


A method to solve the problem of missing data, outlier data, and noisy data to improve the performance of human and information interaction.

***Mojtaba Mazoochi** : Assistant Professor, ICT Research Institute, Tehran, Iran. (Corresponding author) mazoochi@itrc.ac.ir

Leila Rabiei : Researcher, ICT Research Institute, Tehran, Iran.

Mohammad Moradi : Ph.D. in Information Technology, ICT Research Institute, Tehran, Iran.

Received: 2022/12/14 Received in revised form: 2023/02/09 Accepted: 2023/02/12 Published online: 2023/03/03

Abstract

Introduction: Errors in data collection and failure to pay attention to data that is noisy in the collection process for any reason cause problems in data-based analysis and, as a result, wrong decision-making. Therefore, solving the problem of missing or noisy data before processing and analysis is of vital importance in analytical systems. The purpose of this paper is to provide a method to identify noisy data, outliers, and missing data and provide a suitable solution for these data.

Methods: This study is applied research. Data mining techniques including binning smoothing and regression models have been used to identify and replace outlier and noisy data.

Results: The results of the tests performed in the real environment related to the data of social networks show the proper performance of the proposed method. It has also been shown that the proposed method has higher accuracy compared to the methods of binning smoothing, average and linear regression. So that for the data related to the tweet section, the mean squared error obtained for the proposed method was equal to 0.04, the binning smoothing method was equal to 0.38, the linear regression method was equal to 0.05 and the average method was equal to 0.06.

Conclusion: The method presented in this article can initially identify outlier data through one-third and two-thirds normal, and then replace the outlier data with a linear regression model, which results in improving the performance of using and processing information and improving human-information interaction.

Keywords: Noisy Data, Outliers, Missing Data, Smoothing, Binning Method, Regression Model

Conflicts of Interest: Not reported.

Funding: It did not have a financial sponsor.

How to cite this article

APA: Mazoochi, M, Rabiei, L, Moradi, M. (2023). A method to solve the problem of missing data, outlier data, and noisy data to improve the performance of human and information interaction. *Human Information Interaction*, 9(4); 13-25. (Persian)

Vancouver: Mazoochi, M, Rabiei, L, Moradi, M. (2023). A method to solve the problem of missing data, outlier data, and noisy data to improve the performance of human and information interaction. *Human Information Interaction*, 9(4); 13-25. (Persian)



ارائه روشی برای حل مشکل داده‌های گم شده، پرت و نویزی به منظور بهبود عملکرد تعامل انسان و اطلاعات

***مجتبی مازوچی** ^{ID}: استادیار، پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران. (نویسنده مسئول) mazoochi@itrc.ac.ir

لیلا ربیعی ^{ID}: پژوهشگر، پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران.

محمد مرادی ^{ID}: دکتری مهندسی فناوری اطلاعات، پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران.

چکیده

نوع مقاله: مقاله پژوهشی

زمینه و هدف: خطا در جمع‌آوری داده‌ها و عدم توجه به داده‌هایی که در پروسه جمع‌آوری به هر دلیل دچار نویز شده‌اند باعث ایجاد اشکال در تحلیل‌های مبتنی بر داده و به تبع آن، تصمیم‌سازی‌های اشتباه می‌گردد؛ لذا رفع مشکل داده‌های گم شده و یا نویزی، قبل از انجام مراحل پردازش و تحلیل دارای اهمیت حیاتی در سامانه‌های تحلیلی است. هدف این مقاله، ارائه روشی به منظور شناسایی داده‌های نویزی، پرت و داده‌های گم شده و ارائه راهکاری مناسب برای هموارسازی این داده‌ها است.

تاریخ دریافت: ۱۴۰۰/۰۹/۲۳

تاریخ بازنگری: ۱۴۰۰/۱۱/۲۰

تاریخ پذیرش: ۱۴۰۰/۱۱/۲۳

تاریخ انتشار: ۱۴۰۰/۱۲/۱۲

روش پژوهش: این پژوهش بر مبنای هدف، از نوع کاربردی است. به منظور تحلیل داده‌ها از تکنیک‌های داده‌کاوی شامل هموارسازی پیاله‌ای و مدل رگرسیون به منظور شناسایی و جاگذاری داده‌های پرت و نویزی استفاده شده است.

نتایج: نتایج آزمایش‌های انجام شده در محیط واقعی مربوط به داده‌های شبکه‌های اجتماعی، نشان‌دهنده عملکرد مناسب روش پیشنهادی است. همچنین نشان داده شده است که روش پیشنهادی دارای دقت بالاتری در مقایسه با روش‌های هموارسازی پیاله‌ای، میانگین و رگرسیون خطی است. به طوری که برای داده‌های مربوط به بخش توثیت، میانگین مربعات خطای به دست آمده برای روش پیشنهادی برابر ۰.۰۰۴، روش هموارسازی پیاله‌ای برابر ۰.۰۳۸، روش رگرسیون خطی برابر ۰.۰۰۵ و روش جایگزینی با میانگین برابر ۰.۰۰۶ بوده است.

نتیجه‌گیری: روش ارائه شده در این مقاله، می‌تواند در ابتدا از طریق یک‌سوم و دوسوم نرمال، داده‌های پرت را شناسایی کند و سپس با مدل رگرسیون خطی به جایگزینی داده‌های پرت پردازش که در نتیجه سبب بهبود عملکرد استفاده و پردازش اطلاعات و بهبود تعامل انسان و اطلاعات خواهد شد.

