ICWR2024
April 24-25, 2024, Tehran, Iran
10th International Conference on Web Research

A Consensus-based Auto-scaling Approach For Serverless Environments

Mobina Kashaniyan, Mehrdad Ashtiani, Amirhossein Ghassemi

Iran University of Science and Technology, School of Computer Engineering
mobina_kashanian@comp.iust.ac.ir , m_ashtiani@iust.ac.ir , a_ghassemi@comp.iust.ac.ir

Abstract

Efficient management of computing resources has always been a significant concern for users. An automatic scaling system can help in managing hardware resources by adapting to the system's performance history. It can increase or decrease resources automatically, without human intervention, based on predefined criteria. This ensures smooth program execution without any disruption caused by changes in the operating environment. This study focuses on serverless environments, which rely on functions. We model these functions using graph theory, analyze their dependencies, and identify the most critical bottlenecks in the graph. We then use two approaches, supervised and unsupervised, to predict the scalability of bottleneck resources. To be more sure of the scaling decision, the consensus mechanism compares the predictions of the models, and the best model's result is considered the final scaling decision, which creates consistency between the results obtained from the methods. Results show that supervised approaches perform better than unsupervised approaches in the automatic scaling problem. The models implemented in this research can determine the scaling result with 98% accuracy, which is a 2.5% improvement compared to previous works.

Keywords: Autoscaling, Cloud Environment, Machine Learning, Workload Prediction, Ensemble Learning.

ارائه یک رویکرد مقیاس‌پذیری خودکار منابع در محیط‌های بدون سرویس‌دهنده

مبینا کاشانیان، مهرداد آشتیانی، امیرحسین قاسمی

دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر
Mobina_kashanian@comp.iust.ac.ir , m_ashtiani@iust.ac.ir , a_ghassemi@comp.iust.ac.ir

چکیده

مدیریت صحیح منابع همواره یکی از دغدغه‌های کاربران هنگام استفاده از محیط‌های محاسباتی بوده‌است. یک سامانه مقیاس‌پذیر خودکار می‌تواند با توجه به تاریخچه‌ی عملکرد سامانه، منابع سخت‌افزاری را مدیریت کند و در زمان مناسب برای جلوگیری از اختلال درروند اجرای برنامه با وفق دادن سامانه با محیط عملیاتی جدید منابع را به‌صورت خودکار، بدون دخالت انسان و بر اساس معیارهای تنظیم‌شده افزایش یا کاهش دهد. در این پژوهش، با درنظرگرفتن ویژگی منحصربه‌فرد محیط‌های بدون سرویس‌دهنده که فعالیتی براساس توابع دارند، توابع را با نظریه گراف مدل‌سازی کرده و با تحلیل وابستگی‌های آن‌ها و پیداکردن مهم‌ترین گلوگاه در گراف، دو رویکرد با نظارت و بدون نظارت را برای پیش‌بینی مقیاس‌پذیری منابع گلوگاه در نظر می‌گیریم و برای اطمینان بیشتر به تصمیم مقیاس‌پذیری، پیش‌بینی مدل‌ها توسط سازوکار اجماع‌نظر مقایسه می‌گردد و نتیجه‌ی بهترین مدل به‌عنوان تصمیم نهایی مقیاس‌پذیری درنظرگرفته می‌شود که به‌نوعی میان نتایج حاصل‌شده از روش‌ها سازگاری ایجاد کند. نتایج نشان می‌دهند که رویکردهای با نظارت در مقایسه با رویکردهای بدون نظارت در مسئله مقیاس‌پذیری خودکار بهتر عمل می‌کنند و مدل‌های پیاده‌سازی شده در این پژوهش، می‌توانند تا دقت ۹۸٪ نتیجه‌ی مقیاس‌پذیری را تعیین کنند که نسبت به کارهای پیشین انجام‌شده بهبود ۲.۵ درصدی داشته است.

کلمات کلیدی

مقیاس‌پذیری خودکار، محیط بدون سرویس‌دهنده، یادگیری ماشین، رویکرد بدون نظارت، رویکرد با نظارت

