Journal of New Technologies in Distributed Systems and Algorithmic Computing
Islamic Azad University of Sabzevar
E-ISSN: 3115-705X
https://sanad.iau.ir/journal/ntds

Research Article

Design and implementation of distributed systems for big data processing using artificial intelligence algorithms

Rahim Karimi*[1]

[1]Department of Mathematics Education, Farhangian University, Tehran, Iran, **Rahim.karimi@iau.ir**

Correspondence

***Rahim Karimi**, Department of Mathematics Education, Farhangian University, Tehran, Iran,

Rahim.karimi@iau.ir

Main Subjects: Distributed Systems
Received: 22 May 2025
Revised:
Accepted: 28 July 2025

https://doi.org/10.82195/NTDS.2025.1207690

Abstract

In this paper, the design and implementation of distributed systems for big data processing using artificial intelligence algorithms are examined. With the rapid growth of data volumes in today's world, the use of such systems and AI-based algorithms for data processing has become increasingly important. The findings indicate that these methods can significantly improve the performance of big data processing and offer substantial benefits for organizations and enterprises. The application of distributed systems and artificial intelligence algorithms for big data processing can thus lead to notable enhancements in the efficiency and effectiveness of various systems and applications. Consequently, distributed systems provide a versatile and flexible approach to computing. By leveraging the collective power of multiple nodes, these systems can easily handle complex tasks. Although challenges such as data consistency may arise, the scalability and fault tolerance of distributed systems make them a valuable tool in today's digital landscape. This study demonstrates that employing distributed systems and AI algorithms can bring considerable improvements in the performance and efficiency of diverse systems and applications.

Keywords: Distributed Systems, Big Data Processing, Artificial Intelligence Algorithms, Performance Improvement, Systems Efficiency.

فناوری های نوین در سیستم های توزیع شده و محاسبات الگوریتمی

دانشگاه آزاد اسلامی واحد سبزوار

E-ISSN: 3115-705X
https://sanad.iau.ir/journal/ntds

پژوهشی

طراحی و پیاده‌سازی سیستم‌های توزیع شده برای پردازش بیگ دیتا با استفاده از الگوریتم‌های هوش مصنوعی

رحیم کریمی* [1]

[1] گروه آموزش ریاضی، دانشگاه فرهنگیان، تهران، ایران،
rahim.karimi@iau.ir ،

نویسنده مسئول
*رحیم کریمی، دکتری مهندسی فناوری اطلاعات، گروه آموزش ریاضی، دانشگاه فرهنگیان، تهران، ایران،
rahim.karimi@iau.ir ،

عنوان اصلی: سیستم‌های توزیع شده
تاریخ دریافت: ۱ خرداد ۱۴۰۴
تاریخ بازنگری:
تاریخ پذیرش: ۶ مرداد ۱۴۰۴

https://doi.org/10.82195/NTDS.2025.1207690

چکیده:

در این مقاله، به بررسی طراحی و پیاده‌سازی سیستم‌های توزیع شده برای پردازش بیگ دیتا با استفاده از الگوریتم‌های هوش مصنوعی پرداخته شده است. با توجه به رشد رو به افزایش حجم داده‌ها در دنیای امروز، استفاده از این سیستم‌ها و الگوریتم‌های هوش مصنوعی برای پردازش داده‌ها اهمیت بیشتری یافته است. نتایج به دست آمده نشان می‌دهد که این روش‌ها می‌توانند بهبود قابل توجهی در عملکرد پردازش بیگ دیتا ایجاد کنند و از مزایای قابل توجهی برای سازمان‌ها و شرکت‌ها به دنبال داشته باشند. که استفاده از سیستم‌های توزیع شده و الگوریتم‌های هوش مصنوعی برای پردازش بیگ دیتا می‌تواند بهبود قابل توجهی در عملکرد و کارایی سیستم‌ها و برنامه‌های مختلف ایجاد کند. در نتیجه، سیستم های توزیع شده یک رویکرد همه کاره و انعطاف پذیر برای محاسبات ارائه می دهند. این سیستم ها با استفاده از توان جمعی گره های متعدد، می توانند وظایف پیچیده را به راحتی انجام دهند. در حالی که چالش‌هایی مانند سازگاری داده‌ها ممکن است ایجاد شود، مزایای مقیاس‌پذیری و تحمل خطا، سیستم‌های توزیع‌شده را به ابزاری ارزشمند در چشم‌انداز دیجیتال امروزی تبدیل می‌کند. این مقاله نشان می‌دهد که استفاده از سیستم‌های توزیع شده و الگوریتم‌های هوش مصنوعی می‌تواند بهبود قابل توجهی در عملکرد و کارایی سیستم‌ها و برنامه‌های مختلف ایجاد کند.

کلید واژه‌ها: سیستم‌های توزیع شده، پردازش بیگ دیتا، الگوریتم‌های هوش مصنوعی، بهبود عملکرد، کارایی سیستم‌ها.

۱-مقدمه

کلان داده[1] به مجموعه‌ای از کلان داده‌ها و پیچیده اشاره دارد که به صورت مستمر و با سرعت بالا تولید می‌شوند. این داده‌ها معمولاً از منابع مختلفی مانند سنسورها، دستگاه‌های مختلف، شبکه‌های اجتماعی، وب‌سایت‌ها، دیتابیس‌ها و سایر منابع جمع‌آوری می‌شوند. کلان داده به دلیل حجم بالا، سرعت تولید، و تنوع اطلاعات موجود در آن شناخته می‌شود. این داده‌ها معمولاً به صورت ساختارمند و غیرساختارمند (مانند متن، تصاویر، صدا و ویدیو) وجود دارند. از دیدگاه فناوری، کلان داده معمولاً با استفاده از فناوری‌های پردازش توزیع‌شده، پایگاه داده‌هایNoSQL ، ابزارهای تحلیل داده، فناوری‌های ذخیره‌سازی ابری و ابزارهای مدیریت داده و سایر فناوری‌های مرتبط پردازش می‌شوند [۱].

استفاده از کلان داده به شرکت‌ها و سازمان‌ها امکان می‌دهد تا از اطلاعات موجود در کلان داده‌ها بهره‌برداری کنند و از آن‌ها برای تصمیم‌گیری‌های بهتر و پیش‌بینی‌های دقیق‌تر استفاده کنند. به عنوان مثال، از کلان داده می‌توان برای تحلیل رفتار مشتریان، پیش‌بینی روند بازار، بهبود فرآیندهای تولید و سرویس‌دهی، تحلیل داده‌های پزشکی و بهبود خدمات بهداشتی و سلامت استفاده کرد [۲]. کلان داده با ویژگی‌های خاصی که از جمله حجم بالا، سرعت تولید و تنوع داده‌ها است، چالش‌های منحصر به فردی را برای سازمان‌ها و محققان دارد. برخی از این چالش‌ها عبارتند از [۳، ۴]:

- حجم بالا: مدیریت و ذخیره سازی کلان داده‌ها به چالش کشیدن سیستم‌های ذخیره‌سازی و پردازش داده می‌پردازد. این امر نیازمند فناوری‌های پردازش توزیع‌شده و ذخیره‌سازی ابری است.
- سرعت تولید: داده‌های کلان به سرعت بالای تولید و به روزرسانی نیاز دارند که این موضوع نیازمند تکنولوژی‌های پردازش و ذخیره‌سازی با سرعت بالا و قابلیت بالای مقیاس‌پذیری است.
- تنوع داده‌ها: داده‌های کلان ممکن است از منابع مختلف و با فرمت‌های مختلفی مانند متن، تصاویر، ویدیو و صدا باشند. چالش اصلی در اینجا این است که چگونه این داده‌های متنوع را یکپارچه کرده و تحلیل کنیم.
- امنیت و حریم خصوصی: حفظ امنیت و حریم خصوصی داده‌های کلان یک چالش اساسی است. مدیریت دسترسی، رمزنگاری، و حفاظت از داده‌ها مسائلی است که باید به آنها توجه شود.
- تحلیل و استفاده از داده: تحلیل و استفاده از داده‌های کلان نیازمند تولنایی‌های تحلیلی پیچیده و مدل‌سازی پیشرفته است. همچنین، اطمینان از صحت و قابل اعتماد بودن داده‌ها نیز یک چالش است.
- مدیریت فرآیندها و استراتژی‌ها: تصمیم‌گیری‌های مرتبط با مدیریت داده‌های کلان و تعیین استراتژی‌های مناسب برای بهره‌برداری از این داده‌ها نیازمند داشتن دانش و تجربه کافی است.
- هزینه: پیاده‌سازی و مدیریت یک سیستم کلان داده هزینه‌بر است و نیازمند سرمایه‌گذاری‌های قابل توجهی است.

این چالش‌ها نشان‌دهنده اهمیت و نیاز به داشتن استراتژی‌ها و فناوری‌های مناسب برای مدیریت داده‌های کلان است. با وجود اهمیت بسزایی که کلان داده‌ها و پردازش آن‌ها در سازمان‌های گوناگون ایفا می‌کند. همچنان مسائلی در خصوص نحوه مدیریت آن‌ها قرار دارد. در این میان حوزه مالی یکی از حوزه‌های فعالی است که با بحث کلان داده‌ها ارتباط زیادی دارد. بررسی‌های صورت گرفته نشان می‌دهد هنوز چالش‌های زیادی بر سر راه بهره‌گیری از کلان داده‌ها در امور مالی قرار دارد. بدین منظور، مسئله اصلی این پژوهش شناسایی کاربردهای و چالش‌های کلان داده‌ها در مالی است [۵].

