения // Компьютерные исследования и разработки. – С. 1-15.
5. *Гао Мин, Чэнь Гоцяан.* Алгоритм распределения нагрузки для безсерверных вычислений в условиях пограничных вычислений // Исследования в области компьютерных приложений. – 2024. № 41 (03). – С. 811-817+841.
6. *Чжан Гоушэн.* Исследование архитектуры интеграции безсерверных вычислений и микросервисов на основе Kubernetes // Журнал Китайской академии электронных наук. – 2023. – № 18 (01). – С. 48-55.
7. *Ван Синь, Чжао Кай, Цинь Бин.* Обзор применения WebAssembly для безсерверных вычислений на границе // Инженерия и приложения в области вычислительной техники. – 2023. № 59 (11). – С. 28-36.
8. *Цинь Ли, Се Дун, Ху Ян.* Исследование оценки типичных приложений научной информации с точки зрения безсерверных вычислений // Научно-технические исследования в области обороны. – 2022. – № 43 (04). – С. 6-11.
9. *Чжэн Айди, Ван Юнмэй, Сюй Цзинь, Цзо Хайин.* Применение теории моментальной помощи в продолжении ухода за пациентами с фибрилляцией предсердий и хронической сердечной недостаточностью // Китайская журнал общей медицины. – 2022. – № 20 (07). – С. 1259-1262.
10. *Ян Байай, Чжао Шань, Лю Фан.* Обзор технологий безсерверных вычислений // Инженерия и науки в области вычислительной техники. – 2022. – № 44 (04). – С. 611-619.
11. *Чжан Янь, Хэ Тао, И Синьсинь, Цао Чан, Кан Кай.* Исследование службы сетевых вычислений в безсерверных вычислениях для пограничных вычислений // Информационные и коммуникационные технологии. – 2022. – № 16 (02). – С. 40-45.
12. *Чэ Юэюань.* Бессерверные вычисления // Компьютеры и сети. – 2022. – № 48 (01). – С. 36-37.
13. *Се Дун, Ху Ян, Цинь Ли.* Применение безсерверных вычислений в области научной информации в эпоху больших данных // Журнал медицинской библиотеки и информации Китая. – 2021. – № 30 (07). –С. 39-45.
14. *Ху Цуньцун.* Современное состояние и вызовы безсерверных вычислений // Технологии безопасности сети и применения. – 2019. – № (12). – С. 84-85.
15. *Сунь Хайхон, Жэнь Цзюньчао.* Первоначальное исследование безсерверных вычислений // Электронные финансы. – 2019. – № (08). – С. 78-79.

References

1. *Karamzade, A. M., Sendi, A. S.* Reducing the cold start delay in serverless computing using lightweight virtual machines // Journal of Network and Computer Applications. – 2024. – № 232. – Pp. 104030-104030.
2. *Mauro, F., Reali, D.* Application of proximal policy optimization for resource management in serverless peripheral computing // Computers. – 2024. – 13 (9). – Pp. 224-224.
3. *Gorbani, M., Arani, M. G.* Review of approaches to using startup delay in serverless computing: prospects based on optimization // Computing. – 2024, (preliminary publication). – Pp. 1-55.
4. *Li Wei, Li Guanhui, Zhao Qinlin, Dai Chenglong, Chen Si.* Elastic scaling algorithm for serverless calculations based on probabilistic distribution // Computer Research and Development. – Pp. 1-15.
5. *Gao Ming, Chen Guoqiang*. Load balancing algorithm for serverless computing in the context of edge computing // Research in computer applications. – 2024. – № 41 (03). – Pp. 811-817+841.
6. *Zhang Gousheng*. A study of the architecture of integration of serverless computing and microservices based on Kubernetes // Journal of the Chinese Academy of Electronic Sciences. – 2023. – № 18 (01). – Pp. 48-55.
7. *Wang Xin, Zhao Kai, Qin Bing*. An overview of the use of WebAssembly for serverless computing at the border // Engineering and applications in the field of computing. – 2023. – № 59 (11). – Pp. 28-36.
8. *Qin Li, Xie Dong, Hu Yang*. A study of the evaluation of typical applications of scientific information from the point of view of serverless computing // Scientific and technical research in the field of defense. – 2022. – № 43 (04). – Pp. 6-11.
9. *Zheng Aidi, Wang Yunmei, Xu Jin, Zuo Haiying*. Application of the theory of instant care in continuing care of patients with atrial fibrillation and chronic heart failure // Chinese Journal of General Medicine. – 2022. – № 20 (07). – Pp. 1259-1262.
10. *Yang Baiai, Zhao Shan, Liu Fan*. Overview of serverless computing technologies // Engineering and computer science. – 2022. – № 44 (04). – Pp. 611-619.
11. *Zhang Yan, He Tao, and Xinxin, Cao Chang, Kang Kai*. Research of network computing service in serverless computing for edge computing // Information and Communication Technologies. – 2022. – № 16 (02). – Pp. 40-45.
12. *Che Yueyuan*. Serverless computing // Computers and networks. – 2022. – № 48 (01). – Pp. 36-37.
13. *Xie Dong, Hu Yang, Qin Li*. The use of serverless computing in the field of scientific information in the era of big data // Journal

of the Medical Library and Information of China. – 2021. – № 30 (07). – Pp. 39-45.

14. *Hu Cuntsong.* The current state and challenges of serverless computing // Network security technologies and applications. – 2019. – № (12). – Pp. 84-85.

15. *Sun Haihong, Ren Junchao.* Initial research on serverless computing // Electronic finance. – 2019. – № (08). – Pp. 78-79.

Информация об авторе

Ван Яо, независимый исследователь, студент Российского университета дружбы народов (г. Москва, Российская Федерация).

Information about the author

Wang Yao, independent researcher, student at the Peoples' Friendship University of Russia (Moscow, Russian Federation).