۱- مقدمه

در فضای محاسبات بدون سرویس‌دهنده سامانه‌هایی وجود دارند که می‌توانند پویا باشند و مقیاس و اندازه خود را در هرزمانی تغییر دهند. این سرویس‌ها نیاز دارند که آزادی عملکرد داشته و منابع خود را در هرلحظه ارتقا دهد تا سامانه بتواند بدون اختلال کار کند، با توجه به این امر امروزه نیاز به مدیریت منابع به‌صورت خودکار در این سامانه‌ها بسیار مشهود است[۱]. این در حالی است که سایر محیط‌های محاسباتی مانند خوشه‌ها و گریدها این امکان را به‌سادگی به کاربران نمی‌دهند و ارتقا منابع در این سامانه‌ها نیازمند روش‌های پیشرفته‌تری نسبت به فضاهای ابری و بدون سرویس‌دهنده است[۲]. یکی از روش‌های مقیاس‌پذیری خودکار بر اساس کنش و واکنش تصمیم به افزایش یا کاهش منابع می‌کند، این روش می‌تواند بر کارایی، عملکرد و هزینه سامانه‌های نرم‌افزاری از طریق تخصیص منابع اضافی اثر بگذارد و حتی می‌تواند منجر به افزایش هزینه‌ها شود، یا برعکس در اثر تخصیص منابع کمتر منجر به کاهش عملکرد سامانه شود که بسیار آسیب زننده‌است. روش دیگر استفاده از راه‌حل‌های مقیاس‌پذیر خودکار است که این روش راهکاری هوشمند برای تغییر پویا مقیاس و تأمین نیازهای منابع است، این روش باید بتولند با یادگیری شرایط محیطی، نیازهای منابع آینده را پیش‌بینی کند[۳]. طرح تخصیص منابع باید مقیاس‌پذیر، قابل انطباق و قابل‌اطمینان در برابر تغییرات در حجم کار باشد که نیازهای منابع آینده را به‌درستی پیش‌بینی کند و اقدامات مقیاس‌پذیری لازم را انجام دهد نگرانی‌های روش قبل را برطرف کند. اما اثربخشی چنین راه‌حل‌هایی به‌شدت به نوع مدل پیش‌بینی و همچنین کیفیت و کمیت داده‌های آموزشی بستگی دارد و هنگامی‌که منابع موردنیاز به‌طور مناسب تخصیص داده شوند، عملکرد و کارایی سامانه به دلیل کمبود منابع کاهش نمی‌یابد. همچنین، سامانه‌های نرم‌افزاری بزرگ نیاز به ارائه درجه بالایی از اطمینان در کیفیت خدمات مانند زمان پاسخ‌گویی، توان عملیاتی و در

دسترس بودن خدمات دارند، بدون چنین اطمینانی، ارائه‌دهندگان خدمات ممکن است کاربران خود را از دست بدهند و با کاهش درآمد مواجه شوند[۴]. در این پژوهش، هدف ارائه یک سازوکار مقیاس‌پذیری خودکار در محیط‌های بدون سرویس‌دهنده است به‌نحوی‌که با کمک یادگیری ماشین بتوان مدلی ارائه داد که با ایجاد اجماع نظر میان نتایج، عملکرد پیش‌بینی تخصیص منابع را بهبود ببخشد. به‌طور خلاصه می‌توان به نوآوری‌های اصلی این کار به شرح زیر اشاره کرد:

۱. در نظر گرفتن ویژگی اصلی یک محیط بدون سرویس‌دهنده و مدل‌سازی فراخوانی توابع با استفاده از تئوری گراف.
۲. یافتن مهم‌ترین گره در گراف برای نظارت بر منابع آن‌ها.
۳. استفاده از سه مدل یادگیری ماشین برای پیش‌بینی نیازهای منابع آینده گره‌ی مهم.
۴. استفاده از الگوریتم اجماع برای یافتن بهترین مدل از بین مقادیر پیش‌بینی‌شده برای اطمینان در تصمیم‌گیری.

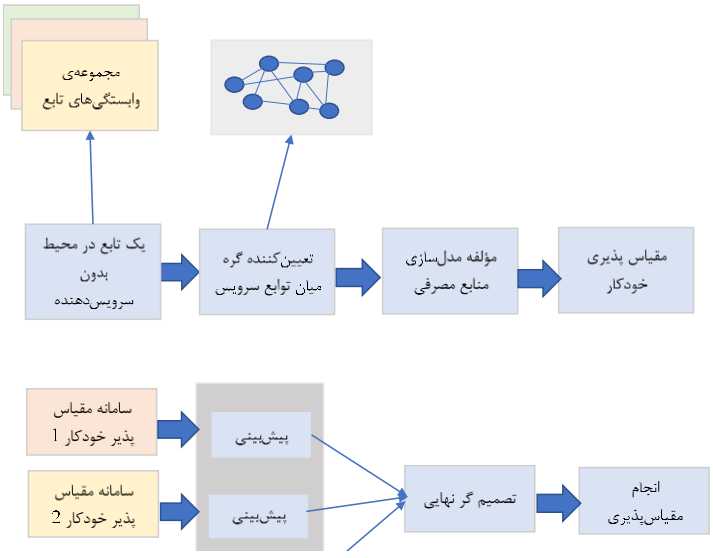
ساختار این مقاله به شرح زیر است: بخش ۲ پیشینه تحقیق را موردبحث قرار می‌دهد. رویکرد پیشنهادی در بخش ۳ ارائه‌شده است و در بخش ۴ به‌صورت جامع مراحل الگوریتم ارائه‌شده است. ارزیابی روش‌های پیشنهادی در بخش ۵ ارائه و موردبحث قرارگرفته است. درنهایت، بخش ۶ مقاله را با ارائه نتایج به پایان می‌رساند.