در علوم کامپیوتر، هوش مصنوعی یا هوش ماشینی به هوشمندی گفته می‌شود که از هرنوع ماشین (و نه انسان) به‌دست بیاید. کتاب‌های مرجع در حوزه‌ی هوش مصنوعی، این علم را دانش مطالعه‌ی کارگزارهای هوشمند می‌دانند که چنین تعریف می‌شوند: «هر دستگاهی که توانایی درک محیط و فعالیت با حداکثر شانس موفقیت را داشته باشد». درمجموع اصطلاح هوش مصنوعی برای توصیف ماشین‌ها یا کامپیوترهایی به‌کار می‌رود که فعالیت‌های شناختی وابسته به ذهن انسان را به‌خوبی انجام دهند. ازمیان فعالیت‌های مهم شناختی می‌توان به «یادگیری» و «حل مسئله» اشاره کرد. حوزه‌ی تحقیق پیرامون هوش مصنوعی، در سال ۱۹۵۶ و آزمایشگاهی در کالج دارتموث متولد شد. جان مک‌کارتی این حوزه را از زیرمجموعه‌ی سایبرنتیک و نظریه‌های سایبرنتیست‌هایی همچون نوربرت وینر خارج کرد و اصطلاح «هوش مصنوعی» به‌نوعی توسط او متولد شد [۵].

[1] Big Data

در عصر کلان داده‌ها، طراحی و پیاده‌سازی سیستم‌های توزیع‌شده نقش مهمی در پردازش کارآمد حجم وسیعی از داده‌ها ایفا می‌کند. الگوریتم‌های هوش مصنوعی به‌عنوان ابزارهای قدرتمندی در این حوزه ظاهر شده‌اند که امکان خودکارسازی وظایف پیچیده و استخراج بینش‌های ارزشمند از مجموعه کلان داده‌ها را فراهم می‌کنند.

وقتی نوبت به طراحی سیستم های توزیع شده برای پردازش کلان داده می شود، باید چندین فاکتور را در نظر گرفت. مقیاس پذیری، تحمل خطا و سازگاری داده ها از جمله چالش های کلیدی هستند که باید مورد توجه قرار گیرند. با استفاده از الگوریتم‌های هوش مصنوعی، مانند یادگیری ماشینی و یادگیری عمیق، می‌توان این سیستم‌ها را برای مدیریت حجم عظیمی از داده‌ها و ارائه تجزیه و تحلیل در زمان واقعی بهینه کرد.

پیاده سازی سیستم های توزیع شده برای پردازش داده های بزرگ با استفاده از الگوریتم های هوش مصنوعی شامل یکپارچه سازی فناوری ها و چارچوب های مختلف استApache Hadoop ،. Sparkو TensorFlow برخی از ابزارهای محبوب مورد استفاده در این زمینه هستند. این پلتفرم ها زیرساخت های لازم را برای پردازش موازی، ذخیره سازی توزیع شده و پردازش کارآمد داده ها فراهم می کنند [۶]. در نتیجه، طراحی و پیاده‌سازی سیستم‌های توزیع‌شده برای پردازش کلان داده‌ها با استفاده از الگوریتم‌های هوش مصنوعی برای سازمان‌هایی که به دنبال مهار قدرت بینش‌های مبتنی بر داده هستند ضروری است. با استفاده از قابلیت‌های هوش مصنوعی، کسب‌وکارها می‌توانند فرصت‌های جدیدی را برای نوآوری، بهینه‌سازی و تصمیم‌گیری در دنیای داده‌محور امروزی باز کنند [۷]. با توجه به اهمیت پیاده‌سازی سیستم‌های توزیع‌شده برای پردازش کلان داده‌ها مسئله اصلی این پژوهش بررسی ملزومات طراحی و پیاده‌سازی سیستم‌های توزیع شده برای پردازش بیگ دیتا با استفاده از الگوریتم‌های هوش مصنوعی است. نوآوری این پژوهش، توسعه سیستم هوشمند خودتنظیمی مبتنی بر یادگیری عمیق برای مدیریت منابع در سیستم‌های توزیع‌شده بیگ دیتا است. در این سیستم، از الگوریتم‌های یادگیری عمیق برای تحلیل وضعیت جاری و پیش‌بینی بار کاری و نیازهای منابع در آینده استفاده می‌شود. سیستم به طور پویا و خودکار، الگوهای بهینه تخصیص منابع (مانند حافظه، پردازنده، شبکه) را در محیط‌های توزیع‌شده تنظیم می‌کند، بدون نیاز به دخالت انسانی مداوم. این رویکرد، برتری‌هایی چون کاهش تأخیر، مصرف بهینه انرژی و افزایش مقیاس‌پذیری در پردازش بیگ دیتا را ایجاد می‌کند.

۲. روش پژوهش

این پژوهش از نظر هدف کاربردی است و با استفاده از روش مرور نظام‌مند انجام شد. بدین منظور به پایگاه‌های اطلاعاتی معتبر همانند Emerald، Google Scholar، Web of Science، Scopus مراجعه شد. همچنین پایگاه‌های داخلی مگیران، پرتال جامع علوم انسانی، نورمگز، ایرانداک و مرکز اطلاعات علمی جهاد دانشگاهی مورد جستجو قرار گرفتند. نخست، فرآیند جستجو جهت شناسایی، استخراج و نیز انتخاب مطالعات و پژوهش‌های مرتبط مورد توجه قرار گرفت. این فرآیند برای شناسایی هرگونه مطالعه مرتبط بالقوه بر اساس پرسش های پژوهش، انجام می‌شود. فرآیند انتخاب شامل چهار مرحله یعنی انتخاب پایگاه‌های اطلاعاتی، جستجوی کلیدواژه‌ها، معیارهای انتخاب پژوهش‌ها و انتخاب پژوهش‌های اولیه بود. برای جستجو از کلیدواژه‌های کلان داده و سیستم‌های توزیع‌شده استفاده شد. پس از جستجو در پایگاه‌های اطلاعاتی تعداد ۴۵۱ مقاله بازیابی شدند. پس از پالایش موارد تکراری که از پایگاه‌های گوناگون شناسایی شده بودند، تعداد ۲۲۹ مقاله باقی ماند. با مطالعه عنوان و چکیده منابع تعداد ۱۸۱ پژوهش به دلیل غیر مرتبط بودن با هدف پژوهش، حذف گردید. انتخاب پژوهش‌های اولیه بر اساس معیارهای ورود و خروج صورت پذیرفت. سپس متن کامل مقالات مطالعه شده و معیارهای ورود و خروج اعمال گردید و در نتیجه، ۲۵ پژوهش حذف گردید. در نهایت، ۲۳ مقاله به مطالعه مروری راه یافتند.

۳. یافته‌های پژوهش

بررسی منابع گوناگون نشان داد، سیستم های توزیع شده به چندین کامپیوتر متصل به هم اشاره دارد که با یکدیگر برای دستیابی به یک هدف مشترک کار می کنند. این سیستم ها وظایف را در گره های مختلف توزیع می کنند و امکان افزایش

کارایی و تحمل خطا را فراهم می کنند. با تقسیم یک کار به وظایف فرعی کوچکتر که توسط گره های مختلف انجام می شود، سیستم های توزیع شده می توانند فرآیندهای پیچیده را به طور موثرتری نسبت به یک سیستم متمرکز انجام دهند [۶].
یکی از مزایای کلیدی سیستم‌های توزیع شده مقیاس‌پذیری است. با افزایش حجم کار، گره‌های اضافی را می‌توان به‌راحتی به سیستم اضافه کرد تا بار اضافی را مدیریت کند. این مقیاس‌پذیری سیستم‌های توزیع‌شده را برای برنامه‌هایی با تقاضای نوسانی ایده‌آل می‌کند و عملکردی روان را حتی در زمان‌های اوج مصرف تضمین می‌کند [۷].

۳-۱. چالش های سیستم های توزیع شده

درحالی‌که سیستم‌های توزیع شده مزایای متعددی را ارائه می‌دهند، اما چالش‌های منحصربه‌فردی را نیز ارائه می‌دهند. یکی از این چالش‌ها اطمینان از سازگاری داده‌ها در تمام گره ها است. با توزیع داده ها در چندین مکان، حفظ ثبات برای جلوگیری از اختلافات و خطاها بسیار مهم است.

۳-۲. انواع سیستم های توزیع شده

انواع مختلفی از سیستم‌های توزیع شده وجود دارد که هر کدام اهداف متفاوتی را انجام می‌دهند. شبکه‌های همتابه‌همتا (P2P)به گره‌ها اجازه می‌دهند بدون سرور مرکزی مستقیماً با یکدیگر ارتباط برقرار کنند. از سوی دیگر، شبکه‌های سرویس‌دهنده - کلینت شامل مشتریانی است که از سرورهای مرکزی خدمات درخواست می‌کنند. علاوه بر این، سیستم‌های محاسباتی توزیع‌شده از چندین رایانه برای کار با هم در یک کار واحد استفاده می‌کنند.

