University of Qom

Sciences and Techniques of Information Management

Volume 11, Issue 2, 2025

A Method for Information Retrieval from Missing Data Using Data Mining Techniques and Genetic Algorithm

Mohammad Moradi
Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Bozorgmehr University of Qaenat, Qaenat, Iran **(Corresponding author)**. m_moradi@buqaen.ac.ir

Mojtaba Mazoochi
Assistant Professor, ICT Research Institute (ITRC), Tehran, Iran. mazoochi@itrc.ac.ir

Abstract

Purpose: In statistical literature, various terms—often used interchangeably—refer to the concept of missing data. These include missing data, lost data, incomplete data, nonresponse data, and others. In statistics, missing data or missing values occur when no data values are recorded for a variable in a given observation. Data are often lost in economic, sociological, and political science research because government or private entities may provide incomplete reports, some study participants may withdraw from participation or avoid answering certain questions, or researchers, technicians, and data collectors may make errors that result in data loss. Missing data can disrupt the distribution of variables, potentially causing model overfitting or underfitting. They can also introduce bias into a dataset, thereby skewing statistical analyses toward biased results and making it difficult to draw meaningful conclusions from the collected data. Moreover, they can lead to incorrect model analysis. Traditionally, the most common method for addressing missing data was simply to remove them, which resulted in low-quality datasets and consequently biased analyses and findings. Today, with scientific advances in various fields and the emergence of powerful statistical methods, missing values in incomplete datasets can be appropriately imputed or estimated prior to modeling. Given the importance of managing missing data, the present study aims to propose a method for improving the accuracy of information and knowledge retrieval from missing data.

Method: The proposed method employs data mining techniques, including clustering and regression, as well as heuristic algorithms such as genetic algorithms. In existing methods, the entire dataset is used to impute missing values. This approach often includes records that are dissimilar to the one with missing data, leading to inaccurate results. In the proposed algorithm, clustering is used to identify similar records. Then, for each cluster, the proportion of missing data for each attribute (column) is calculated. Based on this proportion, either a regression model or a genetic algorithm is applied to recover the missing data.

Findings: The implementation of the proposed method on a dataset with randomly missing data showed that the error rate of the algorithm was 27%. This rate was significantly lower than those of other methods: mean, median, and mode substitution methods (56.5%), the regression method

Cite this article: Moradi, M., Mazoochi, M. (2025). A Method for Information Retrieval from Missing Data Using Data Mining Techniques and Genetic Algorithm. *Sciences and Techniques of Information Management*, 11(2), 173-196. https://doi.org/10.22091/stim.2024.10668.2092

Received: 2024-04-27 ; **Revised:** 2024-06-21 ; **Accepted:** 2024-07-16 ; **Published online:** 2024-07-30

Article type: Research Article **Published by:** University of Qom.

(34.6%), and the support vector machine (SVM) method (42.1%). These results demonstrate higher accuracy in imputing missing data.

Conclusion: Existing methods use the entire dataset to replace missing values, which often leads to the inclusion of dissimilar records and consequently produces inaccurate results. The proposed algorithm addresses this issue by employing clustering to identify similar records and estimate missing data based on records within the same cluster. Additionally, the algorithm incorporates outlier removal, determination of the optimal number of clusters, and other refinements to ensure that abnormal data do not influence the estimation of missing values. Attributes (columns) with more than one-third missing data are removed to prevent unreliable data from affecting the estimation process. Furthermore, regression models within clusters consider related attributes when estimating missing values. The integration of a genetic algorithm, which combines mean, median, mode, and regression models, results in more reliable and accurate outcomes.

Keywords: Keywords: Information recovery, Missing Data, Data Mining, Genetic Algorithm, Clustering, Regression Model.

علوم و فنون مدیریت اطلاعات
دوره ۱۱، شماره ۲، ۱۴۰۴

روشی برای بازیابی اطلاعات از داده‌های گم‌شده با استفاده از تکنیک‌های داده‌کاوی و الگوریتم ژنتیک

محمد مرادی
استادیار گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه بزرگمهر قائنات، قائنات، ایران (**نویسنده مسئول**).
m_moradi@buqaen.ac.ir
مجتبی مازوچی
استادیار، پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران. mazoochi@itrc.ac.ir

چکیده

هدف: در ادبیات آماری، اصطلاحات مختلف و غالباً مترادفی برای مفهوم داده‌های گم‌شده وجود دارد. این اصطلاحات عبارتند از داده‌های گم‌شده، داده‌های از دست رفته، داده‌های ناقص، و داده‌های بی‌پاسخ. در آمار، داده‌های گم‌شده یا مقدارهای گم‌شده زمانی رخ می‌دهند که هیچ مقدار داده‌ای برای یک متغیر در یک مشاهده ذخیره نشده باشد. داده‌ها اغلب در تحقیقات اقتصادی، جامعه‌شناسی، و علوم سیاسی از بین می‌روند، زیرا دولت یا نهادهای خصوصی ممکن است گزارش‌های حساس را ناقص ارائه دهند، یا ممکن است برخی از افراد شرکت‌کننده در مطالعه از ادامهٔ همکاری انصراف دهند، یا از پاسخ دادن به برخی از سؤالات اجتناب کنند، یا محققین، تکنسین‌ها، و جمع‌آوری‌کننده‌های داده‌ها ممکن است اشتباهاتی را انجام دهند که منجر به گم‌شدن داده‌ها شود. داده‌های گم‌شده می‌توانند باعث ایجاد اغتشاش در توزیع متغیر شوند، یعنی می‌توانند باعث بیش‌برازش یا کم‌برازش مدل‌ها شوند. داده‌های گم‌شده می‌توانند باعث یک سوگیری (اریبی) در مجموعه داده شوند و تجزیه و تحلیل آماری را به‌سوی نتایج اریب سوق داده و نهایتاً دستیابی به یک نتیجه‌گیری مفید از داده‌های جمع‌آوری شده را با مشکل مواجه سازند و می‌توانند منجر به تجزیه‌وتحلیل نادرست مدل شوند. پیش از این، برای غلبه بر مشکل داده‌های گم‌شده مرسوم‌ترین روش، حذف داده‌های گم‌شده بود که منجر به داده‌هایی با کیفیت پایین و به تبع آن تحلیل و استخراج نتایج دارای سوگیری می‌شد. امروزه با پیشرفت‌های علمی در حوزه‌های گوناگون و پیدایش روش‌های توانمند آماری، می‌توان پیش از مدل‌سازی داده‌های ناکامل، مقادیر گم‌شده را با مقادیر مناسب جای‌گذاری یا برآورد کرد. با توجه به اهمیت گفته شدهٔ در موضوع مواجهه و مدیریت داده‌های گم‌شده، پژوهش حاضر با هدف ارائهٔ روشی برای بهبود دقت بازیابی اطلاعات و دانش از داده‌های گم‌شده انجام شده است.

روش: در روش پیشنهادی از تکنیک‌های داده‌کاوی شامل خوشه‌بندی و رگرسیون، و همچنین از الگوریتم‌های مکاشفه‌ای شامل الگوریتم ژنتیک استفاده شده است. در روش‌های موجود، برای جایگزینی دادهٔ از دست رفته، از کل

استناد به این مقاله: مرادی، م.، و مازوچی، م. (۱۴۰۴). روشی برای بازیابی اطلاعات از داده‌های گم‌شده با استفاده از تکنیک‌های داده‌کاوی و الگوریتم ژنتیک. *علوم و فنون مدیریت اطلاعات*، ۱۱(۲)، ۱۷۳-۱۹۶. https://doi.org/10.22091/stim.2024.10668.2092

تاریخ دریافت: ۱۴۰۳/۰۲/۰۸ ؛ **تاریخ اصلاح:** ۱۴۰۳/۰۴/۰۱ ؛ **تاریخ پذیرش:** ۱۴۰۳/۰۴/۲۶ ؛ **تاریخ انتشار آنلاین:** ۱۴۰۳/۰۵/۰۹

نوع مقاله: پژوهشی **ناشر:** دانشگاه قم

مجموعه داده استفاده می‌شود. این موضوع سبب در نظر گرفتن رکوردهای غیر مشابه رکورد مربوط به دادهٔ از دست رفته خواهد شد. لذا منجر به نتایج اشتباه خواهد شد. در الگوریتم پیشنهادی، از خوشه‌بندی برای شناسایی رکوردهای مشابه استفاده شده است. سپس، برای هر خوشه، میزان داده‌های گم‌شدهٔ هر صفت (ستون) از مجموعه داده محاسبه شده است. بر اساس میزان دادهٔ از دست رفته، از مدل رگرسیون یا از الگوریتم ژنتیک برای بازیابی اطلاعات از دست رفته استفاده شده است.