۲- پیشینه تحقیق

این بخش به بررسی پژوهش‌های مرتبط و روش‌های مقیاس‌پذیری خودکار می‌پردازد. روش‌های مقیاس‌پذیری خودکار اهدافی مانند بهبود دقت پیش‌بینی، انطباق باSLA یا هزینه منابع را مورد هدف قرار می‌دهند که بر اساس روش آن پژوهش‌ها طبقه‌بندی می‌شوند. قوانین مبتنی بر آستانه به‌طور گسترده برای بهینه‌سازی استفاده از منابع استفاده می‌شود و اجرای موفقیت‌آمیز این قوانین مستلزم توجه دقیق به جزئیات و درک جامع مسئله است. ناتوانی در تنظیم دقیق آستانه‌ها می‌تواند منجر به تخصیص منابع غیربهینه شود و باعث کاهش عملکرد یا ایجاد هزینه‌های غیرضروری شود. رومی و همکاران نشان می‌دهند که آستانه‌ها باید به‌دقت تنظیم شوند تا از نوسانات در سامانه جلوگیری شود[5]. یادگیری تقویتی بدون هیچ دانش قبلی با توجه به حجم کاری ورودی قادر به تعیین بهترین اقدام مقیاس‌پذیری برای هر برنامه هستند. سیستم الستیک داکر برای مقیاس‌پذیری عمودی از یادگیری تقویتی برای بهینه‌سازی منابع استفاده می‌کند[۶]. تئوری صف اغلب برای مدل‌سازی برنامه‌های کاربردی اینترنتی استفاده می‌شود و برای تخمین معیارهای کارایی مانند طول صف یا میانگین زمان انتظار مفید است. نظریه کنترل روشی برای مدیریت و مقیاس‌پذیری خودکار سامانه‌ها، نگه‌داشتن بار پردازنده و سایر متغیرهای کنترل‌شده در سطوح دلخواه از طریق تنظیمات ورودی و محاسبات ریاضی است. تجزیه‌وتحلیل سری زمانی روش دیگری است که شامل بررسی منظم منابع در طول زمان است و برای تشخیص الگو و پیش‌بینی آینده استفاده می‌شود. مارتینز و همکاران این ایده را مطرح کرده است که سامانه، الگوهای بارکاری دوره‌ای یا فصلی که تکرار می‌شود را یاد بگیرد و در پیش‌بینی‌ها بکار گیرد. [۷]. آناستاسیوس و همکاران با استفاده از چندین روش یادگیری تقویتی، به مدیریت خودکار بارهای کاری پویا با تضمین کیفیت خدمات (QoS) پرداختند و در نهایت روشی را انتخاب کردند که در استفاده از منابع بهینه عمل کرد[۸]. آگاروال و همکاران یک عامل یادگیری تقویتی بازگشتی را برای مقیاس خودکار تابع بررسی می‌کنند و آن را با الگوریتم (PPO)[1] مقایسه می‌کنند و یک شبکه LSTM با پیشرفته ترین الگوریتم PPO ارائه کردند و نشان دادند که روابط بازگشتی برای مقیاس‌پذیری توابع بهتر عمل می‌کند[۹].

۳- شرح الگوریتم پیشنهادی

پس از شناخت دقیق محیط بدون سرویس‌دهنده به یکی از مهم‌ترین ویژگی‌های این محیط که بر اساس توابع فعالیت دارند می‌رسیم. در دنیای واقعی، توابع به یکدیگر وابستگی دارند و این وابستگی‌ها باعث کاهش عملکرد سامانه و افزایش پیچیدگی می‌شود و هرچقدر که این وابستگی‌ها بیشتر شوند یک جهنم وابستگی را ایجاد می‌کنند. در این پژوهش، توابع را با استفاده از نظریه گراف مدل‌سازی می‌کنیم و پس‌ازآن مهم‌ترین گره یا گلوگاه گراف را پیدا می‌کنیم و منابع آن گره را تحت نظر می‌گیریم و سپس با استفاده از تعیین آستانه برای آن منابع مقیاس کاهشی، افزایشی یا بدون تغییر را در نظر می‌گیریم و بعدازآن پیش‌بینی مقیاس‌پذیری خودکار و منابع آینده را با سه مدل هوش مصنوعی پیچشی، پرسپترون، حافظه طولانی کوتاه‌مدت روی منابع آن گره انجام می‌دهیم و درنهایت میان نتایج به یک اجماع نظر می‌رسیم که مصرف آینده گره چه خواهد شد و مقیاس‌پذیری آنچه می‌شود این کار را انجام می‌دهیم تا مهم‌ترین گره گراف که بیشترین وابستگی را دارد به کمبود منابع دچار نشود و یا در صورت هدر رفت منابع آن را کاهش دهیم تا بقیه گره‌ها به کمبود منابع دچار نشوند. روش ارائه‌شده از مؤلفه تعیین‌کننده گره میان توابع سرویس، مؤلفه مدل‌سازی منابع مصرفی و مؤلفه تصمیم‌گیرنده مقیاس‌پذیری تشکیل‌شده است. شکل (۱) معماری کلی ایده پژوهش را نشان می‌دهد.



شکل ۱ : معماری کلی ایده پژوهش.

۴- مراحل روش پیشنهادی

۴-۱- مؤلفه تعیین‌کننده گره میان توابع سرویس

محیط بدون سرویس‌دهنده محیطی است که متکی به توابع است و در مباحث نرم‌افزاری توابع همیشه پراستفاده‌ترین عضو برنامه‌نویسی است. وظیفه این مؤلفه، تعیین گره مهم میان توابع در حال اجرا در محیط بدون سرویس‌دهنده است. هنگامی‌که کاربر درخواستی را برای اجرا بر یکی از توابع مدنظر ارسال می‌کند، روابط میان توابع در محیط‌های بدون سرویس‌دهنده موردبررسی قرار می‌گیرد و ساختمان داده گرافی متناظر با روابط توابع در نظر گرفته می‌شود. می‌خواهیم در گراف وابستگی‌ها تابعی که در معرض استفاده بیشتر است را پیدا بکنیم و هرگاه منابع این تابع را برحسب نیاز تأمین کنیم می‌توانیم اجرای پیوسته و دائم و بدون هدر رفت منابع داشته باشیم.