۳-۳. کاربردهای دنیای واقعی سیستم های توزیع شده

سیستم‌های توزیع شده نقش مهمی در فناوری مدرن ایفا می‌کنند. خدمات رایانش ابری، پلتفرم‌های رسانه‌های اجتماعی و بازارهای آنلاین همگی به سیستم‌های توزیع‌شده برای ارائه تجربیات یکپارچه کاربر متکی هستند. با استفاده از قدرت محاسبات توزیع‌شده، این پلتفرم‌ها می‌توانند به طور مؤثر حجم وسیعی از داده‌ها را پردازش کرده و خدمات قابل‌اعتمادی را به کاربران در سراسر جهان ارائه دهند.
امروزه، پردازش کلان‌داده به یک جنبه حیاتی در بسیاری از صنایع، از تجارت الکترونیک گرفته تا مراقبت‌های بهداشتی تبدیل شده است. ازآنجایی‌که حجم داده‌ها به طور تصاعدی درحال‌رشد است، روش‌های سنتی پردازش داده‌ها دیگر کافی نیستند. این منجر به ظهور سیستم‌های توزیع‌شده‌ای شده است که از الگوریتم‌های هوش مصنوعی برای مدیریت کارآمد حجم عظیمی از داده‌ها استفاده می‌کنند.
الگوریتم‌های هوش مصنوعی نقش کلیدی در بهینه‌سازی پردازش داده‌های بزرگ در سیستم‌های توزیع شده ایفا می‌کنند. الگوریتم‌های یادگیری ماشین، مانند شبکه‌های عصبی و درخت‌های تصمیم، می‌توانند مجموعه‌های داده بزرگ را تجزیه‌وتحلیل کنند و بینش‌های ارزشمند را با سرعت و دقت استخراج کنند.
هنگام طراحی سیستم‌های توزیع شده برای پردازش داده‌های بزرگ، درنظرگرفتن عواملی مانند تحمل خطا، مقیاس‌پذیری و تعادل بار ضروری است. این سیستم‌ها با توزیع بار کاری بین گره‌های متعدد، می‌توانند حجم زیادی از داده‌ها را بدون مواجهه با تنگناها مدیریت کنند. پیاده‌سازی سیستم‌های توزیع‌شده با الگوریتم‌های هوش مصنوعی با چالش‌هایی مانند ثبات داده‌ها و هزینه‌های ارتباطی همراه است. بااین‌حال، پیشرفت در فناوری‌هایی مانند Apache Hadoop و Spark غلبه بر این چالش‌ها و ساخت سیستم‌های قوی برای پردازش داده‌های بزرگ را آسان‌تر کرده است.
در نتیجه، طراحی و پیاده‌سازی سیستم‌های توزیع‌شده برای پردازش کلان‌داده‌ها با استفاده از الگوریتم‌های هوش مصنوعی، شیوه تحلیل و استخراج بینش‌های ما از مجموعه کلان‌داده‌ها را متحول کرده است. با استفاده از قدرت هوش مصنوعی و محاسبات توزیع‌شده، سازمان‌ها می‌توانند پتانسیل کامل داده‌های خود را باز کنند و در دنیای مبتنی بر داده امروزی مزیت رقابتی کسب کنند.

در طراحی و پیاده‌سازی سیستم‌های توزیع شده، برخی از الگوریتم‌های هوش مصنوعی که قابل‌استفاده در کلان‌داده‌ها هستند عبارتند از:

جدول ۱. الگوریتم‌های هوش مصنوعی قابل استفاده در کلان داده‌ها
Table 1: AI Algorithms Applicable to Big Data

ردیف	الگوریتم	توضیح الگوریتم
۱	الگوریتم K-Means	این الگوریتم برای خوشه‌بندی داده‌ها به کار می‌رود. با استفاده از این الگوریتم، داده‌ها به چند خوشه مختلف تقسیم می‌شوند به‌گونه‌ای که داده‌های هر خوشه شباهت زیادی به یکدیگر داشته باشند.
۲	شبکه‌های عصبی عمیق	این الگوریتم‌ها بر اساس ساختار شبکه‌های عصبی انسانی طراحی شده‌اند و برای تشخیص الگوها و پیش‌بینی داده‌ها از طریق یادگیری عمیق استفاده می‌شوند.
۳	الگوریتم Random Forest	این الگوریتم بر اساس مفهوم یادگیری ماشین تصمیم‌گیری چندگانه استفاده می‌شود. با استفاده از این الگوریتم، می‌توان بهترین تصمیم‌ها را برای پیش‌بینی داده‌ها از مجموعه تصمیم‌های تصادفی گرفت.
۴	الگوریتم	یک الگوریتم یادگیری ماشین است که برای دسته‌بندی و رگرسیون داده‌ها استفاده می‌شود. این الگوریتم با استفاده از یک هسته (kernel)، داده‌ها را به خطی یا غیرخطی تقسیم می‌کند.
۵	الگوریتم Apriori	این الگوریتم برای استخراج الگوهای تکراری از داده‌ها به کار می‌رود. با استفاده از این الگوریتم می‌توان الگوهای معنی‌دار و مفیدی را در داده‌های کلان شناسایی کرد.
۶	الگوریتم درخت تصمیم	این الگوریتم برای ایجاد مدل‌های تصمیم‌گیری استفاده می‌شود. با استفاده از این الگوریتم، داده‌ها بر اساس سؤالات دودویی تقسیم می‌شوند تا بهترین تصمیم‌ها برای دسته‌بندی داده‌ها گرفته شود.
۷	الگوریتم K-NN	الگوریتم نزدیک‌ترین همسایگان (k-NN) یک روش غیرپارامتری برای طبقه‌بندی است که برای حل بسیاری از مسائل طبقه‌بندی استفاده می‌شود.

الگوریتم‌هایی که در جدول ۱ ارائه شده‌اند، تنها چند نمونه از الگوریتم‌های هوش مصنوعی هستند که در کلان داده‌ها مورد استفاده قرار می‌گیرند و به تحلیل و استخراج اطلاعات مفید از کلان داده‌ها کمک می‌کنند. به‌طورکلی می‌توان بیان داشت، نقش الگوریتم‌های یادگیری ماشین در سیستم های توزیع شده در بهینه سازی عملکرد و کارایی بسیار مهم است. این الگوریتم ها برای انجام وظایف پیچیده مانند پردازش داده ها، تخصیص منابع و زمان بندی وظایف در یک محیط توزیع شده طراحی شده اند. الگوریتم‌های ماشینی در سیستم‌های توزیع‌شده به مدیریت توزیع بار کاری در بین گره‌های متعدد کمک می‌کنند، و تضمین می‌کنند که وظایف به موقع اجرا می‌شوند و منابع به‌طور مؤثر مورد استفاده قرار می‌گیرند. آنها همچنین نقش کلیدی در تحمل خطا و متعادل سازی بار دارند و تضمین می کنند که سیستم می تواند حتی در مواجهه با خرابی ها یا تقاضای بالا به عملکرد خود ادامه دهد. علاوه بر این، الگوریتم‌های ماشینی در سیستم‌های توزیع‌شده، فرآیندهای تصمیم‌گیری هوشمند را امکان‌پذیر می‌کنند، مانند پیش‌بینی منابع مورد نیاز بر اساس داده‌های تاریخی یا بهینه‌سازی اجرای کار بر اساس شرایط فعلی

سیستم. این منجر به بهبود عملکرد سیستم و رضایت کلی کاربر می شود. در نتیجه، نقش الگوریتم‌های ماشین در سیستم‌های توزیع‌شده برای بهینه‌سازی عملکرد، اطمینان از تحمل خطا، و فعال کردن فرآیندهای تصمیم‌گیری هوشمند ضروری است. با استفاده از این الگوریتم‌ها، سیستم‌های توزیع‌شده می‌توانند به سطوح بالاتری از کارایی و قابلیت اطمینان دست یابند که در نهایت منجر به تجربه کاربری بهتر می‌شود. در ادامه توضیحات بیشتری در مورد الگوریتم‌های مذکور ارائه شده است:

الگوریتم‌های یادگیری بدون نظارت: خوشه‌بندی

الگوریتم‌های خوشه‌بندی نیز برای شناسایی الگوهای معمول عملیات ساختمان از داده‌های عملیات ساختمان مانند الگوهای مصرف انرژی ساختمان، الگوی توزیع محیط داخلی و الگوهای عملکرد سیستم انرژی ساختمان استفاده می‌شوند. الگوریتم‌های خوشه‌بندی بر اساس شباهت آماری بین هر یک از دونقطه، همه نقاط یک مجموعه‌داده را به چندین خوشه طبقه‌بندی می‌کنند. نقاط یک خوشه دارای ویژگی‌های آماری مشابه هستند و نقاط در خوشه‌های مختلف دارای ویژگی‌های آماری قابل‌توجهی متفاوت هستند. به‌طورکلی، شرایط مختلف عملکرد ویژگی‌های آماری متفاوتی دارند [۲].

خوشه‌بندی K-means یکی از محبوب‌ترین الگوریتم‌های خوشه‌بندی در حوزه سیستم‌های انرژی ساختمان است. خوشه‌بندی c-means فازی نیز برای شناسایی الگوهای عملکرد ساختمان به کار گرفته شد. الگوریتم‌های خوشه‌بندی دیگر نیز مانند خوشه‌بندی بردار پشتیبانی، خوشه‌بندی حداکثر انتظارات و خوشه‌بندی درخت تصمیم‌گیری استفاده شده است و این اهمیت این الگوریتم‌ها را بیش‌ازپیش مشخص می‌سازد. الگوریتم‌های استخراج نمودار، الگوریتم‌های استخراج متن و الگوریتم‌های قواعد انجمنی پویا نیز برای تشخیص خطای سیستم‌های انرژی ساختمان استفاده شده است [۳].

نحوه کار الگوریتم k-means به شرح زیر است:

مرحله ۱: برای تصمیم گیری در مورد تعداد خوشه ها ، تعداد K را انتخاب می‌شود.

مرحله ۲ K : تا از نقاط را به‌صورت تصادفی یا با محاسبه انتخاب می‌شود. (این می‌تواند غیر از مجموعه‌داده ورودی باشد) بر اساس کد زیر از فاصله‌ی اقلیدوسی برای انتخاب مراکز استفاده شده است.

مرحله ۳: هر نقطه داده را به نزدیکترین مرکز خود اختصاص می‌دهد، که خوشه های K از پیش تعریف شده را تشکیل می‌دهد.

مرحله ۴: میانگین را محاسبه کرده و یک مرکز جدید برای هر خوشه قرار می‌دهد.

مرحله ۵: مراحل سوم را تکرار می‌شود، به این معنی که هر پایگاه داده را به جدیدترین و نزدیکترین مرکز هر خوشه اختصاص می‌دهد.

مرحله ۶: اگر تغییر مجددی اتفاق افتاد، سپس مرحله ۴ مجدد اجرا می‌شود و الگوریتم به پایان می‌رسد.

مرحله ۷: مدل آماده است.