یافته‌ها: نتایج پیاده‌سازی روش پیشنهادی بر روی یک مجموعه داده که حاوی داده‌های گم‌شده به صورت تصادفی بودند نشان داد میزان خطای الگوریتم پیشنهادی برابر ۲۷ درصد است که نسبت به روش استفاده از میانگین، میانه؛ و مد که دارای خطای ۵۶٫۵ درصد، و روش استفاده از رگرسیون که دارای خطای ۳۴٫۶ درصد، و روش ماشین بردار پشتیبان (SVM) که دارای خطای ۴۲٫۱ درصد بود، دقت بالاتری در جانهی داده‌های گم‌شده داشته است.

نتیجه‌گیری: در روش‌های موجود، برای جایگزینی دادهٔ از دست رفته، از کل مجموعه داده استفاده می‌شود. این موضوع سبب در نظر گرفتن رکوردهای غیر مشابه رکورد مربوط به دادهٔ از دست رفته خواهد شد. لذا منجر به نتایج اشتباه خواهد شد. در الگوریتم پیشنهادی، از خوشه‌بندی برای شناسایی رکوردهای مشابه، و محاسبهٔ دادهٔ از دست رفته بر اساس رکوردهای مشابه موجود در خوشه، استفاده شده است. همچنین، در الگوریتم پیشنهادی، حذف داده‌های پرت، تعیین تعداد خوشه‌های بهینه و غیره در نظر گرفته شده است. این موضوع سبب خواهد شد، داده‌های غیرعادی تأثیری در محاسبهٔ داده‌های از دست رفته نداشته باشند. در الگوریتم پیشنهادی، برای هر خوشه، صفاتی (ستون‌ها) که بیش از یک سوم دادهٔ از دست رفته دارند، حذف می‌شوند. این موضوع سبب جلوگیری از تأثیر داده‌های غیرقابل اطمینان در محاسبهٔ داده‌های از دست رفته خواهد شد. همچنین، از مدل رگرسیون در خوشه استفاده می‌شود که سبب می‌شود در محاسبهٔ داده‌های از دست رفته، فیلدهای مربوط در صفات (ستون‌های) دیگر نیز در نظر گرفته شوند. استفاده از الگوریتم ژنتیک در روش پیشنهادی، که منجر به استفادهٔ تلفیقی از میانگین، میانه، مد، و مدل رگرسیون می‌شود، سبب دستیابی به نتایج قابل قبول‌تری خواهد شد.

کلیدواژه‌ها: بازیابی اطلاعات، داده‌های گم‌شده، داده‌کاوی، الگوریتم ژنتیک، خوشه‌بندی، مدل رگرسیون.

۱. مقدمه

گم‌شدگی داده‌ها در تمامی پژوهش‌های علوم اجتماعی، رفتاری، پزشکی و غیره وجود دارد. در آمار، گم شدن داده‌ها به وضعیتی گفته می‌شود که بخشی از مجموعهٔ داده‌ها گزارش نشده باشد. در ادبیات آماری، اصطلاحات مختلف و غالباً مترادفی برای این مفهوم وجود دارد. این اصطلاحات عبارتند از مقادیر گم‌شده[1]، داده‌های گم‌شده[2]، داده‌های ناقص[3]، و داده‌های بی‌پاسخ[4]. داده‌های گم‌شده، تجزیه‌وتحلیل آماری را به‌سوی نتایج اریب سوق داده و نهایتاً دستیابی به یک نتیجه‌گیری مفید از داده‌های جمع‌آوری شده را با مشکل مواجه می‌سازد (کاظمی، کریملو و رهگذر، ۱۳۹۰).

گم‌شدن داده‌ها به صورت‌های مختلفی می‌تواند رخ دهد. به‌عنوان مثال برخی از افراد شرکت‌کننده در مطالعه از ادامهٔ همکاری انصراف می‌دهند، یا از پاسخ دادن به برخی از سؤالات اجتناب می‌کنند. محققین، تکنسین‌ها، جمع‌آوری‌کننده‌های داده‌ها ممکن است اشتباهاتی را انجام دهند. در یک مطالعهٔ گذشته‌نگر به علت نقص مدارک و سوابق ممکن است برخی اطلاعات در دسترس نباشد. همچنین این احتمال وجود دارد که به علت نقص یا ضعف دستگاه و تجهیزات، امکان مشاهده و اندازه‌گیری وجود نداشته باشد و یا در بعضی از مطالعات نظرسنجی، افرادی قادر به اظهارنظر دقیق نباشند و یا به هر دلیلی دادهٔ جمع‌آوری شده برای برخی از افراد شرکت‌کننده در مطالعه گم شود (کاظمی، کریملو و رهگذر، ۱۳۹۰).

روش‌های مختلفی برای برخورد با داده‌های گم‌شده وجود دارد. یک روش، حذف داده‌های گم‌شده است. نقطه ضعف این روش این است که ممکن است در نهایت حجم نمونه بسیار کوچک شود. روش دیگر جایگزینی داده‌های گم‌شده است. جایگزینی به معنای انتساب یک مقدار گم‌شده با مقدار دیگری بر اساس یک برآورد معقول است. بدین صورت که از داده‌های دیگر برای ایجاد مجدد مقدار گم‌شده استفاده می‌شود. روش جایگزینی با استفاده از تکنیک‌های مختلفی مانند جایگزینی با میانگین داده‌ها، جایگزینی با بزرگ‌ترین عنصر، جایگزینی با کوچک‌ترین عنصر، و غیره امکان‌پذیر است. انتساب یک کار پیچیده است زیرا باید جوانب مثبت و منفی را سنجید. انتساب غیردقیق می‌تواند سبب سوگیری داده‌ها و دستیابی به نتایج اشتباه شود (باقی‌یزدل و همکاران، ۱۳۹۵). لذا، هدف این پژوهش، ارائهٔ روشی برای بازیابی اطلاعات از داده‌های از دست‌رفته با در نظر گرفتن

1. Missing Values
2. Missing Data
3. Incomplete Data
4. Non-Response

معیار دقت است. بدین منظور از تکنیک‌های داده‌کاوی نظیر خوشه‌بندی[1] و مدل رگرسیون[2] و نیز از الگوریتم‌های مکاشفه‌ای[3] شامل الگوریتم ژنتیک[4] استفاده شده است. استفاده از خوشه‌بندی و مدل رگرسیون سبب خواهد شد تنها از داده‌های مشابه برای جانهی داده‌های گم‌شده استفاده شود. همچنین استفاده از الگوریتم ژنتیک سبب خواهد شد در مواقعی که حجم داده‌های گم‌شده زیاد است (بیش از ۲۰ درصد کل داده‌ها) در زمان کمتر، نتایج مناسبی ارائه شود.

در ادامه در بخش ۲، پژوهش‌های مرتبط بیان شده است. در بخش ۳ به توضیح الگوریتم پیشنهادی پرداخته شده است. در بخش ۴ به پیاده‌سازی و ارزیابی الگوریتم پیشنهادی بر روی یک مجموعه داده تصادفی پرداخته شده است و نتایج حاصل از پیاده‌سازی الگوریتم پیشنهادی با الگوریتم‌های دیگر مقایسه شده است. در نهایت در بخش ۵ به نتیجه‌گیری پرداخته شده است.

۲. پیشینه پژوهش

در این بخش به پژوهش‌های مرتبط با بازیابی اطلاعات از داده‌های گم‌شده پرداخته شده است. تادا، سوزوکی و اوکادا[5] (۲۰۲۲) به روش محاسبه مقدار گم‌شده برای داده‌های ماتریس چندکلاسه بر اساس مجموعه فقره‌های بسته پرداختند. در پژوهش آن‌ها، دو روش مبتنی بر مجموعه فقره‌های بسته، شامل روش جانهی مبتنی بر مجموعه اقلام بسته[6] و روش جانهی بهبود یافته مبتنی بر مجموعه اقلام بسته[7] برای دستیابی به مقدار گم‌شده با استفاده از فضای ویژگی محلی برای داده‌های ماتریس چندکلاسه پیشنهاد شده است. روش جانهی مبتنی بر مجموعه اقلام بسته مقادیر گم‌شده را با استفاده از مجموعه فقره‌های بسته استخراج شده از هر کلاس تخمین می‌زند. روش جانهی بهبود یافته مبتنی بر مجموعه اقلام بسته یک روش بهبودیافته روش جانهی مبتنی بر مجموعه اقلام بسته است که در آن یک فرایند کاهش ویژگی معرفی شده است. نتایج تجربی نشان می‌دهد که کاهش ویژگی به‌طور قابلِ توجهی زمان محاسباتی را کاهش می‌دهد و دقت انتساب را بهبود می‌بخشد. علاوه بر این، نشان داده شده است که در مقایسه با روش‌های موجود، روش جانهی بهبود یافته مبتنی بر مجموعه اقلام بسته دقت انتساب بالاتری را ارائه می‌کند اما به زمان محاسباتی بیشتری نیاز دارد.