۴-۱-۱- درجه مرکزیت Degree Centerality

درجه مرکزیت در گراف‌ها یک معیار است که میزان مهم بودن یک گره در گراف را اندازه‌گیری می‌کند. این معیار میزان تعداد اتصال‌هایی که به یک گره متصل هستند را اندازه‌گیری می‌کند و گره‌هایی که درجه مرکزیت بالاتری دارند، به‌عنوان گلوگاه در گراف تلقی می‌شوند.

$$D_i = \frac{1}{n}\sum_{1=0}^{n} X_i \qquad (۱)$$

در (۱) X_i نشان‌دهندهمیزان اتصال هر گره است.

۴-۱-۲- بردار مرکزیت ویژه

بردار مرکزیت ویژه محاسبه می‌کند که چقدر یک گره توسط گره‌های دیگر مهم می‌شود و گره‌هایی که ارتباطات بسیاری با گره موردنظر دارند، می‌توانند به‌عنوان مهم‌ترین گره‌ها در شبکه شناخته شوند و با ماتریس مجاورت گراف و قوانین ضرب ماتریسی و دترمینان ماتریس محاسبه می‌شود.

$$C_v = (\frac{1}{\lambda})\sum(A(u,v) * C(u)) \qquad (۲)$$

$$A * C_v = \lambda * C_v \qquad (۳)$$

در (۲) $\mathbf{C_v}$ بردار مرکزیت ویژه $\mathbf{v}$ را نشان می‌دهد. $\boldsymbol{\lambda}$ بزرگ‌ترین مقدار ویژه ماتریس مجاورت گراف، $\mathbf{A(u,v)}$ عنصر ماتریس مجاورت میان دو گره و $\mathbf{C(u)}$ مقدار بردار مرکزیت ویژه گره $\mathbf{u}$ است. در (۳) $\mathbf{A}$ ماتریس مجاورت گراف و $\boldsymbol{\lambda}$ مقدار ویژه است.

۴-۲- مؤلفه مدل‌سازی منابع مصرفی

ایده اصلی این مؤلفه این است که با داشتن میزان حافظه و پردازنده می‌توان مقیاس‌پذیری را بر اساس قوانین مبتنی بر آستانه تعیین کرد که آستانه‌ها بر اساس شناخت صحیح از سامانه تعیین می‌شوند و خبرگان سامانه می‌توانند در صورت نیاز آستانه‌ها را اضافه یا کم کنند. در این مؤلفه، ابتدا برای هر پارامتر تعیین‌کننده در مسئله، مقادیر آستانه تعریف می‌شوند و سپس بر اساس مقایسه‌ی بین پارامترها، تصمیم‌گیری مربوط به مقیاس‌پذیری انجام می‌شود.

۴-۳- مؤلفه تصمیم‌گیرنده مقیاس‌پذیری

هدف و تمرکز اصلی این پژوهش رسیدن به مقیاس‌پذیری است، برای مؤلفه مقیاس‌پذیری دو روش را در نظر می‌گیریم روش با نظارت و روش بدون نظارت. درروش بدون نظارت مدل یادگیری ماشین از روی شباهت داده‌ها به یک مقیاس می‌رسد و همه‌ی نمونه‌ها را بر اساس شباهت گروه‌بندی می‌کند اما درروش یادگیری با نظارت ما یک محرک یادگیرنده داریم که در مؤلفه مدل‌سازی منابع مصرفی بر اساس آستانه، آن محرک را ایجاد کردیم.

۴-۳-۱- روش یادگیری بدون نظار

درروش بدون نظارت، روش‌های تجزیه‌وتحلیل مؤلفه‌های اصلی، خوشه‌بندی، نقشه‌های خودسازمان‌دهی در نظر گرفته‌شده‌اند. تجزیه‌وتحلیل مؤلفه‌های اصلی[۲] یک روش کاهش ابعاد بدون نظارت است که در تجزیه‌وتحلیل داده‌ها و استخراج ویژگی‌های مهم کمک می‌کند. خوشه‌بندی[۳] یک الگوریتم یادگیری ماشینی است که در آن داده‌ها بر اساس شباهت‌ها در گروه‌های متمایز قرار می‌گیرند. نقشه‌های خودسازمان‌دهی[۴] یک روش کلاسیک برای خوشه‌بندی داده‌ها و تجزیه‌وتحلیل الگوها استفاده می‌شود که مرکز خوشه به‌عنوان نماینده آن گروه عمل می‌کند.

۴-۳-۲- روش یادگیری با نظار

درروش با نظارت، مدل‌های یادگیری ماشین شبکه پرسپترون چندلایه، شبکه پیچشی، شبکه حافظه بلندمدت کوتاه‌مدت در نظر گرفته‌شده‌اند و دلیل استفاده از این سه شبکه، فراوانی بیشتر و عملکرد بهتری است که در مسئله مقیاس‌پذیری دارند. شبکه عصبی پرسپترون چندلایه[۵] نوعی شبکه عصبی است که از چندین لایه از به‌هم‌پیوسته تشکیل‌شده‌است. حافظه طولانی کوتاه‌مدت[۶] یک نوع از شبکه‌های عصبی بازگشتی[۷] است که برای پردازش داده‌های دنباله‌ای و حفظ اطلاعات بلندمدت استفاده می‌شود. شبکه‌های عصبی پیچشی[۸] برای داده‌های ترکیبی استفاده می‌شوند. درروش یادگیری با نظارت تمامی پارامترهای مؤلفه تصمیم‌گیرنده مقیاس‌پذیری به‌عنوان یک معیار تعیین‌کننده محسوب شدند و با توجه به میزان تخطی از آستانه، مقدار مؤلفه مقیاس‌پذیری تعیین شد و شامل سه حالت انجام مقیاس‌پذیری، عدم انجام مقیاس‌پذیری و کاهش مقیاس‌پذیری است.