کلمات کلیدی: داده‌های نویزی، داده‌های پرت، داده‌های گم شده، هموارسازی، روش پیاله‌ای، مدل رگرسیون

تعارض منافع: گزارش نشده است.

منبع حمایت‌کننده: حامی مالی نداشته است.

شبیه استناد به این مقاله

APA: Mazoochi, M, Rabiei, L, Moradi, M. (2023). A method to solve the problem of missing data, outlier data and noisy data in order to improve the performance of human and information interaction. *Human Information Interaction*, 9(4); 13-25. (Persian)

Vancouver: Mazoochi, M, Rabiei, L, Moradi, M. (2023). A method to solve the problem of missing data, outlier data and noisy data in order to improve the performance of human and information interaction. *Human Information Interaction*, 9(4); 13-25. (Persian)



انتشار مجله تعامل انسان و اطلاعات با حمایت مالی دانشگاه فوارزمی انجام می‌شود.

انتشار این مقاله به صورت دسترسی آزاد مطابق با [CC BY-NC-SA 3.0](https://creativecommons.org/licenses/by-nc-sa/3.0/) صورت گرفته است.

هدف از انجام این پژوهش، ارائه روشی مناسب برای حل مشکل داده‌های گم شده و داده‌های نویزی به‌منظور بهبود عملکرد تحلیل داده‌ها و در نتیجه بهبود تعامل انسان و اطلاعات است. در ادامه، ابتدا به‌مرور کارهای مرتبط با پژوهش و بیان ضعف‌ها و محدودیت‌های روش‌های موجود پرداخته شده است. سپس، روش پیشنهادی ارائه شده است. در بخش بعد، روش پیشنهادی پیاده‌سازی شده و مورد ارزیابی نسبت به سایر روش‌ها قرار گرفته است. در نهایت نیز نتیجه‌گیری بیان شده است.

پیشینه تحقیق

تا‌ا و همکاران^۵ (۲۰۲۲) به روش محاسبه مقدار گم شده برای داده‌های ماتریس چند کلاسه بر اساس مجموعه آیت‌های بسته پرداخته‌اند. در این پژوهش، دو روش مبتنی بر مجموعه آیت‌های بسته، *CImpute* و *ICImpute*، برای دستیابی به مقدار گم شده با استفاده از فضای ویژگی محلی برای داده‌های ماتریس چند کلاسه پیشنهاد شده است *CImpute* مقادیر گم‌شده را با استفاده از مجموعه آیت‌های بسته استخراج شده از هر کلاس تخمین می‌زند. *ICImpute* یک روش بهبودیافته *CImpute* است که در آن یک فرایند کاهش ویژگی معرفی شده است. نتایج تجربی نشان می‌دهد که کاهش ویژگی به‌طور قابل 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، *ICImpute* دقت انتساب بالاتری را ارائه می‌کند؛ اما به زمان محاسباتی بیشتری نیاز دارد. ژنگ و همکاران^۶ (۲۰۲۱) به پژوهش در رابطه با مقادیر گم شده در سری‌های زمانی چندمتغیره پرداختند. نویسندگان بیان می‌کنند که یک فاز اضافی برای بهینه‌سازی نویز تصادفی ورودی ژنراتور مورد نیاز است. علاوه بر این، مقادیر منتسب به دلیل دشواری آموزش و فرایند تولید ناپایدار، می‌توانند بسیار متفاوت از مقادیر واقعی باشند؛ بنابراین، آن‌ها یک مدل انتها به انتها برای نسبت‌دادن مقادیر از دست‌رفته در یک سری زمانی چندمتغیره پیشنهاد می‌کنند. آزمایش‌ها بر روی سه مجموعه داده‌های سری زمانی چندمتغیره در دنیای واقعی نشان می‌دهند که مدل پیشنهادی از روش‌های پیشرفته در وظایف انتساب و کاربردهای پایین‌دستی، از جمله طبقه‌بندی و رگرسیون، بهتر عمل می‌کند.

باتوجه به پیشرفت فناوری در دنیای امروز و گسترش روزافزون داده‌ها، تجزیه و تحلیل داده‌ها از اهمیت ویژه‌ای برخوردار است. این تحلیل‌ها با تمرکز بر سنجش ذائقه کاربران و نیز پیش‌بینی‌های مبتنی بر داده، کاربردهای فراوانی در حوزه‌های مختلف از جمله اقتصاد، سیاست، جامعه‌شناسی و کسب‌وکار دارند. داده‌کاوی و تجزیه و تحلیل داده‌ها روشی برای استخراج الگو و تبدیل داده به دانش است (اگاروال و یو، ۲۰۰۵). اگر نمونه‌ها نماینده خوبی از بدنه بزرگ داده نباشند، این فرایند شکست خواهد خورد (ارنینگ و همکاران، ۱۹۹۶).