1. Clustering
2. Regression Model
3. Heuristic Algorithm
4. Genetic Algorithm
5. Tada, Suzuki & Okada
6. Closed Itemset-Based Imputation (CIimpute)
7. Improved Closed Itemset-Based Imputation (ICIimpute)

ژنگ[1] و همکاران (۲۰۲۱) به پژوهشی با عنوان «انتساب مقدار از دست‌رفته در سری‌های زمانی چندمتغیره با شبکه‌های متخاصم مولد» پرداختند. نویسندگان بیان می‌کنند که یک فاز اضافی برای بهینه‌سازی نویز تصادفی ورودی ژنراتور نیاز است. علاوه بر این، مقادیر منتسب به دلیل دشواری آموزش شبکه‌های متخاصم مولد[2] و فرایند تولید ناپایدار، می‌توانند بسیار متفاوت از مقادیر واقعی باشند. بنابراین، آن‌ها یک مدل پایان به انتها برای نسبت دادن مقادیر از دست‌رفته در یک سری زمانی چندمتغیره پیشنهاد می‌کنند. به‌طور خاص، یک شبکهٔ رمزگذار در معماری شبکه‌های متخاصم مولد استاندارد معرفی می‌شود که فاز بهینه‌سازی ورودی را حذف می‌کند. آزمایش‌ها بر روی سه مجموعه دادهٔ سری زمانی چند متغیره در دنیای واقعی نشان می‌دهند که مدل پیشنهادی از روش‌های پیشرفته در وظایف انتساب و کاربردهای پایین دستی، از جمله طبقه‌بندی و رگرسیون، بهتر عمل می‌کند.

لی[3] و همکاران (۲۰۲۱) به ارائهٔ یک روش ترکیبی که تجزیهٔ حالت تجربی و یک شبکهٔ حافظهٔ کوتاه‌مدت را برای پیش‌بینی داده‌های سیگنال اندازه‌گیری گم‌شده سیستم‌های پایش سلامت سازه[4] جفت می‌کند، پرداخته‌اند. در مطالعهٔ آن‌ها، تجزیهٔ حالت تجربی با یک شبکهٔ یادگیری عمیق حافظه کوتاه مدت برای بازیابی داده‌های سیگنال اندازه‌گیری شده ترکیب شده است. روش ترکیبی پیشنهادی وظیفهٔ انتساب داده‌های گم‌شده را به‌یک کار پیش‌بینی سری زمانی تبدیل می‌کند که سپس با یک استراتژی «تقسیم‌وغلبه»[5] حل می‌شود. مفهوم اصلی این استراتژی، پیش‌بینی دنباله‌های داده‌های سیگنال اندازه‌گیری شدهٔ خام است که به جای پیش‌بینی مستقیم، با تجزیهٔ حالت تجربی تجزیه می‌شوند، زیرا تجزیه می‌تواند به مدل‌سازی تغییرات دوره‌ای نامنظم داده‌های سیگنال اندازه‌گیری شده کمک کند. علاوه بر این، شبکهٔ حافظهٔ کوتاه‌مدت بلندمدت در مدل ترکیبی می‌تواند همبستگی‌های دوربرد بیشتری از دنباله‌های فرعی را نسبت به شبکهٔ عصبی مصنوعی سنتی به خاطر بسپارد. سه مدل پیش‌بینی پرکاربرد، یعنی میانگین متحرک خودرگرسیونی یکپارچه[6]، رگرسیون بردار پشتیبان[7]، و مدل‌های شبکهٔ عصبی مصنوعی[8] نیز به‌عنوان مدل‌های معیار پیاده‌سازی می‌شوند. داده‌های شتاب خام جمع‌آوری‌شده از یک پل کابلی برای ارزیابی عملکرد روش پیشنهادی برای از دست رفتن دادهٔ

1. Zhang
2. Generative Adversarial Networks (GANs)
3. Li
4. Structural Health Monitoring (SHM)
5. Divide and Conquer
6. Autoregressive Integrated Moving Average
7. Support Vector Regression (SVR)
8. Artificial Neural Network

سیگنال اندازه‌گیری‌شده، استفاده می‌شود. نتایج بازیابی داده‌های سیگنال اندازه‌گیری‌شده نشان می‌دهد که روش ترکیبی پیشنهادی از دو منظر عملکرد عالی دارد. اول، تجزیهٔ حالت تجربی می‌تواند دقت مدل پیش‌بینی حافظهٔ کوتاه‌مدت هسته را بهبود بخشد. دوم، مدل حافظهٔ کوتاه‌مدت بلندمدت از سایر مدل‌های معیار بهتر عمل می‌کند، زیرا می‌تواند تغییرات میکروسکوپی بیشتری را در مقادیر اندازه‌گیری‌شده مطابقت دهد. آزمایش‌های انجام‌شده در مطالعهٔ آن‌ها همچنین نشان می‌دهد که الگوهای تغییر داده‌های سیگنال اندازه‌گیری‌شدهٔ خام پیچیده هستند، و بنابراین استخراج ویژگی‌ها قبل از مدل‌سازی مهم است.

لیو[1] و همکاران (۲۰۲۰) در پژوهش خود با عنوان «مقدار گم‌شده برای داده‌های حسگر اینترنت اشیاء[2] صنعتی با شکاف‌های بزرگ» بیان می‌کنند که داده‌های ازدست‌رفته یکی از بزرگترین مشکلات برای پیش‌پردازش داده‌ها در معماری اینترنت اشیاء است. به دلیل جمع‌آوری داده‌های حسگر با فرکانس بالا، داده‌های از دست رفته در اینترنت اشیاء چالش‌های جدیدی را به همراه دارد. برای پرداختن به این موضوع، نویسندگان بر روی انتساب داده‌های گم‌شده برای شکاف‌های بزرگ در داده‌های سری زمانی تک متغیره تمرکز می‌کنند و یک چارچوب تکراری با استفاده از تکرار شکاف چندگانه[3] برای ارائهٔ مناسب‌ترین مقادیر پیشنهاد شده است. شکاف ابتدا به چند قطعه تقسیم می‌شود تا فرایند انتساب مقدار گم‌شده را مقداردهی اولیه کند و سپس، به‌طور مکرر بازسازی شکاف و الحاق شکاف برای به دست آوردن نتایج انتساب نهایی اجرا می‌شود. روش پیشنهادی با استفاده از داده‌های حسگر جمع‌آوری‌شده از کارخانه‌های تولید واقعی در استرالیا بررسی شده است و نتایج مقایسه نشان می‌دهد که روش پیشنهادی از نظر خطای ریشهٔ میانگین مربع از روش‌های پیشرفته بهتر عمل می‌کند. در شرایط مختلف طول شکاف، رویکرد پیشنهادی به‌طور مداوم میزان خطا را بیشتر از الگوریتم پایه کاهش می‌دهد و کاهش خطا زمانی که طول شکاف‌ها افزایش می‌یابد بیشتر است، که نشان می‌دهد عملکرد به‌طور قابل توجهی بهبود یافته است.

بیزمن[4] و همکاران (۲۰۱۹) به پژوهشی با عنوان «مقدار از دست‌رفته برای جداول» پرداختند. آن‌ها بیان می‌کنند که روش‌های انتساب مقدار گم‌شدهٔ کنونی بر روی داده‌های عددی یا دسته‌بندی تمرکز می‌کنند و مقیاس‌بندی آن‌ها در مجموعه‌های داده با میلیون‌ها ردیف دشوار است. آن‌ها روشی

1. Liu
2. Internet of Things (IoT)
3. Itr-MS-STLecImp
4. Biessmann

را ارائه می‌کنند که می‌تواند برای جداول با انواع داده‌های ناهمگن، از جمله متن بدون ساختار اعمال شود. روش پیشنهادی، استخراج‌کنندهٔ ویژگی‌های یادگیری عمیق[1] را با تنظیم خودکار فراپارامتر ترکیب می‌کند. این به کاربران بدون پیش‌زمینهٔ یادگیری ماشین[2]، مانند مهندسان داده، امکان می‌دهد تا مقادیر گم‌شده را با کم‌ترین تلاش در جدول‌هایی با انواع داده‌های ناهمگون نسبت دهند. آن‌ها روش پیشنهادی را با روش‌های موجود مقایسه می‌کنند و نتایج نشان از رضایت‌بخش بودن روش ارائه شده دارد.

فصیحی و قلیزاده‌گزور (۱۴۰۰) به جانهی داده‌های گم‌شده با بهره‌مندی از رویکردهای آماری پرداخته‌اند. برای نشان‌دادن طرز کار روش‌های پیشنهادی، از روش ارائه‌شده برای ارزیابی مجموعه‌ای از مدارس متوسطهٔ دولتی یونان که برخی دارای مقادیر گم‌شده در ورودی یا خروجی هستند، بهره برده شده است.

رشیدی‌نژاد (۱۳۹۴) به ارائه یک روش پیشنهادی جانهی داده‌های گم‌شده بر اساس میانگین داده‌ها و براوردگر والش[3] پرداخته است. همچنین نشان داده است که براوردگر روش پیشنهادی بهتر از براوردگر حاصل از روش جانهی نسبتی است و میانگین مربع خطای براوردگر حاصل کم‌تر از براوردگر به دست آمده بر اساس روش جانهی نسبتی است.