درخت تصمیم‌گیری: یادگیری درخت تصمیم یکی از روش‌های مدل‌سازی پیش‌بینی‌کننده است که در آمار، داده‌کاوی و یادگیری ماشین استفاده می‌شود. از درخت تصمیم استفاده می‌کند تا از مشاهدات مربوط به یک مورد به نتیجه‌گیری در مورد ارزش مورد (که در برگ نشان‌داده‌شده است) برسد [۷]. الگوریتم درخت تصمیم در دسته یادگیری نظارت شده قرار می‌گیرد. می‌توان از آنها برای حل مسائل رگرسیون و طبقه‌بندی استفاده کرد. درخت تصمیم از نمایش درختی برای حل این مشکل استفاده می‌کند که در آن هر گره برگ با یک برچسب کلاس مطابقت دارد و ویژگی‌ها در گره داخلی درخت نشان داده می‌شوند. ما می‌توانیم هر تابع بولی را روی ویژگی‌های گسسته با استفاده از درخت تصمیم نمایش دهیم.

وقتی از یک گره در درخت تصمیم استفاده می‌کنیم تا نمونه‌های آموزشی را به زیرمجموعه‌های کوچک‌تر تقسیم کنیم، آنتروپی تغییر می‌کند. افزایش اطلاعات معیاری برای این تغییر در آنتروپی است.

الگوریتم K-NN: الگوریتم نزدیک‌ترین همسایگان یک روش غیرپارامتری برای طبقه‌بندی است که برای حل بسیاری از مسائل طبقه‌بندی استفاده می‌شود. رأی اکثریت همسایگان آن یک شیء را طبقه‌بندی می‌کند و شیء به کلاس رایج‌ترین در بین k نزدیک‌ترین همسایگان خود اختصاص داده می‌شود. بنابراین، این یک نوع یادگیری مبتنی بر نمونه است، که در آن تابع

فقط به صورت محلی تقریبی است و همه محاسبات تا طبقه بندی به تعویق می‌افتد. اغلب از یک نوع فازی از الگوریتم k-NN استفاده می‌شود [۸].

در آمار، الگوریتم k-نزدیک‌ترین همسایه (k-NN) یک روش یادگیری نظارت شده ناپارامتریک است که ابتدا توسط Evelyn Fix و Joseph Hodges در سال ۱۹۵۱ توسعه یافت، [۱] و بعداً توسط Thomas Cover گسترش یافت.[۲] برای طبقه بندی و رگرسیون استفاده می شود. در هر دو مورد، ورودی شامل k نزدیکترین مثال آموزشی در یک مجموعه داده است. خروجی بستگی به این دارد که از k-NN برای طبقه بندی یا رگرسیون استفاده شود:

در طبقه‌بندی k-NN، خروجی یک عضویت در کلاس است. یک شیء با رای کثرت همسایه‌هایش طبقه‌بندی می‌شود و شیء به کلاسی که در میان k نزدیک‌ترین همسایه‌هایش رایج‌تر است نسبت داده می‌شود (k یک عدد صحیح مثبت است، معمولاً کوچک). اگر k = 1، شیء به‌سادگی به کلاس آن نزدیک‌ترین همسایه اختصاص داده می‌شود.

در رگرسیون k-NN، خروجی مقدار ویژگی برای شیء است. این مقدار میانگین مقادیر k نزدیک‌ترین همسایه است.

k-NN نوعی طبقه‌بندی است که در آن تابع فقط به‌صورت محلی تقریبی می‌شود و تمام محاسبات تا ارزیابی تابع به تعویق می‌افتد. ازآنجایی‌که این الگوریتم برای طبقه‌بندی به فاصله متکی است، اگر ویژگی‌ها واحدهای فیزیکی متفاوتی را نشان دهند یا در مقیاس‌های بسیار متفاوتی باشند، عادی‌سازی داده‌های آموزشی می‌تواند دقت آن را به طور چشمگیری بهبود بخشد.[۳][۴] هم برای طبقه‌بندی و هم برای رگرسیون، یک تکنیک مفید می‌تواند تعیین وزن به سهم همسایگان باشد، به‌طوری‌که همسایه‌های نزدیک‌تر بیشتر از همسایگان دورتر به میانگین کمک می‌کنند. به‌عنوان‌مثال، یک طرح وزن‌دهی رایج شامل دادن وزن ۱/d به هر همسایه است که d فاصله تا همسایه است [۵]. همسایه‌ها از مجموعه‌ای از اشیا گرفته می‌شوند که کلاس (برای طبقه‌بندی k-NN) یا مقدار ویژگی شیء (برای رگرسیون k-NN) برای آنها شناخته شده است. این را می‌توان به‌عنوان مجموعه آموزشی برای الگوریتم در نظر گرفت، اگرچه هیچ مرحله آموزشی واضحی موردنیاز نیست.

همان‌طور که اندازه مجموعه‌داده‌های آموزشی به بی‌نهایت نزدیک می‌شود، طبقه‌بندی‌کننده نزدیک‌ترین همسایه نرخ خطای کمتر از دوبرابر نرخ خطای بیز (حداقل میزان خطای قابل‌دستیابی باتوجه‌به توزیع داده‌ها) را تضمین می کند.

ماشین‌های بردار پشتیبان: مدل‌های یادگیری تحت نظارت با الگوریتم‌های یادگیری مرتبط هستند. یعنی ماشین‌های بردار پشتیبان به مجموعه آموزشی نیاز دارد، مانند D در مورد ما. سپس، هر ورودی به یکی یا یکی از دودسته تعلق می‌گیرد و الگوریتم آموزش ماشین‌های بردار پشتیبان مدلی را ایجاد می‌کند که هر نمونه ورودی جدید را به یک دسته یا دسته دیگر اختصاص می‌دهد و آن را به یک‌طبقه بندی کننده خطی دوتایی غیراحتمالی تبدیل می‌کند [۹].

جایی که مقدار y برابر با 1 یا 1- و هر x_i برابر با یک مقدار حقیقی بعدی است. هدف پیدا کردن ابرصفحه جداکننده با بیشترین فاصله از نقاط حاشیه‌ای است.

شبکه‌های عصبی مصنوعی: شبکه‌های عصبی مصنوعی طبقه‌بندی‌کننده‌های بسیار غیرخطی هستند که کاربردهای زیادی در حوزه‌های گسترده دارند. ساختار آنها سعی می‌کند شبیه عملکرد مغز انسان با نورون‌ها و سیناپس‌ها باشد. به طور خاص، این شبکه‌ها شامل یک‌لایه ورودی است که سیگنال‌های ورودی را به‌عنوان داده دریافت می‌کند، یک یا چندلایه پنهان نورون که این داده‌ها را به روش غیرخطی پردازش می‌کند و یک لایه خروجی که نتیجه طبقه بندی نهایی را ارائه می‌دهد [۱۰].

شبکه‌های عصبی کانولوشنی: این شبکه‌ها از الگوهای یادگیری ماشین در ساختارهای عمیق استفاده می‌کنند. ابتدا مجموعه‌ای از ویژگی‌های مناسب را از داده‌های خام استخراج می‌کند، با استفاده از تحولات روی سیگنال‌های ورودی که آنها را به لایه‌های عمیق منتقل می‌کند، درحالی‌که در لایه آخر یک‌طبقه بندی برای اختصاص داده‌های ورودی به کلاس‌ها اما با استفاده از ویژگی‌های عمیق انجام می‌شود. توسط لایه‌های کانولوشن مشخص شده است [۱۱].

بیز ساده: طبقه‌بندی‌کننده‌های بیس خانواده‌ای از طبقه‌بندی‌کننده‌های احتمالی هستند که بر اساس به‌کارگیری قضیه بیز با مفروضات استقلال قوی بین ویژگی‌ها استفاده می‌شوند. این طبقه‌بندی‌کننده‌ها بسیار مقیاس‌پذیر هستند و به تعدادی پارامتر خطی در تعداد متغیرها (ویژگی‌ها پیش‌بینی‌کننده‌ها) در یک مشکل یادگیری نیاز دارند [۱۲].

۴-۳. الگوریتم‌های پردازش سیگنال

تبدیل فوریه گسسته[1]

در ریاضیات، تبدیل فوریه گسسته (DFT) یک دنباله محدود از نمونه‌های بافاصله مساوی از یک تابع را به دنباله‌ای با طول یکسان از نمونه‌های با فواصل مساوی تبدیل فوریه گسسته (DTFT) تبدیل می‌کند که یک مقدار مختلط است. تابع فرکانس فاصله زمانی که از DTFT نمونه‌برداری می‌شود، متقابل مدت‌زمان توالی ورودی است. یک DFT معکوس یک سری فوریه است که از نمونه‌های DTFT به‌عنوان ضرایب سینوسی پیچیده در فرکانس‌های DTFT مربوطه استفاده می‌کند. دارای مقادیر نمونه مشابه با دنباله ورودی اصلی است؛ بنابراین DFT یک نمایش دامنه فرکانس از توالی ورودی اصلی است. اگر دنباله اصلی تمام مقادیر غیرصفر یک تابع را در بر بگیرد، DTFT آن پیوسته (و دوره‌ای) است و DFT نمونه‌های گسسته یک‌چرخه را ارائه می‌دهد. اگر دنباله اصلی یک‌چرخه از یک تابع تناوبی باشد، DFT تمام مقادیر غیرصفر یک‌چرخه DTFT را ارائه می‌دهد.