پرت‌کاوی به مشکل پیدا کردن الگوها در مجموعه داده‌های بزرگ که مطابق با رفتار مورد انتظار نیستند، اشاره دارد (کننارزیک، ۲۰۱۱). در واقع پایگاه داده ممکن است شامل اشیای داده‌ای باشد که با رفتار عمومی یا مدل داده منطبق نیستند. چنین اشیای داده‌ای که به‌طور برجسته متفاوت یا در تضاد با مجموعه داده باقی‌مانده هستند، نقاط دورافتاده یا پرت نامیده می‌شوند (هان و کمبر، ۲۰۰۶). داده‌های پرت و داده‌هایی که دارای نویز هستند، در بسیاری از مجموعه داده‌ها دیده می‌شوند. به‌عنوان مثال ممکن است مدیر یک فروشگاه اینترنتی قصد تحلیل سن کاربران خود را به‌منظور دریافت این دانش که افراد در بازه سنی مختلف، بیشتر به کدام محصولات تمایل نشان می‌دهند را داشته باشد. این احتمال وجود دارد که مشتریان سه‌واً سن خود را به‌جای ۲۵ سال، ۲۵۰ سال درج کنند و یا این که شخصی به‌جای این که سن خود را درج کند، سه‌واً سال تولد خود را وارد نماید. به این دست از داده‌ها که معمولاً با باقی داده‌ها ناسازگار هستند، داده‌های پرت گفته می‌شود و مجموعه داده را دارای نویز می‌داند. نویزها که به داده‌های غیرطبیعی نیز شهرت دارند، سبب خراب شدن آمارها می‌شوند.

همچنین، این احتمال وجود دارد که در مراحل جمع‌آوری داده‌ها برخی داده‌ها از دست‌رفته و به اصطلاح گم شوند. عدم توجه به داده‌هایی که در پروسه جمع‌آوری به هر دلیل گم و یا دچار نویز شده‌اند باعث ایجاد اشکال در تحلیل‌های مبتنی بر داده و به تبع آن، تصمیم‌سازی‌های اشتباه می‌گردد؛ لذا رفع مشکل داده‌های گم شده و یا نویزی، قبل از انجام مراحل پردازش و تحلیل دارای اهمیت حیاتی در سامانه‌های تحلیلی است (کیانی و منتظری، ۱۳۹۴).

⁴ Han & Kamber

⁵ Tada et al

⁶ Zhang et al

¹ Aggarwal & Yu

² Arning et al

³ Kantardzic

شکاف، رویکرد پیشنهادی به طور 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 به طور قابل توجهی بهبود یافته است.

بیزمن و همکاران^۴ (۲۰۱۹) به پژوهشی با عنوان «مقادیر ازدست‌رفته برای جداول» پرداخته‌اند. آن‌ها بیان می‌کنند که روش‌های انتساب مقدار گم شده کنونی بر روی داده‌های عددی یا دسته‌بندی تمرکز می‌کنند و مقیاس‌بندی آن‌ها در مجموعه‌های داده با میلیون‌ها ردیف دشوار است. آن‌ها روشی را ارائه می‌کنند که می‌تواند برای جداول با انواع داده‌های ناممکن، از جمله متن بدون ساختار اعمال شود. روش پیشنهادی، استخراج‌کننده ویژگی‌های یادگیری عمیق را با تنظیم خودکار فرآیند ترکیب می‌کند. این به کاربران بدون پیش‌زمینه یادگیری ماشین، مانند مهندسان داده، امکان می‌دهد تا مقادیر گم شده را با کمترین تلاش در جدول‌هایی با انواع داده‌های ناممکن نسبت دهند. آن‌ها روش پیشنهادی را با روش‌های موجود مقایسه می‌کنند و نتایج نشان از رضایت‌بخش بودن روش ارائه شده دارد.

روش‌های بسیاری برای هموارسازی و جایگزینی داده‌های پرت و یا ناسازگار با سایر داده‌ها وجود دارد. یکی از روش‌های مرسوم، استفاده از هموارسازی پیمانه‌ای است که در بسیاری مواقع نسبت به جایگزینی با میانگین، عملکرد بهتری دارد (سادیک و گرونوالد^۵، ۲۰۱۰). مشکل این روش این است که علاوه بر از بین بردن خطا روند، ممکن است روی داده‌های واقعی نیز تأثیر بگذارد. روش ماشین بردار پشتیبان نیز یکی دیگر از راه‌های برازش داده‌های پرت است (هنگهای و همکاران^۶، ۲۰۰۵). اما این روش هنگام حذف هر داده، تمام ویژگی‌های آن را نیز حذف می‌کند درحالی‌که ممکن است برخی از ویژگی‌ها مشمول داده پرت نباشند. یکی دیگر از راه‌های جایگزینی مقادیر پرت وقتی که چند دسته داده با ارتباط معنادار میان آن‌ها وجود داشته باشد، استفاده از برازش رگرسیون خطی است (ژو و همکاران^۶، ۲۰۰۳). مشکل این روش این است که همه داده‌ها را روی یک خط راست برازش می‌کند و لذا برای حالتی که داده‌ها نوسان زیادی دارند، تخمین مناسبی به حساب نمی‌آید. یک روش مرسوم، حذف داده‌های پرت است؛ اما این روش به‌ویژه زمانی که تعداد داده‌های پرت زیاد باشند موجب خدشه در روند آماری تحلیل‌ها می‌شود. یک روش دیگر، جایگزین کردن داده‌های پرت با میانگین سایر