عابدپور[4] و همکاران (۲۰۲۴) یک الگوریتم ترکیبی یادگیری ماشین و الگوریتم ژنتیک برای اختصاص منابع مورد نیاز برنامه‌های کاربردی اینترنت اشیاء در محیط رایانش ابری[5] ارائه دادند. در روش پیشنهادی با استفاده از تکنیک‌های خوشه‌بندی داده‌کاوی نیازمندی‌های برنامه‌های کاربردی دسته‌بندی می‌شوند و در عین حال منابع ناهمگن مه نیز با نظم خاصی از لحاظ تشابه کمی گروه‌بندی می‌شوند. سپس الگوریتم ژنتیک بهترین اختصاص منابع گروهی به منابع درخواستی برنامه‌های کاربردی نگاشت می‌دهد تا هدررفت منابع را کمینه‌سازی کند.

از آنجایی‌که الگوریتم‌های ژنتیک از سوی خطر تله بهینه محلی و همگرائی زودرس تهدید می‌شوند، برای رفع این مشکل خادمی دهنوی[6] و همکاران (۲۰۲۴) یک الگوریتم ژنتیک هیبریدی برای حل مسئلهٔ زمان‌بندی وظایف برنامه‌های کاربردی در رایانش ابر ارائه دادند. نوآوری روش

1. Deep Learning
2. Machine Learning
3. Walsh Appraisal
4. Abedpour
5. Cloud Computing
6. Khademi Dehnavi

پیشنهادی ارائهٔ یک الگوریتم تپه‌نوری[1] در جستجوی محلی ژنتیک در کنار جستجوی سراسری بود که کیفیت نهایی پاسخ را بهبود بخشید.

ونگ[2] و همکاران (۲۰۲۳) یک الگوریتم ژنتیک توسعه‌یافته مبتنی بر خوشه‌بندی برای مسئلهٔ مسیریابی خودرو پیشنهاد دادند. در این پژوهش، یک الگوریتم ترکیبی دو مرحله‌ای با ترکیب یک الگوریتم خوشه‌بندی و الگوریتم ژنتیک برای یافتن راه‌حل‌های بهینه پارتو[3] و حل بهینه‌سازی ایجاد شده است. بر اساس الگوریتم ترکیبی ارائه شده، پیچیدگی محاسباتی مسیریابی خودروها کاهش یافته است.

یی‌فی[4] و همکاران (۲۰۲۳) به شناسایی آسیب سازه در سدها با استفاده از خوشه‌بندی ترکیبی کی- میانگین[5] و الگوریتم ژنتیک پرداختند. در این مطالعه، یک روش جدید برای شناسایی آسیب ساختاری توسعه یافته است که از یک تکنیک مدل‌سازی جایگزین پیشرفته برای هدایت یک استراتژی بهینه‌سازی ترکیبی جدید، یعنی ترکیبی از بهینه‌ساز خوشه‌بندی کی- میانگین و الگوریتم ژنتیک استفاده می‌کند. این روش، کارایی استراتژی بهینه‌سازی در یافتن مقدار بهینهٔ تابع هدف را تا حد زیادی بهبود می‌بخشد.

آواد[6] و همکاران (۲۰۲۳) به طبقه‌بندی و تشخیص قوی داده‌های پزشکی بزرگ با استفاده از خوشه‌بندی پیشرفته کی- میانگین موازی و رگرسیون لجستیک[7] پرداختند. در این مطالعه بیان شده است از آن جا که مدل رگرسیون دارای محدودیت‌ها و مشکلات عملکردی با داده‌های بزرگ پزشکی است، این پژوهش یک رویکرد قوی برای طبقه‌بندی داده‌های پزشکی بزرگ و تشخیص تصویر با پیش‌پردازش خوشه‌بندی کی- میانگین ارائه می‌دهد. افزایش سرعت و کارایی از نتایج روش ارائه شده بوده است.

لیو[8] و همکاران (۲۰۱۹) به تأثیر داده‌های پرت[9] بر خوشه‌بندی پرداختند. در مطالعهٔ آن‌ها، تجزیه‌وتحلیل خوشهٔ مشترک و مشکل تشخیص نقاط پرت در نظر گرفته شده است و الگوریتم

1. Hill Climbing
2. Wang
3. Pareto
4. YiFei
5. K-means
6. Awad
7. Logistic Regression
8. Liu
9. Outlier

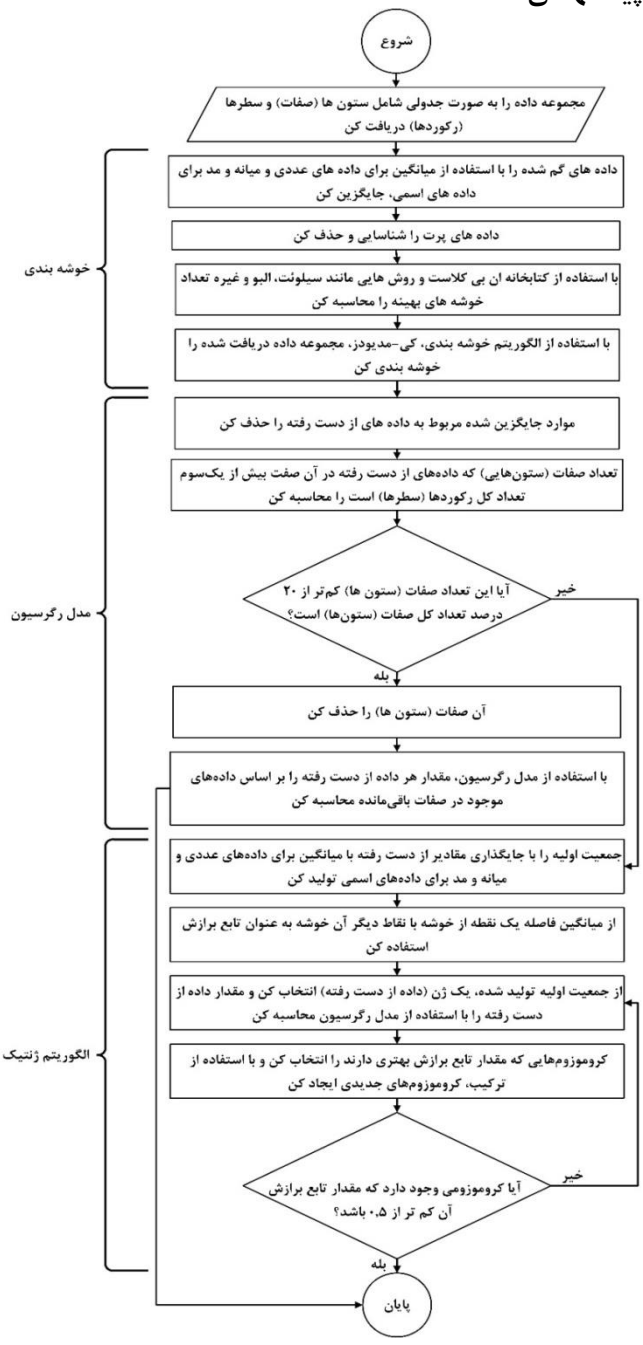
خوشه‌بندی با حذف داده‌های پرت[1] پیشنهاد شده است. نتایج آزمایشات نویسندگان بر روی مجموعه‌های داده‌های متعدد در حوزه‌های مختلف، اثربخشی و کارایی این روش را به‌طور قابل‌توجهی نسبت به روش‌های پیشرفته از نظر اعتبار خوشه‌ای و تشخیص پرت نشان می‌دهد.

از آنجایی که تعیین تعداد بهینهٔ خوشه‌ها، انتخاب تصادفی مراکز اولیه، عدم تشخیص خوشه‌های غیرکروی، و تأثیر منفی نقاط پرت از چالش‌های اصلی الگوریتم خوشه‌بندی کی-میانگین هستند، حیدری[2] و همکاران (۲۰۲۴)، برای مقابله با این مسائل، سه الگوریتم ساده و هوشمند با تغییر ساختار الگوریتم‌های کی-میانگین و کی-واسط[3] پیشنهاد دادند. تفاوت این الگوریتم‌ها در انتخاب مراکز اولیه و شرط توقف است. با استفاده از روش ارائه شده، الگوریتم‌های خوشه‌بندی ذکرشدهٔ دیگر نسبت به نقاط پرت حساس نیستند و می‌توانند خوشه‌هایی را با اشکال غیرکروی شناسایی کنند.