۴-۴- مؤلفه اجماع نظر میان پیش‌بینی‌ها

اگر چندین شبکه عصبی مختلف با ساختارها، پارامترها و بهینه سازهای مختلف را استفاده کنیم، می‌توانیم با روش‌های اجماع نظر به بهترین خروجی برسیم. میانگین‌گیری مدل بیزی یک روش اجماع است که پیش‌بینی‌های چند مدل را ترکیب می‌کند و همزمان عدم قطعیت مرتبط با هر مدل را در نظر می‌گیرد. شکل (۲) به کد این اجماع را نشان می‌دهد.

۵- ارزیابی روش پیشنهادی

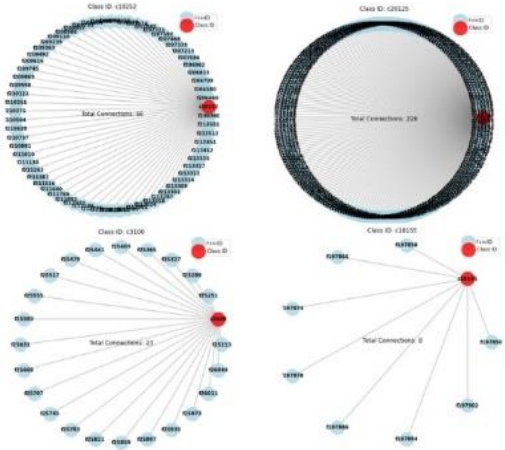
برای پیاده‌سازی و ارزیابی روش ارائه‌شده در این گزارش، از زبان پایتون و برای پیاده‌سازی مدل‌های یادگیری ماشین ، از چارچوب کراس بر روی بستر گوگل کولب استفاده کردیم. داده‌های این پژوهش واقعی هستند و همگی از منابع مختلف مانند کگل[۹] و گیت‌هاب[۱۰] جمع‌آوری گردیده‌اند و متناسب با نیاز پژوهش

با یکدیگر ادغام شدند. برای مؤلفه تعیین‌کننده گره میان توابع سرویس با استفاده از داده‌ها، وابستگی‌های بین توابع را به گراف تبدیل کردیم و هرچقدر که گره‌ای بار محاسباتی سنگینی را به دوش بکشد منابع بیشتری را اشغال می‌کند بنابراین با شناخت این تابع می‌توانیم آن را به بهترین شکل مدیریت کنیم و بیشتر از هر گره دیگری منابع آن را ارتقا دهیم. روش اولی که برای در نظر گرفتن مهم‌ترین گره پیاده‌سازی شد روش درجه مرکزیت بود و روش دوم مرکزیت بردار ویژه است تا به نتیجه‌ای که درروش قبل گرفتیم مطمئن‌تر شویم. با توجه به شکل (۳) و شکل (۴) در هر دو روش توانستیم به مهم‌ترین گره در گراف برسیم که می‌توانیم با اطمینان بالایی نتیجه بگیریم که مهم‌ترین گره در گراف است.

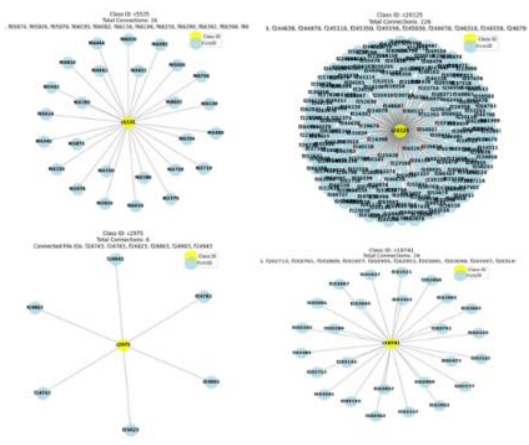
Algorithm 1 — **Consensus of predictions**

INPUTS:
$n_{mlp_{t+1}}$ ← Prediction of mlp neural network t+1
$n_{cnn_{t+1}}$ ← Prediction of cnn neural network t+1
$n_{lstm_{t+1}}$ ← Prediction of lstm neural network t+1
OUTPUTS:
b_{avg}
Begin
mu=0
sigma=1
bayes_avg_preds = []
for i in range(len(test_data)):
likelihood_prod = (1 / (sigma * sqrt(2 * pi))) * exp(−1/2 * (($n_{mlp_{t+1}}$[i] − mu) / sigma)^2) * (1 / (sigma * sqrt(2 * pi))) * exp(−1/2 * (($n_{cnn_{t+1}}$[i] − mu) / sigma)^2) * (1 / (sigma * sqrt(2 * pi))) * exp(−1/2 * (($n_{lstm_{t+1}}$[i] − mu) / sigma)^2)
weights = exp(−1/2 * (($n_{mlp_{t+1}}$[i] − mu) / sigma)^2), exp(−1/2 * (($n_{cnn_{t+1}}$[i] − mu) / sigma)^2), exp(−1/2 * (($n_{lstm_{t+1}}$ [i] − mu) / sigma)^2)
total_weight = sum(weights)
weights = [weight / total_weight for weight in weights]
posterior = sum(weight * likelihood_prod for weight in weights)
bayes_avg_preds.append(posterior)
b_{avg} = max(weights))
Return b_{avg}
End

شکل ۲ شبه کد اجماع نظر پیش‌بینی‌ها



شکل ۳: گراف وابستگی‌ها با روش درجه مرکزیت.