در پژوهش لی و همکاران^۱ (۲۰۲۱)، تجزیه حالت تجربی با یک شبکه یادگیری عمیق حافظه کوتاه‌مدت برای بازیابی داده‌های سیگنال اندازه‌گیری شده ترکیب شده است. روش ترکیبی پیشنهادی وظیفه انتساب داده‌های گم شده را به‌عنوان یک کار پیش‌بینی سری زمانی تبدیل می‌کند که سپس با یک استراتژی «تقسیم و غلبه» حل می‌شود. مفهوم اصلی این استراتژی، پیش‌بینی دنباله‌های داده‌های سیگنال اندازه‌گیری شده خام است که به‌جای پیش‌بینی مستقیم، با تجزیه حالت تجربی تجزیه می‌شوند، زیرا تجزیه می‌تواند به مدل‌سازی تغییرات دوره‌ای نامنظم داده‌های سیگنال اندازه‌گیری شده کمک کند. علاوه بر این، شبکه حافظه کوتاه‌مدت و بلندمدت در مدل ترکیبی می‌تواند همبستگی‌های دوربرد بیشتری از دنباله‌های فرعی را نسبت به شبکه عصبی مصنوعی سنتی به‌خاطر بسپارد. سه مدل پیش‌بینی پرکاربرد، یعنی میانگین متحرک یکپارچه اتورگرسیون، رگرسیون بردار پشتیبان و مدل‌های شبکه عصبی مصنوعی نیز به‌عنوان مدل‌های معیار پیاده‌سازی می‌شوند. داده‌های شتاب خام جمع‌آوری شده از یک پل کابلی برای ارزیابی عملکرد روش پیشنهادی برای ازدست‌رفتن داده سیگنال اندازه‌گیری شده استفاده می‌شود. نتایج بازیابی داده‌های سیگنال اندازه‌گیری شده نشان می‌دهد که روش ترکیبی پیشنهادی، عملکرد عالی را نشان می‌دهد.

لیو و همکاران^۲ (۲۰۲۰) در پژوهش خود بیان می‌کنند که داده‌های ازدست‌رفته یکی از بزرگ‌ترین مشکلات برای پیش‌پردازش داده‌ها در معماری اینترنت اشیا است. به دلیل جمع‌آوری داده‌های حسگر با فرکانس بالا، داده‌های ازدست‌رفته در اینترنت اشیا چالش‌های جدیدی را به همراه دارد. برای پرداختن به این موضوع، نویسندگان بر روی انتساب داده‌های گم شده برای شکاف‌های بزرگ در داده‌های سری زمانی تک‌متغیره تمرکز می‌کنند و یک چارچوب تکراری با استفاده از تکرار شکاف چندگانه به نام *Itr-MS-STLeImp* برای ارائه مناسب‌ترین مقادیر پیشنهاد شده است. شکاف ابتدا به چند قطعه تقسیم می‌شود تا فرایند انتساب مقدار گم شده را مقداردهی اولیه کند و سپس، به طور مکرر بازسازی شکاف و الحاق شکاف برای به‌دست‌آوردن نتایج انتساب نهایی اجرا می‌شود. روش پیشنهادی با استفاده از داده‌های حسگر جمع‌آوری شده از کارخانه‌های تولید واقعی در استرالیا بررسی شده است و نتایج مقایسه نشان می‌دهد که روش پیشنهادی از نظر خطای ریشه میانگین مربع از روش‌های پیشرفته بهتر عمل می‌کند. در شرایط مختلف طول

⁴ Sadik & Gruenwald

⁵ Honghai et al

⁶ Zhou et al

¹ Li et al

² Liu et al

³ Biessmann et al

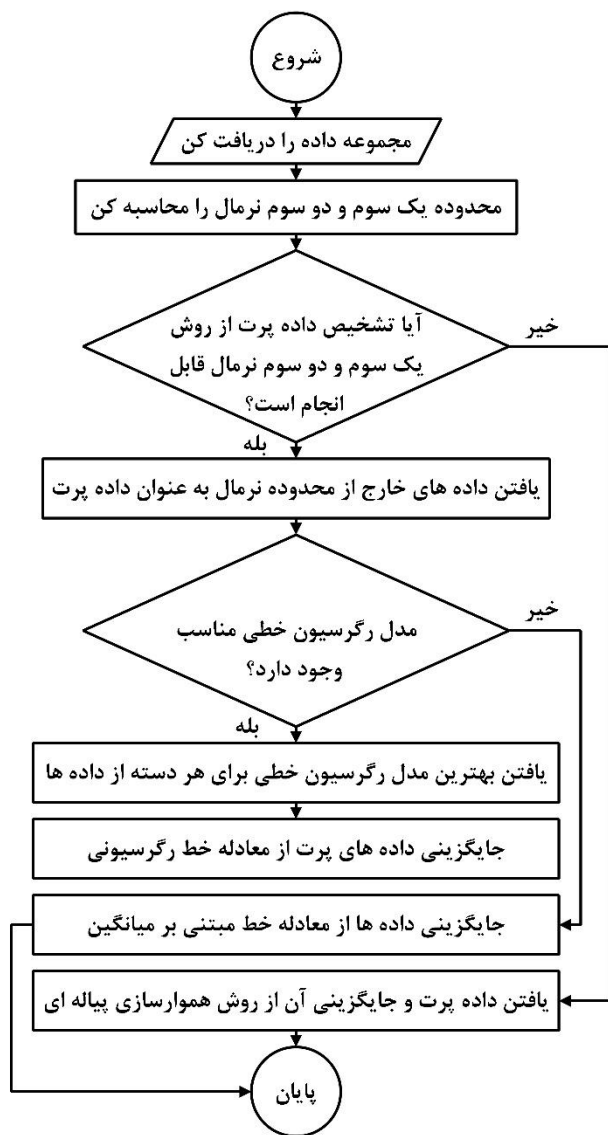
روش پژوهش

شکل ۱، ساختار کلی روش پیشنهادی را نشان می‌دهد. همان طوری که در این شکل مشاهده می‌شود، روش پیشنهادی شامل دو بخش اصلی است:

۱. شناسایی داده‌های از دست‌رفته و داده‌های پرت
۲. هموارسازی داده‌ها

در ادامه، هر کدام از این مراحل توضیح داده شده است.