همان‌طور که در پیشینهٔ پژوهش اشاره شد برخی روش‌های ارائه شده برای بازیابی اطلاعات از داده‌های گم‌شده مانند روش ارائه شده توسط تادا، سوزوکی و اوکادا (۲۰۲۲) و لی و همکاران (۲۰۲۱) به زمان محاسباتی بالایی نیاز دارد. برخی روش‌ها مانند روش ارائه شدهٔ لیو و همکاران (۲۰۲۰) برای بازیابی مقادیر گم‌شده، بر روی داده‌های عددی یا دسته‌بندی تمرکز می‌کنند و مقیاس‌بندی آن‌ها در مجموعه‌های داده با میلیون‌ها ردیف دشوار است. روش جایگزین کردن داده‌های از دست‌رفته با میانگین سایر داده‌ها مانند روش ارائه شدهٔ رشیدی‌نژاد (۱۳۹۴) با این که در بسیاری از مواقع، کارآمد است اما موجب اختلال در توزیع و کم برآورد شدن واریانس می‌شود. همچنین روش استفاده از برازش رگرسیون خطی، همهٔ داده‌ها را روی یک خط راست برازش می‌کند و لذا برای حالتی که داده‌ها نوسان زیادی دارند، تخمین مناسبی به حساب نمی‌آید. در این پژوهش سعی شده است با تکنیک‌های داده‌کاوی نظیر خوشه‌بندی، به شناسایی داده‌های مشابه و سپس تصمیم‌گیری دربارهٔ جایگزینی داده‌های ازدست‌رفته براساس روش‌هایی مانند مدل رگرسیون پرداخته شود. همچنین زمانی که تعداد داده‌های ازدست‌رفته زیاد باشد، از الگوریتم‌های مکاشفه‌ای نظیر الگوریتم ژنتیک استفاده شود. استفاده از خوشه‌بندی و شناسایی داده‌های مشابه در روش پیشنهادی، سبب بهبود مشکل تخمین داده‌های گم‌شده و افزایش دقت می‌شود. همچنین، الگوریتم ژنتیک در روش پیشنهادی سبب حل مشکل زمان محاسباتی بالا خواهد شد.

1. Clustering with Outlier Removal (COR) Algorithm
2. Heidari
3. K-medoids

۳. الگوریتم پیشنهادی



شکل ۱- فلوچارت الگوریتم پیشنهادی

در این بخش، روش پیشنهادی که شامل خوشه‌بندی، حذف داده‌های پرت، مدل رگرسیون، الگوریتم ژنتیک است، توضیح داده شده است. شکل ۱ (شکل بالا) فلوچارت الگوریتم پیشنهادی را نشان می‌دهد. مجموعه داده شامل داده‌های ازدست‌رفته به‌صورت جدولی شامل ستون‌ها (صفات) و سطرها (رکوردها) دریافت می‌شود.

۳-۱. خوشه‌بندی

یکی از مواردی که در روش‌های موجود جایگزینی داده‌های ازدست‌رفته وجود دارد این است که جایگزینی دادهٔ ازدست‌رفته بر اساس کل داده‌های موجود انجام می‌شود. به‌عنوان مثال از میانگین، میانه یا مد کل مجموعه داده‌ها استفاده می‌شود. مشکل این روش این است که از رکوردهایی که مشابه با رکورد مربوط به فیلد ازدست‌رفته نیست، استفاده می‌شود. این موضوع، باعث جایگزینی دادهٔ ازدست‌رفته بر اساس دادهٔ اکثریت می‌شود نه بر اساس تشابهی که دادهٔ ازدست‌رفته با سایر داده‌های مشابه دارد. لذا، سبب جایگزینی اشتباه دادهٔ ازدست‌رفته خواهد شد. برای غلبه بر این مشکل، نیاز است از خوشه‌بندی برای شناسایی رکوردهای مشابه با رکوردی که دارای فیلد ازدست‌رفته است، استفاده شود. برای خوشه‌بندی نیاز است به نکاتی توجه قرار گیرد که در ادامه بیان شده است.

جای‌گذاری اولیهٔ مقادیر ازدست‌رفته. برای امکان خوشه‌بندی، نیاز است داده‌های ازدست‌رفته با مقادیری جای‌گذاری شوند. برای این که این مقادیر تأثیری در خوشه‌بندی نداشته باشند، می‌توان از جای‌گذاری مقادیر ازدست‌رفته با مقدار اکثریت استفاده کرد. لذا از میانگین برای داده‌های عددی و میانه و مد برای داده‌های اسمی استفاده خواهد شد. توجه شود این جای‌گذاری مقادیر ازدست‌رفته، فقط برای امکان خوشه‌بندی است و پس از خوشه‌بندی، مقادیر انتساب‌یافته حذف خواهند شد.

شناسایی داده‌های پرت. وجود داده‌های پرت، سبب خوشه‌بندی اشتباه و در نتیجه جای‌گذاری داده‌های ازدست‌رفته با مقادیر اشتباه خواهد شد. لذا نیاز است قبل از خوشه‌بندی، داده‌های پرت، شناسایی و حذف شوند.

تعیین تعداد خوشه بهینه. یکی از مواردی که در خوشه‌بندی باید به آن توجه داشت، تعیین تعداد خوشه‌های بهینه است. برای شناسایی تعداد خوشه‌های بهینه از کدنویسی در زبان برنامه‌نویسی آر[1] و کتابخانه انبی‌کلاست[2] استفاده می‌شود. روش‌های استفاده‌شده‌در تعیین تعداد خوشه‌های بهینه

1. R Programming
2. NbClust

می‌توانند مواردی مانند سیلوئت[1]، آرنج[2]، آماره شکاف[3] و شبه[4] باشند.

الگوریتم خوشه‌بندی. پس از مشخص شدن تعداد خوشه‌های بهینه باید الگوریتمی برای خوشه‌بندی انتخاب شود. علی‌رغم شناسایی داده‌های پرت و حذف آن‌ها، باز هم ممکن است تعدادی داده پرت باقی مانده باشند. لذا به دلیل حساس بودن الگوریتم کی-میانگین به داده‌های پرت از الگوریتم کی-واسط استفاده خواهد شد.

حذف موارد جایگزین شدهٔ مربوط به داده‌های ازدست‌رفته: پس از خوشه‌بندی، داده‌های فرضی که برای انجام خوشه‌بندی تخصیص داده شده بودند، حذف می‌شوند.

۳-۲. مدل رگرسیون

برای هر خوشه، مراحل زیر نیاز است انجام شود:

۱. تعداد صفاتی (ستون‌هایی) که داده‌های ازدست‌رفته در آن صفت بیش از یک‌سوم تعداد کل رکوردها (سطرها) است را محاسبه کن.

۲. در صورتی که تعداد صفات (ستون‌های) محاسبه شده در مرحلهٔ (۱) کم‌تر از ۲۰ درصد تعداد کل صفات (ستون‌ها) است، آن صفات (ستون‌ها) را حذف کن و با استفاده از مدل رگرسیون، مقدار هر دادهٔ ازدست‌رفته را براساس داده‌های موجود در صفات باقی‌مانده محاسبه کن.

۳. در صورتی که تعداد صفات (ستون‌های) محاسبه شده در مرحلهٔ (۱) بیش از ۲۰ درصد تعداد کل صفات (ستون‌ها) است، از الگوریتم ژنتیک استفاده کن.

۳-۳. الگوریتم ژنتیک. برای هر خوشه، مراحل زیر را انجام بده:

تولید جمعیت اولیه. جمعیت اولیه را با جای‌گذاری مقادیر ازدست‌رفته با میانگین برای داده‌های عددی و میانه و مد برای داده‌های اسمی تولید کن.

تابع برازش. از میانگین فاصله یک نقطه از خوشه با نقاط دیگر آن خوشه به عنوان تابع برازش[5] استفاده کن که با فرمول ۱ محاسبه می‌شود.

$$a(i) = \frac{1}{n_i} \sum_{l=1}^{n_i} d(x_i , x_l)$$ **فرمول ۱**

این معیار را می‌توان ملاکی برای ارزیابی تعلق نقطه x_i در خوشه‌اش در نظر گرفت. هر چه مقدار

1. Silhouette
2. Elbow
3. Gap Statistic
4. Pseudo
5. Fitness Function

$a(i)$ کوچکتر باشد، میزان تعلق این نقطه به خوشه‌اش بیشتر است.

جهش. از جمعیت اولیهٔ تولید شده، یک ژن (داده از دست رفته) انتخاب کن و مقدار دادهٔ ازدست‌رفته را با استفاده از مدل رگرسیون محاسبه کن. سپس، مقدار تابع برازش برای کروموزوم جدید را محاسبه کن.

ترکیب. کروموزوم‌هایی که مقدار تابع برازش بهتری دارند را انتخاب کن و با استفاده از ترکیب، کروموزوم‌های جدیدی ایجاد کن.

شرط توقف. تا رسیدن به کروموزومی که مقدار تابع برازش آن کمتر از ۰٫۵ باشد مراحل جهش و ترکیب را تکرار کن.

