Прогнозирование нагрузки в Центрах обработки данных с использованием модели NARX

Я.В. Метелкин[1], Ю.В. Хицкова[2], К.А. Маковий[1]

[1]*Воронежский государственный технический университет, 20-летия Октября 84, Воронеж, Россия, 394006*
[2]*Воронежский государственный университет, Университетская площадь 1, Воронеж, Россия, 394018*

Аннотация
В статье представлена проблема оптимизации использования аппаратных ресурсов центра обработки данных. Предложен подход к прогнозированию нагрузки на сервер с использованием нейронных сетей и его реализация в системе компьютерной математики Matlab. В исследовании использовался визуальный анализ графиков исходных данных и корреляционный анализ.

Ключевые слова
Виртуализация, NARX, нейронные сети, оптимизация.

1. Введение

Управление аппаратными ресурсами в центрах обработки данных имеет важное значение для обеспечения эффективной работы, а также минимизации затрат на модернизацию и техническое обслуживание.

Одним из способов предоставления облачных сервисов является технология DAAS (Desktop as a Service), основанная на инфраструктуре виртуальных рабочих столов (VDI). В исследованиях оптимизации использования технологии VDI можно выделить два аспекта: минимизацию финансовых затрат и повышение экологичности ИТ-инфраструктуры (концепция зеленых вычислений) [1].

Существуют два подхода к решению данной проблемы: статический [2] и динамический [3]. В настоящее время для решения оптимизационных задач успешно используются биоинспирированные алгоритмы - генетические алгоритмы, искусственные иммунные системы, нейронные сети [4].

2. Использование модели NARX для прогнозирования нагрузки в центрах обработки данных

Причиной выбора НС является возможность точного прогнозирования нелинейных временных рядов. Процент загрузки сервера на каждом последующем шаге зависит от количества аппаратных ресурсов, используемых на предыдущем, поэтому мы предлагаем использовать сеть NARX.

Одним из главных преимуществ таких структур является то, что в качестве входных параметров они могут принимать динамические входные данные, представленные наборами временных рядов, а также различные внешние параметры. Управляющее уравнение для этого типа нейросетевой модели:

$$y'(t) = f(y(t-1), \dots, y(t-dy), u(t-1), \dots, u(t-du)), \quad (1)$$

где $y'(t)$ - прогнозируемое значение $y(t)$, dy и du представляют собой входную и выходную временную задержку соответственно. Модель NARX обеспечивает лучший прогноз, чем другие модели NN, поскольку она использует дополнительную информацию, содержащуюся в предыдущих значениях $u(t)$.

В рамках данного исследования мы использовали набор данных, содержащий трассировку нагрузки GWA-T-12 Bitbrains [5]. Мы решили выделить несколько классов в соответствии с процентом нагрузки с шагом 5: например, 0-5% - 0 класс, 5-10% - 1 класс и т. д., а целью прогнозирования является определение класса нагрузки во времени. Рабочая нагрузка циклична в контексте рабочей недели, поэтому в качестве внешнего параметра мы использовали номер дня недели. Этот параметр принимает значения от 1 до 5 по номеру дня недели.

В качестве алгоритма обучения был выбран метод Левенберга-Марквардта. Этот алгоритм обратного распространения был использован для минимизации ошибки, а также веса модели NARX.

3. Заключение

В ходе обучения удалось достичь MSE 1.08554 e-4. На рисунке 1 показано соотношение реальных данных загрузки и прогнозируемых. Как видно, есть проблема, которую необходимо решить в будущем, связанная с прогнозированием пиковых значений.

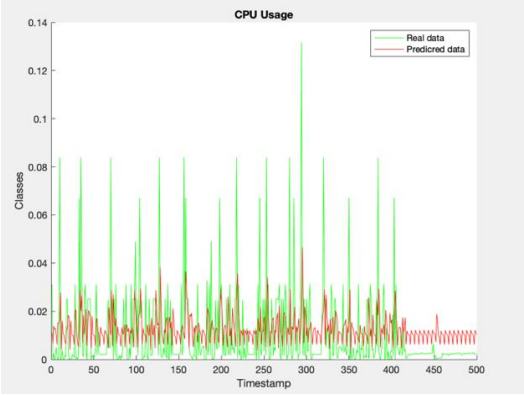

Рисунок 1: NARX прогнозируемые и реальные данные

Проведенное исследование показывает, что нейросетевой подход может быть успешно использован для прогнозирования нагрузки на сервер в долгосрочной перспективе, что позволит в комплексе с различными алгоритмами распределения ресурсов оптимизировать работу ЦОДа.

4. Литература

[1] Mi, H. Online self-reconfiguration with performance guarantee for energy-efficient large-scale cloud computing data centers / H. Mi, H. Wang, G. Yin, Y. Zhou, D. Shi, L. Yuan // Proceedings of the IEEE International Conference on Services Computing. – 2010. – P. 514-521.

[2] Makoviy, K. Server hardware resources optimization for virtual desktop infrastructure implementation / K. Makoviy, D. Proskurin, Yu. Khitskova, Ya. Metelkin // CEUR Workshop Proceedings. – 2017. – Vol. 1904. – P. 178-183. DOI: 10.18287/1613-0073-2017-1904-178-183.

[3] Wolke, A. More than bin packing: On dynamic resource allocation strategies in cloud computing / A. Wolke, B. Tsend-Ayush, C. Pfeiffer, M. Bichler // Information Systems. – 2015. – Vol. 52. – P. 83-95.

[4] Astakhova, I.F. Model and algorithm of an artificial immune system for the recognition of single symbols / I.F Astakhova, S.A. Ushakov, Ju.V. Hitskova // Advances in Computer Science Proceedings of the 6th European Conference of Computer Science. – 2015. – P. 127-131.

[5] GWA-T-12 Bitbrains. Delft University of Technology [Electronic resource]. – Access mode: http://gwa.ewi.tudelft.nl/datasets/gwa-t-12-bitbrains (12.12.2019).