داده‌هاست (ترویانسکایا و همکاران^۱، ۲۰۰۱). این روش با این که در بسیاری از مواقع، کارآمد است؛ اما موجب اختلال در توزیع و کم برآورد شدن واریانس می‌شود؛ لذا چنانچه بیان شد، هر کدام از روش‌های موجود برای هموارسازی و جایگزینی داده‌های نویزی دارای اشکالات و محدودیت‌هایی است. در این پژوهش سعی شده است، روشی پیشنهاد شود که بر محدودیت‌های روش‌های موجود فائق آید و همچنین خطای کم‌تری داشته باشد.



شکل ۱. فلوچارت روش پیشنهادی

¹ Troyanskaya et al

برابر انحراف معیار داده‌ها و حد پایین را مقدار میانگین منهای سه برابر انحراف معیار داده‌ها در نظر می‌گیرند و داده‌های خارج از این محدوده به‌عنوان داده پرت در نظر گرفته می‌شوند. اما در حالتی که انحراف معیار عدد بزرگی باشد ممکن است حد پایین صفر یا منفی شود که در این صورت برای حد پایین از دو سوم نرمال یعنی مقدار میانگین منهای دو برابر انحراف معیار داده‌ها و یا از یک‌سوم نرمال یعنی مقدار میانگین منهای انحراف معیار داده‌ها استفاده می‌کنیم. به عبارتی، سه مقدار برای حدود بالایی و سه مقدار برای حدود پایینی داده‌ها داریم که با کنترل داده‌ها می‌توان دریافت که بیشتر داده‌ها بین کدام دو کران قرار دارند و نهایتاً داده‌های ناسازگار را شناسایی کرد. روش کار به این صورت است که ابتدا میانگین همه داده‌ها در بازه شش ماه قبل و شش ماه بعد از هر داده محاسبه شده و منهای انحراف معیار آن‌ها خواهد شد. اگر این مقدار مثبت شود تمام داده‌های کمتر از آن، به‌عنوان داده پرت در نظر گرفته می‌شود. اما در حالتی که این مقدار باز هم صفر یا منفی باشد برای شناسایی داده پرت، داده‌ها را به چندین پیاله تقسیم می‌کنیم و پیاله پایینی را داده پرت در نظر می‌گیریم. در شکل ۲، کران‌های بالا و پایین برای شناسایی داده پرت با خطوط قرمز مشخص شده‌اند.

شناسایی داده‌های ازدست‌رفته و داده‌های پرت

معمولاً به دلیل ایجاد مشکلات فنی در فرایند جمع‌آوری داده‌ها از قبیل ایجاد مشکل در خزشگرها، شبکه و یا ذخیره‌سازی داده‌ها، ممکن است بخشی از داده‌ها از دست برود. در این حالت، تعداد داده به‌دست‌آمده بسیار کمتر از تعداد داده مورد انتظار خواهد بود. شناسایی داده‌های ازدست‌رفته می‌تواند با بررسی کردن داده‌های جمع‌آوری شده و شناسایی فیلدهای خالی انجام شود.

شناسایی داده‌های پرت

ممکن است به دلایل فنی، بخشی از داده‌ها اشتباه بوده و با سایر داده‌ها ناسازگار باشد. شناسایی این نوع از داده‌ها به‌راحتی انجام نمی‌شود. بدین منظور داده‌های خارج از محدوده نرمال شناسایی می‌شوند. حتی اگر توزیع داده‌ها به‌صورت نرمال نباشد؛ اما محدوده نرمال می‌تواند محدوده مناسبی برای شناسایی داده‌های پرت باشد. حد بالایی محدوده نرمال را مقدار میانگین به‌علاوه سه



شکل ۲. کران‌های بالا و پایین برای شناسایی داده پرت

از رگرسیون، ابتدا میان چند دسته از داده‌ها که دوه‌دو رابطه معناداری با هم دارند مدل رگرسیونی خطی در نظر می‌شود و

سپس برای برازش داده پرت، معادله خطی انتخاب می‌شود که بیشترین معیار مربعات R را داشته باشد. برای مثال، در بستر اینستاگرام، بین تعداد پست‌ها، تعداد لایک‌ها و تعداد کامنت‌ها رابطه معناداری وجود دارد. اگر همه معیارهای مربعات R در

هموارسازی داده‌ها

در روش پیشنهادی، ابتدا به‌صورت پیش‌فرض از برازش داده‌ها با رگرسیون خطی استفاده می‌شود. اما در شرایط مختلف که این روش کارایی کمتری دارد به‌تناسب از روش‌های دیگری با خط مبتنی بر میانگین و هموارسازی پیاله‌ای استفاده خواهد شد. در استفاده

معادله خط مبتنی بر میانگین که حالت بهبودیافته‌ی جهانی داده پرت با میانگین است استفاده می‌کنیم. این روش در بخش تست و ارزیابی به تفصیل بیان شده است. در شکل ۳، داده‌های سمت چپ جدول که به صورت رنگی نشان داده شده‌اند، داده‌های نویزی شناسایی شده و داده‌های هم‌رنگ آن‌ها در سمت راست جدول از معادله خط L که همان خط مبتنی بر میانگین است به دست آمده‌اند.