۴. پیاده‌سازی و ارزیابی الگوریتم پیشنهادی

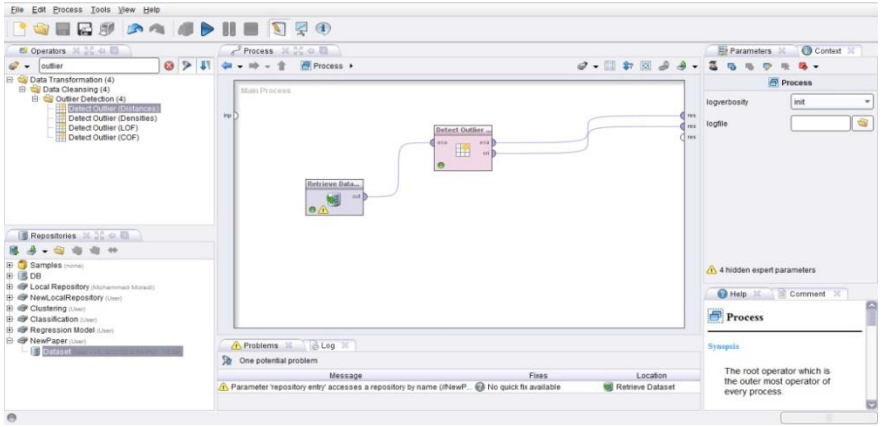
در این بخش به پیاده‌سازی و ارزیابی الگوریتم پیشنهادی پرداخته شده است. بدین منظور از مجموعه داده‌های سوابق بالینی نارسایی قلبی واقع در سایت UCI[1] استفاده شده است. تعداد رکوردها (نمونه‌ها) در این مجموعه داده برابر ۲۹۹ است. همچنین این مجموعه داده دارای ۱۳ صفت است. برای پیاده‌سازی و ارزیابی الگوریتم پیشنهادی، از ۱ درصد از کل مجموعه داده، به‌صورت تصادفی به‌عنوان مقادیر ازدست‌رفته در نظر گرفته شد. سپس طبق الگوریتم پیشنهادی، داده‌های از دست رفته با میانگین، میانه، و مد جای‌گذاری شد. شکل ۲، بخشی از فیلدهای تصادفی در نظر گرفته شده به‌عنوان داده‌های ازدست‌رفته (هایلایت شده با رنگ زرد) و جای‌گذاری هر یک بر اساس میانگین، میانه، و مد را نشان می‌دهد.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	age	anaemia	creatinine	diabetes	ejection_fr	high_blood	platelets	serum_cre	serum_so	sex	smoking	time	DEATH_EVENT					
2	75	0	582	0	20	1	265000	1.9	130	1	0	4	1					
3	55	0	7861	0	38	0	263358	1.1	136	1	0	6	1					
4	65	0	146	0	20	0	162000	1.3	129	1	1	7	1					
5	36	1	111	0	20	0	210000	1.9	137	1	0	7	1					
6	65	1	160	1	20	0	327000	2.7	116	0	0	8	1					
7	90	1	47	0	40	1	204000	2.1	132	1	1	8	1					
8	75	1	246	0	15	0	127000	1.2	137	1	0	10	1					
9	60	1	315	0	60	0	454000	1.1	131	1	1	10	1					
10	65	0	157	0	65	0	263358	1.5	138	0	0	10	1					
11	80	1	123	0	35	1	388000	9.4	133	1	1	10	1					
12	75	1	81	0	38	1	368000	4	131	1	1	10	1					
13	62	0	231	0	25	1	253000	0.9	140	1	1	10	1					
14	45	1	981	0	30	0	136000	1.1	137	1	0	11	1					
15	50	1	168	0	38	1	276000	1.1	137	1	0	11	1					
16	82	1	379	0	50	0	47000	1.3	136	1	0	13	1					
17	87	1	149	0	38	0	262000	0.9	140	1	0	14	1					
18	45	0	582	0	14	0	166000	0.8	127	1	0	14	1					
19	70	1	125	0	25	1	237000	1	140	0	0	15	1					
20	48	1	582	1	55	0	87000	1.9	121	0	0	15	1					
21	65	1	52	0	25	1	276000	1.3	137	0	0	16	0					
22	65	1	128	1	30	1	297000	1.6	136	0	0	20	1					

شکل ۲- بخشی از فیلدهای تصادفی در نظر گرفته شده به‌عنوان مقادیر ازدست‌رفته و جای‌گذاری آن‌ها با میانگین، میانه، و مد

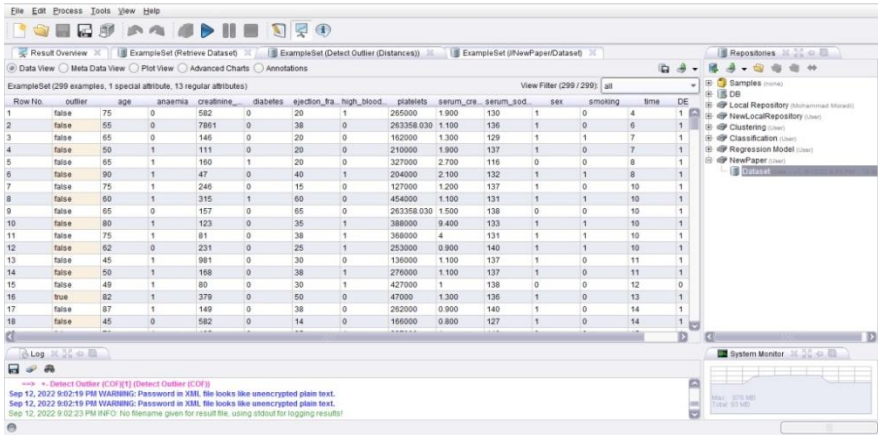
1. https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records#

طبق گام بعدی الگوریتم پیشنهادی، نیاز است داده‌های پرت، شناسایی و حذف شوند. بدین منظور از نرم‌افزار رپیدماینر[1] استفاده شد. شکل ۳ عملگرهای[2] استفاده شده برای شناسایی و حذف داده‌های پرت را نشان می‌دهد.



شکل ۳- عملگرهای استفاده شده در نرم‌افزار رپیدماینر برای شناسایی و حذف داده‌های پرت

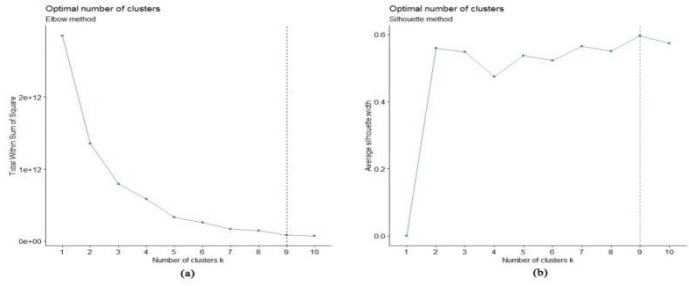
همچنین شکل ۴ نتیجه شناسایی داده‌های پرت را نشان می‌دهد. برای مثال در این مجموعه داده، رکورد ۱۶، به‌عنوان داده پرت تشخیص داده شده است. تعداد ۸ رکورد به‌عنوان داده‌های پرت تشخیص داده شد که از مجموعه داده حذف شد.



شکل ٤- تشخیص داده‌های پرت برای مجموعه دادهٔ مطالعه‌شده

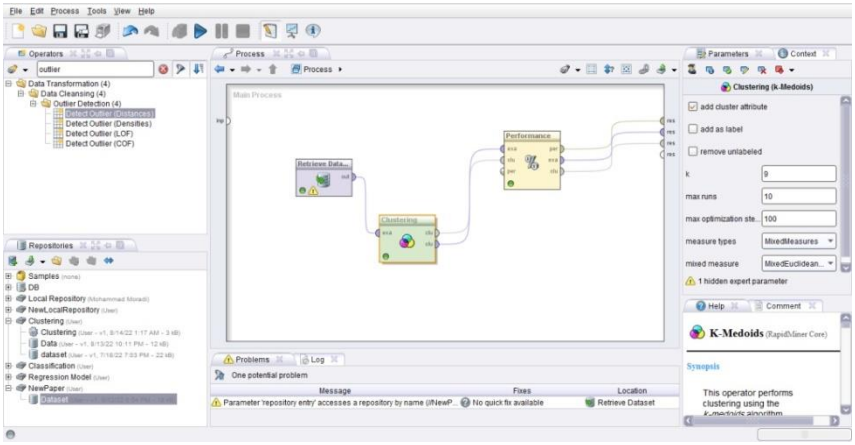
1. Rapid Miner
2. Operator

سپس با استفاده از کتابخانه انبی‌کلاست و کدنویسی در زبان برنامه‌نویسی آر و استفاده از روش‌های سیلوئت، آرنج، آماره شکاف، شاخص سی[۱]، توپ[۲]، دودا[۳]، بیل[۴] و سی‌اچ[۵]، تعداد خوشهٔ بهینه برای مجموعه دادهٔ بررسی‌شده، محاسبه شد و اکثر روش‌ها، ۹ خوشه را به‌عنوان تعداد خوشه بهینه مشخص کردند. شکل ۵، تعداد خوشه‌های مشخص شده بر اساس دو روش آرنج و سیلوئترا نشان می‌دهد.