شکل ٤ : گراف وابستگی‌ها با روش مرکزیت بردار ویژه

برای آزمایش مؤلفه مدل‌سازی منابع مصرفی، مقدار تخطی از میزان مشخص مصرف پردازنده در چهار هسته پردازنده، حافظه، زمان اجرای تابع و همچنین مقدار بی‌استفاده ماندن پردازنده، حافظه را ملاک قرار داده‌ایم. برای تصمیم‌گیری در مورد مقیاس‌پذیری دو راه پیش‌روی داشتیم که توابع را برحسب شباهت در یک گروه‌بندی قرار دهیم (رویکردی بدون نظارت) و روش دقیق‌تر، روش یادگیری با نظارت است که برای این روش نیازمند داده‌های آموزشی که در آن مقیاس‌پذیری دخیل باشد وجود داشت، به همین منظور دو روش را پیاده‌سازی کردیم و بعدازآن پیش‌بینی مقیاس‌پذیری و منابع را برای گام زمانی آینده توسط روش با نظارت انجام دادیم و در بین خروجی‌های پیش‌بینی‌شده با روش میانگین‌گیری بیزی اجماع نظر گرفتیم.

۶- تحلیل نتایج

در این بخش نتایج روش پیشنهادی همراه با نمودارهای آن ارائه و برای بررسی کارایی و دقت مؤلفه پیش‌بینی‌کننده منابع، روش پیشنهادی با روش دیگری در همین حوزه مقایسه می‌شود.

روش‌های بدون نظارت در مسئله پژوهش ما که مسئله‌ای حساس است نامناسب است چراکه گروه‌بندی به‌درستی انجام نشده‌است و خطای زیادی دارد. در پژوهش ما هدف بهبود پیش‌بینی مقیاس‌پذیری است، بنابراین روش‌های با نظارت بسیار بهتر از روش‌های بدون نظارت عمل می‌کنند. نتایج به تفکیک روش‌ها در شکل (۵) تا (۷) قابل مشاهده است.

برای ارزیابی مدل‌های یادگیری با نظارت، ۸۰ درصد مجوعه‌داده به‌عنوان داده تمرین و ۱۰ درصد را به‌عنوان داده آزمون و ۱۰درصد را نیز برای ارزیابی استفاده کرده‌ایم و سپس مقادیر تابع‌های هزینه مانند میانگین خطای مربعات و میانگین خطای مطلق و دقت را روی مجموعه داده آزمون به دست آوردیم. شبکه حافظه طولانی کوتاه‌مدت توانسته است نسبت به سایر شبکه‌های عصبی در این مسئله طبقه‌بندی که برحسب سه برچسب مقیاس‌پذیری انجام می‌شد بهتر عمل بکند. همچنین برای اطمینان از خروجی شبکه‌ها از ارزیابی K-Fold Cross Validation استفاده کردیم و داده‌ها را به‌صورت تصادفی در ۵ مرحله آزمایش کردیم و خروجی هر مرحله نیز در جدول (۱) مشخص است.

شکل (۸) مقایسه‌ای از نتایج دقت و خطای پیش‌بینی یادگیری مقیاس‌پذیری برای مدل‌های با نظارت را نشان می‌دهد.

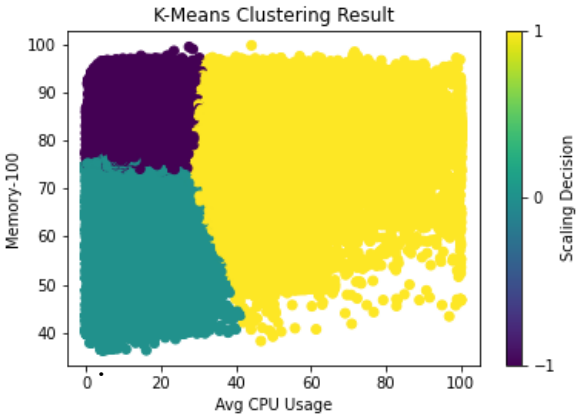
جدول (۲) مقایسه‌ای از نتایج پیش‌بینی منابع درگام کنونی و آینده را برای مدل‌های با نظارت نشان می‌دهد.