انتخاب‌های دوبه‌دو از میان دسته‌های مختلف، از مقدار $0/5$ معینی کم‌تر باشند، بیانگر این است که بین هیچ کدام از دسته‌های مختلف داده‌ها، برازش خطی کارایی وجود ندارد. همچنین در حالتی که هر دو داده متقابل در دودسته منتخب برای مدل رگرسیونی داده پرت باشند نمی‌توان از این مدل برای برازش استفاده کرد. ضمناً ممکن است داده برازش شده از خط رگرسیونی کم‌تر از داده اولیه باشد که مطلوب نیست. برای سه حالت اخیر از

| | | | | | | | |
|------------|----------|-------------|-------------|------------|----------|-------------|----------|
| 140 | 4379 | 1379541 | 54411 | 1400-01-21 | 4379 | 1379541 | 54411 |
| 141 | 2207 | 996533 | 38336 | 1400-01-22 | 2207 | 996533 | 38336 |
| 142 | 5621 | 1220302 | 54154 | 1400-01-23 | 5621 | 1220302 | 54154 |
| 143 | 897 | 392344 | 20140 | 1400-01-24 | 897 | 1328328.622 | 57190 |
| 144 | 326 | 693281 | 22059 | 1400-01-25 | 1388.592 | 1654243.393 | 59396.85 |
| 145 | 988 | 551593 | 19023 | 1400-01-26 | 988 | 1500795.289 | 55905.45 |
| 146 | 1578 | 644379 | 32743 | 1400-01-27 | 5384.976 | 1601282.527 | 71683.45 |
| 147 | 363 | 306259 | 17965 | 1400-01-28 | 1506.696 | 1235098.567 | 54688.75 |
| 148 | 153 | 201563 | 6886 | 1400-01-29 | 836.376 | 1121712.799 | 41947.9 |
| 149 | 277 | 323532 | 14584 | 1400-01-30 | 1232.184 | 1253805.226 | 50800.6 |
| 150 | 277 | 338241 | 14708 | 1400-01-31 | 1232.184 | 1269735.073 | 50943.2 |
| 151 | 1748 | 532980 | 21717 | 1400-02-01 | 1748 | 1480637.41 | 59003.55 |
| 152 | 20 | 29599 | 1242 | 1400-02-02 | 411.84 | 935475.787 | 35457.3 |
| 153 | 8 | 6728 | 144 | 1400-02-03 | 373.536 | 910706.494 | 34194.6 |
| MEAN | 2356.162 | 2567984.701 | 96460.01299 | | | | |
| STD | 1982.402 | 1657275.958 | 62265.16154 | | | | |
| MEAN- STD | 373.7608 | 910708.7434 | 34194.85145 | | | | |
| MEAN+3*STD | 6320.965 | | | | | | |

شکل ۳. داده‌های جایگزین شده با معادله خط L

پردازش و تحلیل داده‌ها و مصورسازی است. یکی از مشکلات مهم و رایج در تحلیل شبکه‌های اجتماعی که باعث ایجاد اشتباهات و انحرافات فراوان در تحلیل نتایج و در نتیجه تصمیم‌سازی‌های غلط می‌گردد فقدان بخشی از داده‌ها و وجود داده‌های نویزی به دلیل سیاست‌های متولیان بسترهای این شبکه‌ها و همچنین ایجاد مشکلات فنی در فرایند جمع‌آوری داده‌ها است. بستر ذکاوت هم از این قاعده مستثنی نیست. فقدان بخشی از داده‌ها در هنگام جمع‌آوری در برخی از بازه‌های زمانی، باعث می‌گردد که اطلاعات آماری مستخرج از آن‌ها، نویزی شده و عملاً تحلیل‌های آماری، دارای اشتباهات و انحرافات زیاد شوند. در این بخش، تأثیر روش پیشنهادی برای رفع این مشکل بررسی شده است.

ابتدا داده‌های نویزی شناسایی و مقدار صفر برای آن‌ها لحاظ شد تا توسط الگوریتم هموارسازی مقدار مناسب آن جایگزین گردد. در مرحله اول هموارسازی در بازه‌های زمانی یک‌ماهه روی بسترهای اینستاگرام، تلگرام و توییتر پیاده‌سازی شد. در اینستاگرام، برای هر ماه، تعداد کل پست‌ها، میانگین تعداد لایک

به‌علاوه همان‌طور که در بخش شناسایی داده پرت گفته شد گاهی ممکن است یافتن داده‌های پرت از محدوده‌های نرمال میسر نباشد که در این حالت، از تقسیم‌بندی داده‌ها به چندین پیاله استفاده می‌شود. برای جانهی داده‌ها در این مواقع، چون پراکنده‌ی داده‌ها زیاد است، کل داده شش‌ماهه را به ده پیاله تقسیم می‌کنیم. سپس تمام داده‌های پایین‌ترین دهک را به‌عنوان داده پرت لحاظ کرده و همه داده‌های پایین‌ترین دهک را با ماکزیمم همان پیاله جایگزین می‌کنیم.