DFT مهم‌ترین تبدیل گسسته است که برای انجام تحلیل فوریه در بسیاری از کاربردهای عملی استفاده می‌شود. در پردازش سیگنال دیجیتال، تابع هر مقدار یا سیگنالی است که در طول زمان تغییر می‌کند، مانند فشار موج صوتی، سیگنال رادیویی، یا خوانش دمای روزانه، نمونه‌برداری شده در یک بازه زمانی محدود (اغلب توسط یک تابع پنجره تعریف می‌شود.) در پردازش تصویر، نمونه‌ها می‌توانند مقادیر پیکسل‌ها در امتداد یک ردیف یا ستون یک تصویر شطرنجی باشند. DFT همچنین برای حل مؤثر معادلات دیفرانسیل جزئی و انجام عملیات‌های دیگر مانند کانولوشن یا ضرب اعداد صحیح بزرگ استفاده می‌شود.

ازآنجایی‌که با حجم محدودی از داده سروکار دارد، می‌توان آن را با الگوریتم‌های عددی یا حتی سخت‌افزار اختصاصی در رایانه‌ها پیاده‌سازی کرد. این پیاده‌سازی‌ها معمولاً از الگوریتم‌های تبدیل فوریه سریع (FFT) کارآمد استفاده می‌کنند؛ [۳] تا جایی که اصطلاحات «FFT» و «DFT» اغلب به جای یکدیگر استفاده می‌شوند. پیش از استفاده کنونی، ابتدائی سازی "FFT" ممکن است برای اصطلاح مبهم "تبدیل فوریه محدود" نیز استفاده شده باشد.

تجزیه های موجک

تجزیه موجک[2] جدیداً تکنیک‌های پردازش سیگنال چند مقیاسی اضافه شده است. بر خلاف اهرام گاوس و لاپلاس، آنها یک تصویر کامل ارائه می‌دهند و تجزیه را بر اساس مقیاس و جهت انجام می‌دهند. آنها با استفاده از بانک‌های فیلتر آبشاری که در آن فیلترهای پایین گذر و بالاگذر محدودیت‌های خاص خاصی را برآورده می‌کنند، اجرا می‌شوند. درحالی‌که مفاهیم پردازش سیگنال کلاسیک درک عملیاتی از چنین سیستم‌هایی را ارائه می‌دهند، ارتباطات قابل‌توجهی با کار در ریاضیات کاربردی و روان فیزیک وجود دارد که درک عمیق‌تری از تجزیه موجک و نقش آنها در بینایی ارائه می‌دهد. ازنقطه‌نظر ریاضی، تجزیه موجک معادل بسط سیگنال در یک موجک است. ویژگی‌های منظم و لحظه ناپدیدشدن فیلتر پایین گذر بر شکل توابع پایه تأثیر می‌گذارد و ازاین‌رو توانایی آن‌ها برای نمایش مؤثر تصاویر معمولی را دارد. از منظر روانی، مراحل اولیه پردازش اطلاعات بصری انسان ظاهراً شامل تجزیه تصاویر شبکیه به مجموعه‌ای از اجزای باند گذر مربوط به مقیاس‌ها و جهت‌گیری‌های مختلف است.

تولید و استخراج ویژگی

---

[1] Discrete Fourier Trans

[2] Wavelet decompositions

در یادگیری ماشین، تشخیص الگو و پردازش تصویر،استخراج ویژگی از مجموعه اولیه داده‌های اندازه‌گیری شروع می‌شود و مقادیر مشتق شده (ویژگی‌ها) را ایجاد می‌کند که آموزنده و غیرضروری است، مراحل یادگیری و تعمیم بعدی را تسهیل می‌کند و در برخی موارد منجر می‌شود به تفسیرهای بهتر انسانی استخراج ویژگی مربوط به کاهش ابعاد است. استخراج ویژگی شامل کاهش تعداد منابع موردنیاز برای توصیف مجموعه‌ای بزرگ از داده‌ها است. هنگام انجام تجزیه‌وتحلیل داده‌های پیچیده، یکی از مشکلات عمده ناشی از تعداد متغیرهای درگیر است. تجزیه‌وتحلیل با تعداد زیادی از متغیرها به‌طورکلی به مقدار زیادی حافظه و قدرت محاسباتی نیاز دارد، همچنین ممکن است باعث شود الگوریتم طبقه‌بندی برای آموزش نمونه‌ها مناسب باشد و به نمونه‌های جدید ضعیف شود. استخراج ویژگی یک اصطلاح کلی برای روش‌های ایجاد ترکیبی از متغیرها برای حل این مشکلات است درحالی‌که هنوز داده‌ها را بادقت کافی توصیف می‌کنید. بسیاری از تمرین‌کنندگان یادگیری ماشین معتقدند که استخراج بهینه ویژگی‌ها، کلید ایجاد مدل مؤثر است [۸].

روش‌های استخراج ویژگی، علاوه بر ویژگیها و بافت سیگنال تغییر یافته و بدون تغییر ، توصیفگرهای ساختاری و نمودار را شامل می شود. استخراج ویژگی با استخراج ویژگی ها از داده های ورودی ، دقت مدل های آموخته شده را افزایش می دهد. این مرحله از چارچوب کلی با حذف داده های اضافی ، ابعاد داده ها را کاهش می دهد. البته باعث افزایش آموزش و سرعت استنباط می شود. روش های استخراج ویژگی ها با انجام ترکیبات و تبدیل مجموعه ویژگی های اصلی ، ویژگی های جدید ایجاد شده را به دست می آورند [۹].

اصول[1]

تجزیه‌وتحلیل مؤلفه اصلی (PCA) یک تکنیک محبوب برای تجزیه‌وتحلیل مجموعه‌داده‌های بزرگ حاوی تعداد زیادی از ابعاد ویژگی‌ها در هر مشاهده، افزایش تفسیرپذیری داده‌ها درحالی‌که حداکثر مقدار اطلاعات را حفظ می‌کند، و امکان تجسم داده‌های چند بعدی را فراهم می کند. به طور رسمی، PCA یک تکنیک آماری برای کاهش ابعاد یک مجموعه داده است. این امر با تبدیل خطی داده ها به یک سیستم مختصات جدید انجام می شود که در آن (بیشتر) تغییرات در داده ها را می توان با ابعاد کمتری نسبت به داده های اولیه توصیف کرد. بسیاری از مطالعات از دو جزء اصلی اول برای ترسیم داده‌ها در دو بعد و شناسایی بصری خوشه‌هایی از نقاط داده نزدیک به هم استفاده می‌کنند. تجزیه و تحلیل مؤلفه های اصلی در بسیاری از زمینه ها مانند ژنتیک جمعیت، مطالعات میکروبیوم، علوم جوی و غیره کاربرد دارد.

انتخاب ویژگی

انتخاب ویژگی[2] یک رویکرد مهم برای کاهش ابعاد داده‌های با ابعاد بالا است. در سال‌های اخیر، الگوریتم‌های انتخاب ویژگی‌های زیادی پیشنهاد شده است. بااین‌حال، اکثر آنها فقط از اطلاعات موجود در فضای داده استفاده می‌کنند. آنها اغلب از اطلاعات مفید موجود در فضای ویژگی غافل می‌شوند و معمولاً از اطلاعات مربوط به هندسه زیرین داده‌ها سوءاستفاده نمی‌کنند [۱۰].

انتخاب ویژگی فرایندی است که در آن ویژگی‌ها به‌صورت خودکار یا دستی انتخاب می‌شوند و بیشترین نقش را در متغیر یا خروجی پیش‌بینی موردنظر دارند. وقوع ویژگی‌های اضافی یا نامربوط درداده‌های به‌دست‌آمده، دقت مدل‌ها را کاهش می‌دهد و باعث می‌شود مدل بر اساس ویژگی‌های نامربوط یاد بگیرد. بر اساس همبستگی متقابل از روش انتخاب ویژگی فیلتر استفاده می‌شود. هر دو روش پیچاندن و فیلتر مزایای خود را دارند و همچنین ضربه‌ها [۱۱].

[1] Principal component analysis

[2] Feature Selection

روش‌های اصلی انتخاب ویژگی

در کل سه نوع انتخاب ویژگی وجود دارد: روش‌های بسته‌بندی[1] (انتخاب جلو، عقب و گام‌به‌گام)، روش‌های فیلتر[2] (روش آنووا، همبستگی پیرسون، آستانه واریانس) و روش‌های جاسازی شده[3] (همانند درخت تصمیم)

روش‌های بسته‌بندی مدل‌ها را با زیر مجموعه خاصی از ویژگی‌ها محاسبه می‌کنند و اهمیت هر ویژگی را ارزیابی می‌کند. سپس آنها زیرمجموعه‌ای متفاوت از ویژگی‌ها را امتحان می‌کنند تا به زیرمجموعه بهینه برسند. دو اشکال این روش زمان محاسبه بزرگ‌داده‌ها با ویژگی‌های زیاد است و این که وقتی تعداد داده‌های زیادی وجود ندارد، به مدل برتری می‌بخشد [۱۲].

روش‌های فیلتر از معیاری غیر از میزان خطا برای تعیین مفیدبودن آن ویژگی استفاده می‌کنند. به‌جای تنظیم یک مدل (مانند روش‌های بسته‌بندی)، زیرمجموعه‌ای از ویژگی‌ها از طریق رتبه‌بندی آنها با یک روش توصیفی مفید انتخاب می‌شود. مزایای روش‌های فیلتر این است که زمان محاسبه بسیار پایینی دارند و بر داده‌ها بیش از حد مناسب نیستند. بااین‌حال، یک اشکال این است که آنها در برابر هر گونه تعامل یا ارتباط بین ویژگی‌ها کور هستند.

روش‌های جاسازی شده، انتخاب ویژگی را به‌عنوان بخشی از فرایند ایجاد مدل انجام می‌دهند. این امر به‌طورکلی منجر به ایجاد محیطی شاد بین دو روش انتخاب ویژگی می‌شود که قبلاً توضیح داده شد ، زیرا انتخاب همراه با فرایند تنظیم مدل انجام می شود [۱۳].