۶-۱- بررسی با پژوهش مرتبط

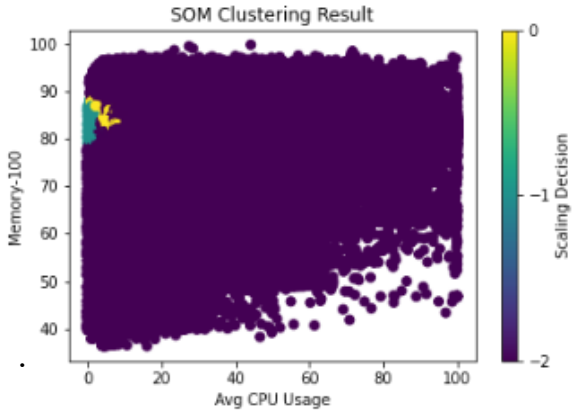
درنهایت نیز پژوهش خود را با روشی که در پژوهش [۱۰] ارائه‌شده بود مقایسه کردیم. این مقایسه از این جهت انجام شد که مدل‌هایی که استفاده شده بود (پیچشی و حافظه کوتاه مدت طولانی) در هر دو پژوهش یکسان بوده است. پژوهش مشابه هر دو مدل را ادغام کرده و از مدل ترکیبی پیچشی- حافظه کوتاه‌مدت طولانی استفاده کرده و معیارهای خطا را برای مصرف CPU گزارش کرده است. در این مقاله ما، سه مدل هوش‌مصنوعی را که به طور خاص پیچشی، حافظه کوتاه مدت طولانی و شبکه پرسپترون چندلایه هستند را به صورت جداگانه آزمایش کرده و سپس برای مقایسه عادلانه فقط تاثیر دو مدل

پیچشی و حافظه کوتاه مدت طولانی را در نظر گرفتیم و نشان دادیم که با در نظر گرفتن شرایط مشابه مدل پیشنهادی ما بهتر از ادغام شبکه‌ها عمل میکند. این امر نشان‌دهنده کارا بودن ترکیب مکانیزم‌های اجماع نظر در مدل‌های یادگیری ماشین است.در جدول (۳) خلاصه‌ای از گزارش‌ها را مشاهده می‌کنید و منظور از CI بازه اطمینان است.

مقایسه نشان می‌دهد مدل پیشنهادی با مقدار RMSE برابر ۰.۶۹ برای آموزش شبکه پیچشی و ۰.۱۹ برای شبکه حافظه طولانی کوتاه‌مدت خطای پایین‌تری نسبت به مقاله موردبحث با مقدار ۷.۱۴ دارد و مدل پیشنهادی با نرخ پایین‌تری از MAEو RMSE وMSE همراه است و بازه اطمینان نیز گزارش شده‌است.



شکل ۵ : مقیاس‌پذیری با روش خوشه‌بندی

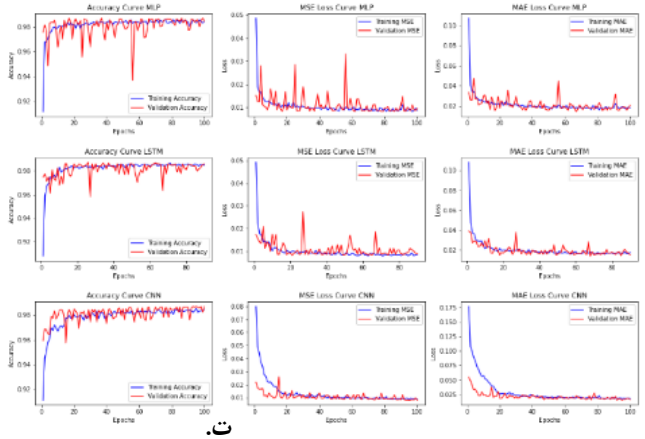


شکل ٦ : مقیاس‌پذیری با روش نقشه‌های خودسازمان‌دهی

شکل ۷ : مقیاس‌پذیری با روش تجزیه‌وتحلیل مؤلفه اصلی

۷- نتیجه‌گیری

در این پژوهش به بررسی ارائه یک رویکرد مقیاس‌پذیری خودکار در محیط بدون‌سرویس‌دهنده پرداخته شد و در ادامه ارزیابی جامعی بر روی الگوریتم پیشنهادی پیاده‌سازی شده با دو رویکرد با نظارت و بدون نظارت انجام شد.



ت.

شکل ۸ : میانگین خطای مطلق و مربعات و دقت مراحل تمرین و آزمون مامی شبکه‌ها درروش بانظار

جدول ۱) : نتایج دقت و خطای پیش‌بینی یادگیری مقیاس‌پذیری

MAE	*MSE*	*F1-Score*	*Recall*	*Precision*	*Accuracy*	
0.02	0.04	0.98	0.98	0.98	0.984	*MLP-Fold 1*
0.01	0.02	0.99	0.99	0.98	0.99	*MLP-Fold 2*
0.02	0.03	0.98	0.99	0.98	0.9875	*MLP-Fold 3*
0.01	0.02	0.99	0.99	0.98	0.989	*MLP-Fold 4*
0.02	0.02	0.98	0.99	0.98	0.987	*MLP-Fold 5*
0.01	0.01	0.99	0.99	0.99	0.992	*LSTM-Fold 1*
0.01	0.02	0.99	0.99	0.98	0.9895	*LSTM-Fold 2*
0.01	0.01	0.99	0.99	0.98	0.9905	*LSTM-Fold 3*
0.01	0.01	0.99	0.99	0.98	0.9905	*LSTM-Fold 4*
0.01	0.01	0.99	0.99	0.98	0.9905	*LSTM-Fold 5*
0.02	0.02	0.98	0.99	0.98	0.9875	*CNN-Fold 1*
0.02	0.03	0.98	0.99	0.98	0.9865	*CNN-Fold 2*
0.05	0.09	0.97	0.97	0.97	0.9715	*CNN-Fold 3*
0.01	0.02	0.99	0.99	0.98	0.99	*CNN-Fold 4*
0.01	0.02	0.99	0.99	0.98	0.9895	*CNN-Fold 5*