پیاده‌سازی و ارزیابی روش پیشنهادی

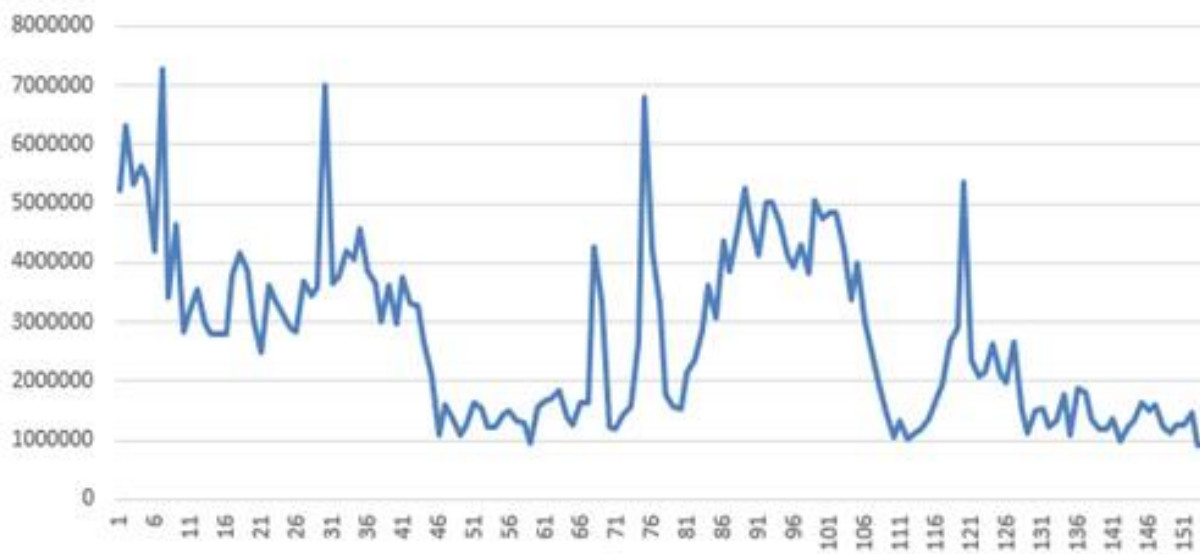
به‌منظور پیاده‌سازی و ارزیابی روش پیشنهادی، از بستر ذکاوت که با همکاری نویسندگان این مقاله و جمعی از پژوهشگران دیگر در راستای ذائقه‌سنجی و تحلیل کاربران شبکه‌های اجتماعی توییتر، تلگرام و اینستاگرام، توسعه داده شده است، استفاده گردید. این بستر دارای اجزای مختلف از قبیل جمع‌آوری داده‌های شبکه‌های اجتماعی، یکپارچه‌سازی داده‌ها، پیش‌پردازش،

نویزی از روش هموارسازی پباله‌ای شناسایی شده بود مقادیر صفر با نزدیک‌ترین مقدار یعنی کوچک‌ترین داده غیرصفر جایگزین شد. اما در بررسی بیشتر مشخص شد همه داده‌ها در طول یک ماه نویزی نبوده‌اند و فقط در برخی از روزها جمع‌آوری داده دچار مشکل شده است؛ بنابراین در این مرحله، مجموعه داده‌ها به صورت روزانه در نظر گرفته شد. داده‌های نویزی مانند مرحله اول شناسایی و در صورت امکان، با رگرسیون خطی جایگزین شدند. در برخی از موارد، داده‌هایی که با میانگین جایگزین شدند در چند روز متوالی اتفاق افتاده بودند. این امر باعث می‌شد تا نموداری که روند انتشار پست‌ها یا سایر متغیرها را نشان می‌دهد در این بازه‌ها به صورت خط مستقیم درآید. یک نمونه از روند انتشار قبل و بعد از هموارسازی به ترتیب در شکل ۴ و شکل ۵ نشان داده شده است.

و کامنت، تعداد پست‌های شامل یک هشتگ خاص، موضوع خاص و میانگین لایک و کامنت آن هشتگ، هر کدام به عنوان یک متغیر در نظر گرفته شده و در ستون‌های مجزا قرار گرفتند. برای بستر توئیتر نیز تعداد کل پست‌ها، تعداد ری‌توییت‌ها، متوسط لایک و کامنت، تعداد توئیٹ شامل یک هشتگ خاص، موضوع خاص، تعداد ری‌توییت و تعداد لایک و کامنت آن در نظر گرفته شد. سپس رگرسیون خطی میان تمام ستون‌ها به دست آمد و معادله رگرسیون خطی که بیشترین معیار مربعات R را داشت برای جایگزینی صفرهای ایجاد شده استفاده شد. در حالتی که هیچ یک از معیارهای R مقداری بیشتر از ۵۰ درصد نداشت و نیز در حالتی که دو داده متقابل در دو ستونی که رگرسیون روی آن‌ها نوشته شده هر دو صفر شوند، مقادیر صفر را با مقدار میانگین داده‌های غیرصفر جایگزین کردیم. همچنین در حالتی که داده



شکل ۴. روند انتشار تعداد لایک‌های اینستاگرام قبل از هموارسازی



شکل ۵. روند انتشار تعداد لایک‌های اینستاگرام بعد از هموارسازی

$$y = \frac{y_n - y_1}{x_n - x_1} (x - x_1) + y_1 \quad (1)$$

به منظور بررسی عملکرد روش پیشنهادی، یک مجموعه داده واقعی دوماهه بدون نویز در نظر گرفته شد. سپس ۱۰ درصد از داده‌ها، به صورت تصادفی با ضریب ۰/۵ دست‌کاری شده و روش یافتن داده نویزی روی این مجموعه داده جدید آزمایش شد. با توجه به محدودیت فضا، بخشی از این داده در شکل ۶ نمایش داده شده است.