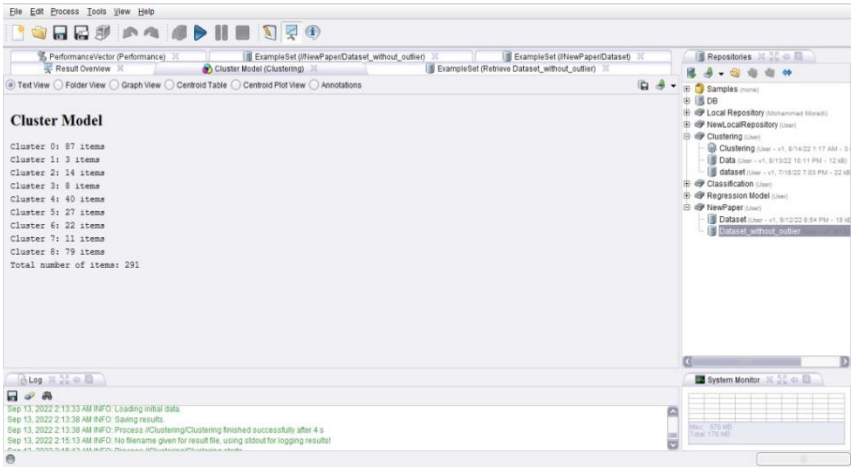
شکل ۵- روش‌های استفاده‌شده در تعیین تعداد بهینهٔ خوشه‌ها: (الف) روش آرنج (ب) روش سیلوئت

پس از مشخص شدن تعداد خوشه بهینه، به خوشه‌بندی مجموعه داده مطالعه‌شده با استفاده از نرم‌افزار رپیدماینر پرداخته شد. شکل ۶، عملگرهای استفاده شده، و شکل ۷، نتایج خوشه‌بندی و تعداد رکورد درون هر خوشه را نشان می‌دهد.



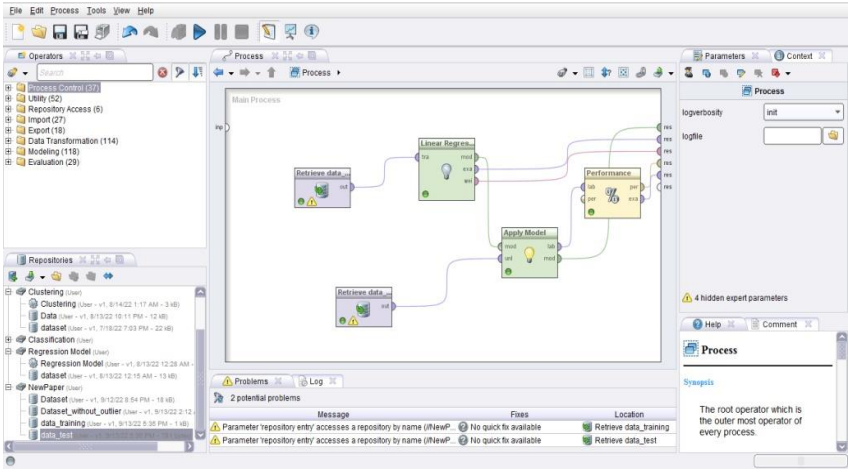
شکل ٦- عملگرهای استفاده شده در خوشه‌بندی مجموعه داده با استفاده از نرم‌افزار رپیدماینر

1. Cindex
2. Ball
3. Duda
4. Beale
5. Ch



شکل ۷- نتایج خوشه‌بندی بر روی مجموعه دادهٔ مطالعه‌شده

طبق گام بعدی الگوریتم پیشنهادی، برای هر خوشه، تعداد صفاتی (ستون‌هایی) که داده‌های ازدست‌رفته در آن صفت بیش از یک‌سوم تعداد کل رکوردها (سطرها) بود، محاسبه شد. چون تعداد این صفات کمتر از ۲۰ درصد تعداد کل صفات بود، آن صفات (ستون‌ها) حذف شدند و با استفاده از مدل رگرسیون، مقدار هر دادهٔ ازدست‌رفته براساس داده‌های موجود در صفات باقی‌مانده محاسبه شد. از داده‌های ازدست‌رفته به‌عنوان داده‌های تست و از دیگر داده‌های خوشه، به‌عنوان دادهٔ آموزش استفاده شد. شکل ۸، اوپراتورهای استفاده شده در طراحی رگرسیون مدل در نرم‌افزار رپیدماینر را نشان می‌دهد.



شکل ۸- عملگرهای استفاده‌شده در طراحی مدل رگرسیون در نرم‌افزار رپیدماینر

همچنین از فرمول ۲، برای محاسبهٔ درصد خطا برای داده‌های عددی[1]، از فرمول ۳ برای محاسبه درصد خطا برای داده‌های دودویی[2]، و از فرمول ۴ برای محاسبهٔ درصد خطای نهایی استفاده شد.

$$\text{PErrorN} = \frac{|a-b|}{a}$$ **فرمول ۲**

که a، مقدار واقعی و b، مقدار تخمین زده شده برای داده از دست رفته می‌باشد.

$$\text{PErrorB} = \frac{NF}{NF+NT}$$ **فرمول ۳**

که NF، تعداد مواردی است که دادهٔ ازدست‌رفتهٔ دودویی اشتباه تخمین زده شده است. NT، تعداد مواردی است که دادهٔ ازدست‌رفتهٔ دودویی درست تخمین زده شده است.

$$\text{PError} = \frac{PErrorN+PErrorB}{2}$$ **فرمول ۴**

جدول ۱، نتایج مربوط به میانگین درصد خطای پیاده‌سازی روش پیشنهادی و روش استفاده از میانگین، میانه و مد، روش رگرسیون و همچنین روش بردار پشتیبان[3] در جای‌گذاری داده‌های ازدست‌رفته را نشان می‌دهد.

جدول ۱- نتایج مربوط به میانگین درصد خطای پیاده‌سازی روش پیشنهادی و روش استفاده از میانگین، میانه و مد، روش رگرسیون و روش بردار پشتیبان در جای‌گذاری داده‌های ازدست‌رفته

روش	درصد خطا
روش استفاده از میانگین، میانه و مد برای جای‌گذاری داده‌های ازدست‌رفته	۵۶٫۵
رگرسیون	۳۴٫۶
ماشین بردار پشتیبان	۴۲٫۱
روش پیشنهادی	۲۷

همان‌طور که در جدول ۱ مشاهده می‌شود، درصد خطای روش پیشنهادی به‌طور میانگین، حدوداً نصف روش استفاده از میانگین، میانه و مد بوده و همچنین نسبت به روش‌های رگرسیون و ماشین بردار پشتیبان عملکرد بهتری داشته است.

۵. نتیجه‌گیری

در این پژوهش به ارائهٔ الگوریتمی برای بازیابی اطلاعات از داده‌های گم‌شده پرداخته شد. در الگوریتم پیشنهادی از خوشه‌بندی، رگرسیون و الگوریتم ژنتیک استفاده شد. نتایج نشان داد الگوریتم

1. Numerical
2. Binomial
3. Support Vector Machine (SVM)

پیشنهادی می‌تواند دارای خطایی برابر با نصف روش جای‌گذاری با میانگین، میانه و مد (امانوئل[1] و همکاران، ۲۰۲۱) باشد. روش رگرسیون و روش ماشین بردار پشتیبان اگرچه درصد خطای کمتری نسبت به روش میانگین، میانه و مد داشتند، ولی دارای عملکرد بدتری نسبت به روش پیشنهادی بودند. روش جایگزین کردن داده‌های ازدست‌رفته با میانگین سایر داده‌ها مانند روش ارائه شدهٔ رشیدی‌نژاد (۱۳۹۴) با این که در بسیاری از مواقع، کارآمد است اما موجب اختلال در توزیع و کم برآورد شدن واریانس می‌شود. همچنین روش استفاده از برازش رگرسیون خطی، همهٔ داده‌ها را روی یک خط راست برازش می‌کند و لذا برای حالتی که داده‌ها نوسان زیادی دارند، تخمین مناسبی به حساب نمی‌آید. برخی روش‌ها مانند روش ارائه‌شدهٔ لیو و همکاران (۲۰۲۰) برای بازیابی مقادیر گم‌شده، بر روی داده‌های عددی یا دسته‌بندی تمرکز می‌کنند و مقیاس‌بندی آن‌ها در مجموعه‌های داده با میلیون‌ها ردیف دشوار است. همچنین روش ارائه شدهٔ تادا، سوزوکی و اوکادا (۲۰۲۲) و لی و همکاران (۲۰۲۱) به زمان محاسباتی بالایی نیاز دارد. در این پژوهش سعی شده است با تکنیک‌های داده‌کاوی نظیر خوشه‌بندی، به شناسایی داده‌های مشابه و سپس تصمیم‌گیری دربارهٔ جایگزینی داده‌های ازدست‌رفته براساس روش‌هایی مانند مدل رگرسیون پرداخته شود. همچنین زمانی که تعداد داده‌های ازدست‌رفته زیاد باشد، از الگوریتم‌های مکاشفه‌ای نظیر الگوریتم ژنتیک استفاده شود. استفاده از خوشه‌بندی و شناسایی داده‌های مشابه در روش پیشنهادی، سبب بهبود مشکل تخمین داده‌های گم‌شده و افزایش دقت می‌شود. همچنین، الگوریتم ژنتیک در روش پیشنهادی سبب حل مشکل زمان محاسباتی بالا خواهد شد مزایای دیگر روش پیشنهادی در ادامه بیان شده است:

ـ در روش‌های موجود، برای جایگزینی دادهٔ ازدست‌رفته، از کل مجموعه داده استفاده می‌شود. این موضوع سبب در نظر گرفتن رکوردهای غیرمشابه رکورد مربوط به دادهٔ ازدسترفته خواهد شد. لذا منجر به نتایج اشتباه خواهد شد. در الگوریتم پیشنهادی، از خوشه‌بندی برای شناسایی رکوردهای مشابه، و محاسبه دادهٔ ازدست‌رفته براساس رکوردهای مشابه موجود در خوشه، استفاده شده است.

ـ در الگوریتم پیشنهادی، حذف داده‌های پرت، تعیین تعداد خوشه‌های بهینه و غیره در نظر گرفته شده است. این موضوع سبب خواهد شد، داده‌های غیرعادی تأثیری در محاسبهٔ داده‌های ازدست‌رفته نداشته باشند.

ـ در الگوریتم پیشنهادی، برای هر خوشه، صفاتی (ستون‌ها) که بیش از یک سوم دادهٔ ازدست‌رفته دارند حذف می‌شوند. این موضوع سبب جلوگیری از تأثیر داده‌های غیرقابل اطمینان در محاسبه

1. Emmanuel

داده‌های ازدست‌رفته خواهد شد.

ـ در الگوریتم پیشنهادی، از مدل رگرسیون در خوشه استفاده می‌شود که سبب می‌شود در محاسبهٔ داده‌های ازدست‌رفته، فیلدهای مربوط در صفات (ستون‌های) دیگر نیز در نظر گرفته شود.

ـ استفاده از الگوریتم ژنتیک که منجر به استفادهٔ تلفیقی از میانگین، میانه، مد و مدل رگرسیون می‌شود، سبب دستیابی به نتایج قابل قبول‌تری خواهد شد.

از محدودیت‌های روش پیشنهادی می‌توان به استفادهٔ این روش از الگوریتم‌های مکاشفه‌ای اشاره کرد که لزوماً منجر به جواب بهینه نخواهند شد. با این حال، در روش پیشنهادی این مورد تنها زمانی رخ خواهد داد که حجم داده‌های گم‌شده بیش از بیست درصد تعداد کل داده‌ها باشد. حل این مشکل می‌تواند به‌عنوان کار آینده در نظر گرفته شود.

منـابـع

رشیدی‌نژاد، آ. (۱۳۹۴). مقایسه براورد میانگین جامعه بر اساس روش‌های جانهی نسبتی در آمارگیری‌های نمونه‌ای. *نشریه بررسی‌های آمار رسمی ایران*. ۱(۸۶)، ۵۱-۶۴.

باقی یزدل، ر.، جمالی، ا.، خدایی، ا.، حبیبی، م. (۱۳۹۵). روش‌های برخورد با داده‌های گم‌شده: مزایا، معایب، رویکردهای نظری و معرفی نرم‌افزارها. *نامه آموزش عالی*. ۹(۳۳)، ۱۱-۳۷.

فصیحی، ب.، عزیزی، ح.، قلیزاده گزور، ز. (۱۴۰۰). تحلیل پوششی داده‌ها با داده‌های گمشده. *پژوهش‌های نوین در تصمیم‌گیری*، ۶(۱)، ۲۰۱-۲۲۹. https://doi.org/ 20.1001.1.24766291.1400.6.1.9.7

کاظمی، ا.، کریملو، م.، رهگذر، م. (۱۳۹۰). مروری بر داده‌های گم‌شده. *مجله مطالعات ناتوانی*. ۱(۱)،۴۷-۵۲.
https://doi.org/20.1001.1.23222840.1390.1.1.3.1

References

Abedpour, K., Hosseini Shirvani, M., & Abedpour, E. (2024). A genetic-based clustering algorithm for efficient resource allocating of IoT applications in layered fog heterogeneous platforms. *Cluster Computing*, 27(2), 1313-1331. https://doi.org/10.1007/s10586-023-04005-x

Awad, F. H., Hamad, M. M., & Alzubaidi, L. (2023). Robust classification and detection of big medical data using advanced parallel K-means clustering, YOLOv4, and logistic regression. *Life*, 13(3), 1-34. https://doi.org/10.3390/life13030691

Baghi Yazdel, R., Jamali, E., Khodaei, E., & Habibi, M. (2016). Methods of Dealing with Missing Data: Advantages, Disadvantages, Theoretical Approaches and Application of Software. *Higher Education Letter*, 9(33), 11-37. [In Persian]

Biessmann, F., Rukat, T., Schmidt, P., Naidu, P., Schelter, S., Taptunov, A., ... & Salinas, D. (2019). DataWig: Missing value imputation for tables. *Journal of Machine Learning Research*, 20(175), 1-6.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big data*, 8, 1-37. https://doi.org/10.1186/s40537-021-00516-9

Fasihi, B., Azizi, H., & Gholizadeh Gazvar, Z. (2021). Data envelopment analysis with missing data. *Modern Research in Decision Making*, 6(1), 201-229. [In Persian] https://doi.org/ 20.1001.1.24766291.1400.6.1.9.7

Heidari, J., Daneshpour, N., & Zangeneh, A. (2024). A novel K-means and K-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers. *Pattern Recognition*, 1-10. https://doi.org/10.1016/j.patcog.2024.110639

Kazemi E, Karimlo M, Rahgozar M. A review of missing data. (2011). *MEJDS*; 1(1), 47-52. https://doi.org/20.1001.1.23222840.1390.1.1.3.1 [In Persian]

Khademi Dehnavi, M., Broumandnia, A., Hosseini Shirvani, M., & Ahanian, I. (2024). A hybrid genetic-based task scheduling algorithm for cost-efficient workflow execution in heterogeneous cloud computing environment. *Cluster Computing*, 1-26. https://doi.org/10.1007/s10586-024-04468-6

Li, L., Zhou, H., Liu, H., Zhang, C., & Liu, J. (2021). A hybrid method coupling empirical mode decomposition and a long short-term memory network to predict missing measured signal data of SHM systems. *Structural Health Monitoring*, 20(4), 1778-1793. https://doi.org/10.1177/1475921720932813

Liu, H., Li, J., Wu, Y., & Fu, Y. (2019). Clustering with outlier removal. *IEEE transactions on knowledge and data engineering*, 33(6), 2369-2379. https://doi.org/10.1109/TKDE.2019.2954317

Liu, Y., Dillon, T., Yu, W., Rahayu, W., & Mostafa, F. (2020). Missing value imputation for industrial IoT sensor data with large gaps. *IEEE Internet of Things Journal*, 7(8), 6855-6867. https://doi.org/10.1109/JIOT.2020.2970467

Rashidi-nejad, A. (2015). Comparison of population average estimates based on relative bias methods in sample surveys. *Iranian Journal of Official Statistics Studies*. 1(86), 51-64. [In Persian]

Tada, M., Suzuki, N., & Okada, Y. (2022). Missing Value Imputation Method for Multiclass Matrix Data Based on Closed Itemset. *Entropy*, 24(2), 1-15. https://doi.org/10.3390/e24020286

Wang, Y., Wei, Y., Wang, X., Wang, Z., & Wang, H. (2023). A clustering-based extended genetic algorithm for the multi depot vehicle routing problem with time windows and three-dimensional loading constraints. *Applied Soft Computing*, 133, 1-10. https://doi.org/10.1016/j.asoc.2022.109922

YiFei, L., Minh, H. L., Khatir, S., Sang-To, T., Cuong-Le, T., MaoSen, C., & Wahab, M. A. (2023). Structure damage identification in dams using sparse polynomial chaos expansion combined with hybrid K-means clustering optimizer and genetic algorithm. *Engineering Structures*, 283, 1-11. https://doi.org/10.1016/j.engstruct.2023.115891

Zhang, Y., Zhou, B., Cai, X., Guo, W., Ding, X., & Yuan, X. (2021). Missing value imputation in multivariate time series with end-to-end generative adversarial networks. *Information Sciences*, 551, 67-82. https://doi.org/10.1016/j.ins.2020.11.035