کشف دانش

کشف دانش[4] یک علم بین‌رشته‌ای است که هدف آن استخراج دانش مفید و کاربردی از مخازن داده های بسیار بزرگ است. به طور عمده ، با توجه به مجموعه داده ها ، یک فرایند کشف دانش در جستجوی موارد زیر است:

طبقه‌بندی‌کننده یک تصمیم‌گیرنده است که می‌تواند داده‌ها را به دسته‌های از پیش تعریف شده تقسیم کند که اغلب کلاس نامیده می‌شوند.

- پیش‌بینی: پیش‌بینی‌کننده یک تابع مناسب است که می‌تواند یک ویژگی هدف را با استفاده از داده‌های باقی‌مانده پیش‌بینی کند.
- خوشه‌بندی: خوشه‌بندی فرایندی است که بر اساس شباهت نقاط داده، داده‌ها را به دسته‌های ناشناخته‌ای تقسیم می‌کند که خوشه نامیده می‌شوند.
- الگوها: الگو یک قاعده قابل‌تشخیص درداده‌ها است که عناصر و یا ویژگی‌های آن در یک طرح قابل‌پیش‌بینی تکرار می‌شود.
- ناهنجاری‌ها: یک ناهنجاری که غالباً بیرونی نامیده می‌شود، اطلاعات غیرمنتظره‌ای است که به طور قابل‌توجهی از بقیه داده‌ها منحرف می‌شود.
- انجمن‌ها: ارتباط پیوند بین دو یا چند پدیده است که در قطعات اطلاعات کدگذاری شده است.
- مدل‌ها: مدل مجموعه‌ای از توابع ریاضی و یا منطقی است که می‌تواند توزیع و رفتار داده‌ها را توصیف کند [۱۴].

الگوریتم‌های خوشه‌بندی نیز برای شناسایی الگوهای معمول استفاده می‌شوند. الگوریتم‌های خوشه‌بندی بر اساس شباهت آماری بین هر یک از دونقطه، همه نقاط یک مجموعه‌داده را به چندین خوشه طبقه‌بندی می‌کنند. نقاط یک خوشه دارای ویژگی‌های آماری مشابه هستند و نقاط در خوشه‌های مختلف دارای ویژگی‌های آماری قابل‌توجهی متفاوت هستند. به‌طورکلی، شرایط مختلف عملکرد ویژگی‌های آماری متفاوتی دارند [۲].خوشه‌بندیی K-means یکی از محبوب ترین الگوریتم های خوشه بندی است. خوشه بندی c-means فازی نیز برای شناسایی الگوهای عملکرد به کار گرفته شد. الگوریتم های خوشه بندی دیگر نیز مانند خوشه بندی بردار پشتیبانی ، خوشه بندی حداکثر انتظارات و خوشه بندی درخت تصمیم گیری استفاده

---

[1] Wrapper method

[2] Filter methods

[3] Embeded method

[4] Knowledge Discovery

شـده اسـت و این اهمیت این الگوریتم‌ها را بیش از پیش مشـخص می‌سـازد. الگوریتم های اسـتخراج نمودار ، الگوریتم های استخراج متن و الگوریتم های قواعد انجمنی پویا نیز برای تشخیص خطای سیستم‌ها استفاده شده است [۲].

درخت الگوی مکرر (FP–growth)

درخت الگوی مکرر (FP-growth) یکی دیگر از الگوریتم‌های رایج قواعد انجمنی اسـت. الگوریتم رشـد FP برای یافتن مجموعه‌های مکرر در پایگاه‌داده تراکنشی مورداستفاده قرار می‌گیرد. رشد FP نشان‌دهنده موارد مکرر در درختان الگوی مکرر یا FP-tree است. به‌طورکلی، رشد FP در استخراج حجم عظیمی از داده‌ها بسیار مفید ظاهر می‌شود [۱۵-۱۶].

الگوریتم یادگیری با نظارت: روش طبقه‌بندی

الگوریتم‌های طبقه‌بندی می‌توانند رابطه پیچیده بین خطاها و علائم را بر اساس داده‌های جمع‌آوری‌شده در شرایط پیچیده و گوناگون بیاموزند. سپس می‌تواند تشخیص دهد که یک وضعیت جدید متعلق به کدام خطا است. دو نوع الگوریتم طبقه‌بندی استفاده شده است، یعنی الگوریتم‌های طبقه‌بندی چندطبقه و الگوریتم‌های طبقه‌بندی یک‌طبقه. ماشین بردار پشتیبانی[1] یکی از پرکاربردترین الگوریتم‌های طبقه‌بندی چندطبقه است. برخی از الگوریتم‌های پیش‌پردازش داده‌ها با ماشین بردار پشتیبانی ادغام شده اند تا کارایی الگوریتم‌های داده‌کاوی افزایش یابد در ادامه فرایند پیش‌پردازش داده‌ها معرفی شده است [۱۷].

ارزیابی مدل

ساده‌ترین روش اندازه‌گیری عملکرد مسائل طبقه‌بندی به‌ویژه هنگامی که خروجی شامل دو یا چند کلاس باشد، استفاده از روش ارزیابی مدل و ماتریس درهم‌ریختگی است. ماتریس درهم‌ریختگی چیزی شبیه یک جدول دوبعدی است. ارزش واقعی و ارزش پیش‌بینی‌شده همان‌طور که در شکل زیر نشان‌داده‌شده است، هر دو بعد مثبت - صحیح (TP)، منفی - صحیح (TN)، مثبت - غلط (FP) و منفی - غلط (FN) هستند [۱۵].

دقت و صحت مدل

دقت[2] و صحت[3] متداول‌ترین الگوریتم‌های کلاس‌بندی هستند که در قالب پیش‌بینی‌های درست تعریف می‌شوند. در واقع دقت درستی پیش‌بینی‌ها را در یک نسبت موارد صحیح به کل موارد درست نشان می‌دهد. در شاخص صحت نیز نسبت مقدار موارد صحیح در کلاس‌ها به کل اعضای پیش‌بینی شده در آن گروه محاسبه می‌شود.

امتیاز F

این امتیاز دقت و صحت را با هم مدنظر قرار می‌دهد. به بیان ریاضی، امتیاز F1 میانگین وزن‌دار از دقت و صحت است. بهترین مقدار برای F1 مقدار یک و بدترین مقدار صفر است

استخراج قواعد انجمنی

در استخراج قواعد انجمنی ارتباط بین متغیرها در میان حجم عظیمی از داده‌های عملیات بسیار بررسی می‌شود. قاعده ارتباط معمولاً به شکل "$A \rightarrow B$" نشان داده می‌شود، جایی که A مقدم و B نتیجه آن است. الگوریتم Apriori یکی از رایج‌ترین الگوریتم‌های قواعد انجمنی برای شناسایی الگوهای معمول است.

---

[1] Support vector machine

[2] Precision

[3] Recall

از دیگر الگوریتم‌های استخراج قوانین مرتبط مانند قواعد انجمنی وزنی، قواعد انجمنی کمی و قواعد انجمنی زمانی استفاده شده است. در مقایسه با الگوریتم‌های معمول قواعد انجمنی، الگوریتم قواعد انجمنی کمی می‌تواند داده‌های عددی و داده‌های دسته‌ای را بدون تشخیص داده‌ها استخراج کند [۱۶].

اخیراً، محققان دریافتند که الگوریتم استخراج نمودار، یعنی تنوع قواعد انجمنی، در استخراج پایگاه‌های داده چند رابطه‌ای بیشتر از الگوریتم‌های معمول قواعد انجمنی مؤثر است. به‌عنوان‌مثال، فن و همکاران یک روش مبتنی بر استخراج نمودار برای نشان‌دادن الگوهای عملکرد معمولی سیستم‌های HVAC پیشنهاد کرد. نمودارها قادر به توصیف دانش به‌صورت تصویری هستند؛ بنابراین، روش‌های مبتنی بر معدن گراف می‌تواند تفسیرپذیری دانش استخراج شده را بهبود بخشد [۱۷].

بسیاری از الگوریتم‌های رگرسیونی برای پیش‌بینی موفقیت‌آمیز شبکه عصبی مصنوعی، رگرسیون بردار پشتیبان (SVR)، میانگین متحرک خودگردان (ARIMA)، شبکه عصبی عمیق (DNN) استفاده شده است.) و غیره به کار می‌روند. به طور به‌طورکلی چهار مرحله است، یعنی تبدیل داده‌ها، انتخاب ویژگی، بهینه‌سازی پارامترهای مدل و مدل آموزش. در مرحله تبدیل داده‌ها، داده‌های عملیات خام تاریخی به‌منظور افزایش دقت مدل پیش‌بینی به یک مقیاس نرمال تبدیل می‌شوند. مرحله استخراج ویژگی در استخراج مرتبط‌ترین متغیرهای مؤثر بر بار انرژی هدف است. سپس از ویژگی‌های استخراج شده برای آموزش مدل استفاده می‌شود. مرحله بهینه‌سازی پارامترهای مدل بهینه‌سازی پارامترهای فوق‌العاده مدل برای به‌دست‌آوردن ساختار مدل بهینه است [۱۸].

رگرسیون خطی

از جمله فنون بررسی ارتباط میان متغیرهای مستقل و وابسته است که با یک متغیر مستقل و یک متغیر وابسته ثابت ادامه می‌یابد.

معیار نیم‌رخ[1]:

یکی دیگر از روش‌های ارزیابی خوشه‌بندی، معیار «نیم‌رخ» است. این معیار هم به پیوستگی[2] درون خوشه‌ها و هم به میزان تفکیک‌پذیری آن‌ها بستگی دارد. مقدار نیم‌رخ برای هر نقطه، میزان تعلق آن را به خوشه‌اش در مقایسه با خوشه مجاور اندازه می‌گیرد. در واقع الگوریتم نیم‌رخ از اطلاعات معیار مفید دیگری برای ارزیابی طبیعی تعداد خوشه‌هاست [۱۹].