جدول ۲) : مقادیر به‌دست‌آمده از معیارهای ارزیابی مدل پیشنهادی

مقادیر به‌دست‌آمده از معیارهای ارزیابی مدل پیشنهادی

Actual Value for t	MLP Prediction for t	LSTM Prediction for t	CNN Prediction for t	Bayesian Avg Prediction for t
10.3520393	10.236974	10.364485	9.263559	9.934422
5.2734375	5.160522	5.405163	4.7756	5.103557
29.3650742	29.177689	29.725983	29.035881	29.304556
Actual Value for t+1	**MLP Prediction for t+1**	**LSTM Prediction for t+1**	**CNN Prediction for t+1**	**Bayesian Avg Prediction for t+1**
64.19699	63.02586	64.61939	46.664474	59.63367
64.19621	62.944954	64.57348	46.92247	59.659576
64.20816	63.03141	64.62794	47.328598	59.811844

جدول ۳) : مقادیر به‌دست‌آمده از معیارهای ارزیابی مدل پیشنهادی

	روش پیشنهادی این پژوهش				روش پیشنهادی پژوهش مرتبط			
	Model	MSE	MAE	RMSE	Model	MSE	MAE	RMSE
Train	MLP	2.40	1.25	1.55	CNN-LSTM	26.90	2.64	5.18
	CNN	0.46	0.58	0.68				
	LSTM	0.07	0.15	0.26				
Test	MLP	2.48	1.27	1.57	CNN-LSTM	51.09	3.72	7.14
	CNN	0.47	0.59	0.69				
	LSTM	0.03	0.15	0.19				
CI	MLP	2.11-2.89	1.22-1.33	1.45-1.70	CNN-LSTM	188.11-208.49	8.56-9.15	13.71-14.47
	CNN	0.44-0.52	0.57-0.61	0.66-0.72				
	LSTM	0.03-0.04	0.14-0.16	0.17-0.22				

نتایج نشان می‌دهد که الگوریتم پیشنهادی با استفاده از ترکیب مدل‌های شبکه عصبی پرسپترون و حافظه طولانی کوتاه‌مدت و پیچشی می‌تواند خطاهای پیش‌بینی را به حداقل برساند و دقت بالاتری را ارائه دهد. مشاهدات به‌وضوح نشان می‌دهد که الگوریتم پیشنهادی با پایین‌ترین مقادیر MAE و RMSE و MSE، عملکرد بهتری نسبت به روش مقایسه شده دارد.

مراجع

[1] Jonas, Eric, Schleier-Smith, Johann, Tsai, Chia-Che, “Cloud Programming Simplified: A Berkeley View on Serverless Computing,” arXiv preprint arXiv: 1902.03383, Feb 9, 2019.

[2] J. V. Bibal Benifa, D. Dejey, “RLPAS: Reinforcement Learning-based Proactive Auto-Scaler for Resource Provisioning in Cloud Environment,” Journal of Mobile Networks Application, vol. 24, no. 4, pp. 1348–1363, 2019.

[3] Almeida, V., Arlitt, M. ,Rolia, J., “Analyzing a web-based system’s performance measures at multiple time scales,” Journal of ACM Sigmetrics Performance Evaluation Review, vol. 30, no. 2, pp. 3–9, Sep. 2002.

[4] Golshani, E., Ashtiani, M., “Proactive auto-scaling for cloud environments using temporal convolutional neural networks,” Journal of Parallel and Distributed Computing, vol. 154, no.4, pp. 119–141, 2021.

[5] Al-Roomi, M., Al-Ebrahim S., Buqrais, S., Ahmad, I., "Cloud Computing Pricing Models: A Survey," *Journal of Grid and Distributed Computing*, vol. 6, no. 5. pp. 93-106, 2013.

[6] Lorido-Botrán, T., Miguel-Alonso, J., Lozano, J. A., “Comparison of Auto-scaling Techniques for Cloud Environments,” Journal of Parallelism, Nov 2013, Paris, France, p.p 56-64.

[7] Martínez, R. G., Li, Z., Lopes, A., Rodrigues, L., "Augure: Proactive reconfiguration of cloud applications using heterogeneous resources," in Proceedings of the 2017 IEEE 16th International Symposium on Network Computing and Applications (NCA), pp. 1-8, Cambridge, MA, USA, November, 2017.

[8] Anastasios Z., Eleni F., Nikos F., Symeon P.,” Reinforcement learning-assisted autoscaling mechanisms for serverless computing platforms”, *Simulation Modelling Practice and Theory*, Vol. 116, 2022.

[9] Siddharth A. and Rodriguez M.A. and Buyya R.,” A Deep Recurrent-Reinforcement Learning Method for Intelligent AutoScaling of Serverless Functions”, 2023.

[10] Ouhame, S., Hadi, Y., Ullah, A., “An efficient forecasting approach for resource utilization in a cloud data center using CNN-LSTM model,” Journal of Neural Computing and Applications, vol. 33, no. 16, pp. 10043–10055, 2021.

زیرنویس

[1] model-free Proximal Policy Optimisation
[2] Principal Component Analysis
[3] Clustering
[4] Self-Organizing Maps
[5] MLP
[6] LSTM (Long Short-Term Memory
[7] RNN
[8] CNN
[9] Kaggle Code Metrics Dataset SoftwareProjectStructure
[10] Github Hardware Performance Archive