کاهش داده

کاهش داده[3] عبارت از تبدیل اطلاعات دیجیتالی عددی یا الفبایی به‌صورت تجربی یا تجربی به یک فرم تصحیح شده، مرتب و ساده شده است. هدف از کاهش داده‌ها می‌تواند دوگانه باشد: کاهش تعداد پرونده‌های داده با حذف داده‌های نامعتبر یا تولید خلاصه داده‌ها و آمار در سطوح مختلف تجمیع برای برنامه‌های مختلف. کاهش داده یا تکنیک‌های کاهش متغیر، به‌سادگی به فرایند کاهش تعداد یا ابعاد ویژگی‌ها در یک مجموعه‌داده اشاره دارد. معمولاً در هنگام تجزیه‌وتحلیل داده‌های با ابعاد بالا (به‌عنوان‌مثال، تصاویر چند پیکسلی از صورت یا متون مقاله، فهرست‌های نجومی و غیره) استفاده می‌شود. بسیاری از روش‌های آماری و یادگیری ماشین برای داده‌های با ابعاد بالا استفاده شده است، مانند مدل برداری و مخلوط برداری، نقشه‌برداری توپوگرافی مولد، کاهش ابعاد نقش مهمی در عملکرد طبقه‌بندی دارد. یک سیستم تشخیص با استفاده از مجموعه‌ای محدود از ورودی‌ها طراحی شده است. درحالی‌که اگر این ویژگی‌های اضافی را اضافه کنیم، عملکرد این سیستم افزایش می‌یابد، اما در برخی موارد یک گنجاندن بیشتر منجر به کاهش عملکرد می‌شود؛ بنابراین کاهش ابعاد ممکن است همیشه یک سیستم طبقه‌بندی را بهبود ندهد [۲۲].کاهش داده‌ها نقش مهمی در عملکرد طبقه بندی دارد. یک سیستم تشخیص با استفاده از مجموعه ای محدود از ورودی‌ها طراحی شده است. در حالی که اگر این ویژگی‌های اضافی را اضافه کنیم، عملکرد این سیستم

---

[1] Silhouette

[2] Cohesion

[3] Data Reduction

افزایش می یابد، اما در برخی موارد یک گنجاندن بیشتر منجر به کاهش عملکرد می شود. بنابراین کاهش ابعاد ممکن است همیشه یک سیستم طبقه بندی را بهبود ندهد.

تبدیل داده‌ها

در علوم رایانه، تبدیل داده‌ها[1] فرایند تغییر قالب، ساختار یا مقادیر داده است. برای پروژه‌های تجزیه‌وتحلیل داده‌ها، داده‌ها ممکن است در دو مرحله از خط لوله داده تبدیل شوند. فرایندهایی مانند یکپارچه‌سازی داده‌ها، انتقال داده‌ها، ذخیره‌سازی داده‌ها و کشمکش داده‌ها همه ممکن است شامل تغییر داده‌ها باشد. روش‌های تبدیل داده ایجاد شده توسط پیشینیان عمدتاً از نظر آماری است که با شرایط غیرطبیعی توالی سروکار دارد. بااین‌حال، نظریه‌های محدود ریاضی یا آماری نمی‌توانند ویژگی‌های اساسی داده‌ها را کاملاً توضیح دهند. در سال‌های اخیر، ترکیب بین‌رشته‌ای سیستم‌های پیچیده به یک موضوع داغ تبدیل شده است. در همین حال، برخی از نظریه‌ها در فیزیک نیز نقش مهمی در زمینه‌های اقتصادی و مالی ایفا می‌کنند [۲۳].

در بررسی روش‌های مختلف مورداستفاده در حوزه سیستم‌های توزیع‌شده‌ی بیگ دیتا با الگوریتم‌های هوش مصنوعی، مشاهده می‌شود که بسیاری از رویکردها بهبودهای قابل‌توجهی را در زمینه‌های خاص ارائه می‌دهند، اما اغلب نقاط ضعف مهمی دارند. بعضی از روش‌ها، مانند شبکه‌های عصبی عمیق، نیازمند داده‌های بسیار زیاد و زمان آموزش طولانی هستند که در محیط‌های زمان واقعی محدودیت‌هایی ایجاد می‌کند. سایر رویکردها، مانند روش‌های مبتنی بر قوانین، ساده‌تر و سریع‌تر هستند اما نمی‌توانند پیچیدگی‌های دینامیک سیستم‌های توزیع‌شده را به‌اندازه کافی مدل‌سازی کنند. مقایسه میان این روش‌ها نشان می‌دهد که هیچ یک به طور کامل برتری مطلق ندارد و انتخاب روش مناسب، وابسته به نیازهای خاص پروژه و محدودیت‌های عملی است. علاوه بر این، غالباً مطالعات بر روی معیارهای عملکرد محدود تمرکز دارند، درحالی‌که جنبه‌هایی مانند مقیاس‌پذیری و قابلیت اطمینان و امنیت کمتر مورد ارزیابی قرار گرفته است.

در حوزه پردازش بیگ دیتا و سیستم‌های توزیع‌شده، دیتاست‌های متعددی برای آزمایش و ارزیابی روش‌ها استفاده شده است، اما اغلب این دیتاست‌ها به‌صورت کامل معرفی نمی‌شوند و در برخی پژوهش‌ها، جزئیات قابل‌قبولی ارائه نمی‌گردد. نمونه‌هایی مانند مجموعه‌داده‌های شبیه‌سازی شده برای ارزیابی کارایی سیستم، داده‌های جمع‌آوری‌شده از شبکه‌های سنجش ابری، و مجموعه‌داده‌های واقعی مانند موارد حوزه اینترنت اشیا و داده‌های ثبت‌شده در مراکز داده‌ها از جمله موارد رایج هستند. شناخت ویژگی‌های هر دیتاست، شامل حجم، نوع داده، مدت‌زمان جمع‌آوری، و مشخصات کیفیت داده‌ها، اهمیت زیادی در تفسیر نتایج دارد که متأسفانه بیشتر منابع به این نکات پرداخته نشده است. پارامترهای مورداستفاده در جدول ۲ قابل مشاهده هستند.

پارامتر	کاربرد - اهمیت	شرح - توضیحات
(Efficiency) کارایی	ارزیابی بهره‌وری کلی سیستم	میزان بهره‌وری منابع در اجرای وظایف
(Energy Consumption) مصرف انرژی	مهم برای سیستم‌های سبز و پایدار	مقدار انرژی مصرف‌شده توسط سیستم در حین عملیات
(Response Time) زمان پاسخ	معیار مهم در سیستم‌های زمان‌حساس	مدت‌زمان لازم برای پاسخگویی سیستم به درخواست‌ها
(Error Rate) نرخ خطا	نشان‌دهنده دقت و صحت عملکرد سیستم	نسبت خطاهای رخ‌داده در نتایج سیستم
(Scalability) مقیاس‌پذیری	ارزیابی قابلیت رشد و توسعه سیستم	توانایی سیستم در مدیریت حجم‌های بزرگ‌داده و کاربران بدون افت کارایی
(Stability) پایداری	مهم در سیستم‌های بلندمدت و هوشمند	ثبات عملکرد سیستم در مواجهه با تغییرات یا حجم‌های متغیر داده‌ها

[1] Data Transformation

(Flexibility) انعطاف‌پذیری	برای سیستم‌های چندمنظوره و آینده‌پذیر	توانایی سیستم در تطابق با نیازهای مختلف و تغییر شرایط
(Interoperability) تعامل‌پذیری	اهمیت در ساختارهای توزیع‌شده و چندسازه‌ای	سازگاری و یکپارچگی سیستم با سایر سامانه‌ها و پلتفرم‌ها
(Training Speed) سرعت آموزش مدل‌ها	مهم در توسعه سریع مدل‌های دینامیک	مدت‌زمان موردنیاز برای آموزش الگوریتم‌های یادگیری ماشین
(Update Rate) نرخ به‌روزرسانی	برای سیستم‌های پویا و در حال تغییر	سرعت به‌روزرسانی مدل‌ها و داده‌های سیستم
(Resource Utilization) بهره‌وری منابع	برای بهبود بهره‌وری کلی سیستم	میزان استفاده مؤثر از CPU، حافظه و سایر منابع سیستم
(Reliability) قابلیت اطمینان	مهم در سیستم‌های حساس و بحرانی	درصد صحت و عملکرد بدون خطا در عملیات سیستم
(Reproducibility) قابلیت تکرار	برای اعتبارسنجی پژوهش و توسعه مدل‌ها	توانایی تکرار نتایج در آزمایش‌های مختلف
(Security) امنیت	حیاتی برای حفاظت داده‌های حساس	میزان ایمنی سیستم در مقابل تهدیدات و حملات
(Operational Cost) هزینه اجرا	برای ارزیابی هزینه – فایده راهکارها	هزینه‌های مرتبط با عملیات و نگهداری سیستم
(Usability) کاربری	مهم در پذیرش و کاربرد عملی سیستم	میزان سهولت استفاده از سیستم توسط کاربران
(Parallel Processing Power) توان پردازش موازی	برای سیستم‌های مقیاس‌پذیر توزیع‌شده	ظرفیت و سرعت انجام عملیات موازی
(Peak Resource Consumption) مصرف منابع در حالت پیک	برای مدیریت و برنامه‌ریزی منابع	حداکثر منابع مصرف‌شده در شرایط اوج فعالیت
(Load Balancing) توازن بار	برای کارایی و پایدارسازی سیستم	توزیع درست و متعادل وظایف و داده‌ها میان سرورها و منابع
(Real-world Performance) ارزیابی کارایی در محیط‌های واقعی	برای اطمینان از کاربردپذیری نتایج در دنیا واقعی	عملکرد سیستم در محیط‌های عملیاتی و واقعی

بر اساس داده‌های به‌دست‌آمده از جدول ۲، معیارهای ارزیابی، نقش کلیدی در سنجش اثربخشی و کارایی هر روش دارند. در مطالعه حاضر، پارامترهایی مانند کارایی(Efficiency) ، مصرف انرژی(Energy Consumption) ، زمان پاسخ (Response Time)، نرخ خطا (Error Rate) و مقیاس‌پذیری (Scalability) مورد اشاره قرار گرفته است. اما تحلیل عمیق درباره کاربرد و اهمیت هر پارامتر، مقایسه میان مدل‌ها بر اساس این معیارها و بررسی تأثیر هر پارامتر بر نتایج کلی، در مقاله مشاهده نمی‌شود. لازم است که در ادامه، این پارامترها با جزئیات بیشتری بررسی شوند تا بتوانیم ارزیابی دقیق‌تری از مدل‌ها و روش‌های پیاده‌سازی‌شده داشته باشیم.

در این نوآوری، سیستم هوشمندی توسعه‌یافته است که به‌صورت خودتنظیم و مبتنی بر یادگیری عمیق، منابع سیستم‌های توزیع‌شده برای پردازش بیگ دیتا را به‌صورت دینامیک و هوشمند مدیریت می‌کند. این سیستم با جمع‌آوری داده‌های لحظه‌ای از سرورها، شبکه و زیرساخت‌های در حال اجرا، الگوهای مصرف منابع را تحلیل و پیش‌بینی می‌کند و بر اساس این پیش‌بینی‌ها، تصمیم می‌گیرد که چگونه منابع را تخصیص دهد، وظایف را توزیع کند و بار کاری را تعادل بخشد. در این راستا، از شبکه‌های عصبی عمیق نظیر LSTM یا مدل‌های ترنسفورمر برای آموزش مدل‌های پیش‌بینی بهره می‌برند تا دقت این پیش‌بینی‌ها به حداکثر برسد. سیستم تصمیم‌گیری خودکار، قادر است به‌صورت پیوسته وارد عمل شود و واکنش سریع به

تغییرات ناگهانی در میزان بار کاری نشان دهد، بدون نیاز به دخالت مستقیم انسان. این رویکرد نه‌تنها موجب بهبود بهره‌وری و کاهش مصرف انرژی می‌شود، بلکه فرایند مدیریت منابع را بسیار انعطاف‌پذیرتر و مقیاس‌پذیرتر می‌سازد. یکی از چالش‌های اصلی این سیستم، حفظ دقت و امنیت داده‌ها است که با به‌کارگیری روش‌های پیشرفته رمزگذاری و حفاظت از حریم خصوصی قابل‌حل است. این نوآوری می‌تواند به‌صورت مستقیم در مراکز داده‌های ابری بزرگ و سیستم‌های اینترنت اشیا در شهرهای هوشمند پیاده‌سازی و بهینه‌سازی شود، و نقش مهمی در توسعه زیرساخت‌های هوشمند و کارآمد ایفا کند.

۱۰. نتیجه‌گیری

مقاله حاضر به بررسی طراحی و پیاده‌سازی سیستم‌های توزیع شده برای پردازش بیگ دیتا با استفاده از الگوریتم‌های هوش مصنوعی پرداخته است. ازآنجایی‌که حجم داده‌ها در دنیای امروزی روبه‌رشد است، استفاده از سیستم‌های توزیع شده و الگوریتم‌های هوش مصنوعی برای پردازش این داده‌ها اهمیت بیشتری پیدا کرده است. نتایج به‌دست‌آمده از این مقاله نشان می‌دهد که استفاده از سیستم‌های توزیع شده و الگوریتم‌های هوش مصنوعی می‌تواند بهبود قابل‌توجهی در عملکرد پردازش بیگ دیتا داشته باشد. این روش‌ها امکان پردازش سریع‌تر و بهینه‌تر داده‌ها را فراهم می‌کنند و از مزایای قابل‌توجهی برای سازمان‌ها و شرکت‌ها به دنبال دارند.

بنابراین، از این مقاله می‌توان نتیجه گرفت که استفاده از سیستم‌های توزیع شده و الگوریتم‌های هوش مصنوعی برای پردازش بیگ دیتا می‌تواند بهبود قابل‌توجهی در عملکرد و کارایی سیستم‌ها و برنامه‌های مختلف ایجاد کند. در نتیجه، سیستم‌های توزیع شده یک رویکرد همه‌کاره و انعطاف‌پذیر برای محاسبات ارائه می‌دهند. این سیستم‌ها با استفاده از توان جمعی گره‌های متعدد، می‌توانند وظایف پیچیده را به‌راحتی انجام دهند. درحالی‌که چالش‌هایی مانند سازگاری داده‌ها ممکن است ایجاد شود، مزایای مقیاس‌پذیری و تحمل خطا، سیستم‌های توزیع‌شده را به ابزاری ارزشمند در چشم‌انداز دیجیتال امروزی تبدیل می‌کند.

منابع

[1] Aminizadeh, S., Heidari, A., Toumaj, S., Darbandi, M., Navimipour, N. J., Rezaei, M., ... & Unal, M. (2023). The applications of machine learning techniques in medical data processing based on distributed computing and the Internet of Things. *Computer methods and programs in biomedicine*, 107745.

[2] Al-Jumaili, A. H. A., Muniyandi, R. C., Hasan, M. K., Paw, J. K. S., & Singh, M. J. (2023). Big data analytics using cloud computing based frameworks for power management systems: Status, constraints, and future recommendations. *Sensors*, *23*(6), 2952.

[3] Khang, A., Gupta, S. K., Rani, S., & Karras, D. A. (Eds.). (2023). *Smart Cities: IoT Technologies, big data solutions, cloud platforms, and cybersecurity techniques*. CRC Press.

[4] Manikandan, N., Tadiboina, S. N., Khan, M. S., Singh, R., & Gupta, K. K. (2023, May). Automation of Smart Home for the Wellbeing of Elders Using Empirical Big Data Analysis. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 1164-1168). IEEE.

[6] Hong, S. C. T.-L., S. D'Oca, D. Yan, S. P. (2016). Advances in research and applications of energy-related occupant behavior in buildings. Electronic Library, 116, 694-704.

[7] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas. Learning where to attend with deep architectures for image tracking. Neural computation, 24(8):2151–2184, 2012

[8] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

[9] Chunduri, R. K., & Cherukuri, A. K. (2021). Scalable algorithm for generation of attribute implication base using FP-growth and spark. Soft Computing, 1-22.

[10] D'Oca, S., Chen, C. F., Hong, T., & Belafi, Z. . (2017). Synthesizing building physics with social psychology: An interdisciplinary framework for context and occupant behavior in office buildings. Energy research & social science, 34, 240-251.
[11] Fan, S. X., F. (2018). Mining big building operational data for improving building energy efficiency: a case study. Build. Serv. Eng. Res. Technol, 39, 117-128.
[12] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
[13] Laender, A. H., Ribeiro-Neto, B. A., Da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. ACM Sigmod Record, 31(2), 84-93.
Loshin, D. (2013). Business Intelligence (Second Edition):
[14] Morgan Kaufmann Mirmozaffari, M., Boskabadi, A., Azeem, G., Massah, R., Boskabadi, E., Dolatsara, H. A., & Liravian, A. (2020). Machine learning clustering algorithms based on the DEA optimization approach for banking system in developing countries. European Journal of Engineering and Technology Research, 5(6), 651-658.
[15] Nabilah, A., Devita, H. P., Van Halen, Y., & Jurizat, A. (2021). Energy Efficiency in Church Building Based on Sefaira Energy Use Intensity Standard. Paper presented at the IOP Conference Series: Earth and Environmental Science.
[16] Poelmans, J., Dedene, G., Verheyden, G., Van der Mussele, H., Viaene, S., & Peters, E. (2010). Combining business process and data discovery techniques for analyzing and improving integrated care pathways. Paper presented at the Industrial Conference on Data Mining.
[17] Qamar Shahbaz Ul Haq. (2016). Data Mapping for Data Warehouse Design: Morgan Kaufmann
[18] Qiu, F. F., Z. Li, G. Yang, P. Xu, Z. Li. (2019). Data mining based framework to identify rule based operation strategies for buildings with power metering system. Build. Simul, 12, 195-205.
[14] Salvador García, J. L., Francisco Herrera. (2014). Data Preprocessing in Data Mining: Springe
[15] Sherman, R. (2015). Business Intelligence Guidebook: Morgan Kaufmann.
Zhang. (2015). A New Data Transformation Method and Its Empirical Research Based on Inverted Cycloidal Kinetic Model. Procedia Computer Science, 55, 485-492.
[16] D. Held, S. Thrun, and S. Savarese. Learning to track at 100 fps with deep regression networks. arXiv preprint arXiv:1604.01802, 2016.
[17] Vatter, J., Mayer, R., & Jacobsen, H. A. (2023). The evolution of distributed systems for graph neural networks and their origin in graph processing and deep learning: A survey. *ACM Computing Surveys*, *56*(1), 1-37.

[18] S. Hong, T. You, S. Kwak, and B. Han. Online tracking by learning discriminative saliency map with convolutional neural network. arXiv preprint arXiv:1502.06796, 2015.
[19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. arXiv preprint arXiv:1603.08155, 2016.
[20] S. E. Kahou, V. Michalski, and R. Memisevic. Ratm: Recurrent attentive tracking model. arXiv preprint arXiv:1510.08660, 2015.
[21] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder. The visual object tracking vot2015 challenge results. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 1–23, 2015.
[22] Olaniyi, O., Okunleye, O. J., & Olabanji, S. O. (2023). Advancing data-driven decision-making in smart cities through big data analytics: A comprehensive review of existing literature. *Current Journal of Applied Science and Technology*, *42*(25), 10-18.

[23] Himeur, Y., Elnour, M., Fadli, F., Meskin, N., Petri, I., Rezgui, Y., ... & Amira, A. (2023). AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. *Artificial Intelligence Review*, *56*(6), 4929-5021.