برای جلوگیری از این مشکل و برای حفظ روند صعود یا نزول هر متغیر، برای جایگزینی صفرها از معادله خط مثبتی بر میانگین استفاده شد. به این صورت که ابتدا تمام داده‌های نویزی شناسایی شدند. سپس کم‌ترین و بیشترین آن‌ها، یعنی x_1 و x_n تعیین شدند. مقدار y_1 کران پایینی شناسایی داده پرت یعنی تفاضل میانگین و انحراف معیار قرار داده شد. مقدار y_n میانگین هندسی کل داده‌ها در نظر گرفته شد. معادله خط واصل بین (x_1, y_1) و (x_n, y_n) از رابطه (۱) محاسبه شد که در این معادله، x داده نویزی اولیه و y داده جایگزین است.

| | | | | | |
|----|---------------|-------------|----|-----------------|-------------|
| 10 | 1399-07-10 | 636944 | 10 | 1399-07-10 | 636944 |
| 11 | 1399-07-11 | 307764 | 11 | 1399-07-11 | 307764 |
| 12 | 1399-07-12 | 659217 | 12 | 1399-07-12 | 659217 |
| 13 | 1399-07-13 | 651665 | 13 | 1399-07-13 | 651665 |
| 14 | 1399-07-14 | 696065 | 14 | 1399-07-14 | 696065 |
| 15 | 1399-07-15 | 666840 | 15 | 1399-07-15 | 666840 |
| 16 | 1399-07-16 | 738398 | 16 | 1399-07-16 | 738398 |
| 17 | 1399-07-17 | 658697 | 17 | 1399-07-17 | 658697 |
| 18 | 1399-07-18 | 312766 | 18 | 1399-07-18 | 312766 |
| 19 | 1399-07-19 | 619436 | 19 | 1399-07-19 | 619436 |
| 20 | 1399-07-20 | 671312 | 20 | 1399-07-20 | 671312 |
| 21 | 1399-07-21 | 663739 | 21 | 1399-07-21 | 663739 |
| 22 | 1399-07-22 | 662170 | 22 | 1399-07-22 | 662170 |
| 23 | 1399-07-23 | 651952 | 23 | 1399-07-23 | 651952 |
| 24 | 1399-07-24 | 319195 | 24 | 1399-07-24 | 319195 |
| 25 | 1399-07-25 | 596779 | 25 | 1399-07-25 | 596779 |
| 26 | 1399-07-26 | 711513 | 26 | 1399-07-26 | 711513 |
| 27 | 1399-07-27 | 758291 | 27 | 1399-07-27 | 758291 |
| 28 | 1399-07-28 | 747862 | 28 | 1399-07-28 | 747862 |
| 29 | 1399-07-29 | 892420 | 29 | 1399-07-29 | 892420 |
| 30 | 1399-07-30 | 964481 | 30 | 1399-07-30 | 964481 |
| 31 | AVERAGE | 640262.4333 | 31 | AVERAGE | 640262.4333 |
| 32 | STD.S | 137358.6961 | 32 | STD.S | 137358.6961 |
| 33 | AVERAGE-STD.S | 502903.7372 | 33 | AVERAGE-2×STD.S | 365545.0411 |

شکل ۶. بخشی از داده‌های نویزی تشخیص داده شده در مجموعه داده مورد مطالعه با روش ارائه شده

نویزی زیاد شده و بیشتر از ده درصد از کل داده‌ها باشد، از کارایی این روش کاسته می‌شود. پس از آزمودن روش یافتن داده‌های نویزی، به بررسی عملکرد روش پیشنهادی برای جایگزینی داده‌های نویزی می‌پردازیم. همان‌طور که در روش پیشنهادی بیان شد در صورت وجود رگرسیون خطی مناسب با معیار مربعات R بالاتر از $0/5$ ، داده‌ها با رگرسیون خطی جایگزین می‌شوند. در غیر این صورت، جایگزینی باتوجه به خط مبتنی بر میانگین و یا هموارسازی پیاله‌ای انجام می‌شود. برای بررسی عملکرد روش پیشنهادی، ابتدا یک مجموعه داده واقعی بدون نویز که مربوط به داده‌های توییت‌ر مهر و آبان ۱۳۹۹ است، در نظر گرفته شد. ۱۰ درصد از داده‌ها به صورت تصادفی نویز داده شدند. سپس جایگزینی داده‌های نویزی با استفاده از روش‌های هموارسازی پیاله‌ای، میانگین داده‌ها، رگرسیون خطی و همچنین روش پیشنهادی در این مقاله انجام شد. به‌عنوان نمونه، در شکل ۷ داده‌های نویزی با استفاده از روش پیشنهادی جایگزین شده‌اند. داده‌های جایگزین شده با رنگ آبی در این شکل مشخص شده‌اند.

در شکل ۶، داده‌های زردرنگ، داده‌هایی هستند که به صورت تصادفی نویزی شده‌اند و داده آبی‌رنگ، کران پایین برای داده غیر نویزی را نشان می‌دهد. در بخش سمت راست شکل ۶، برای کران پایین از میانگین منهای انحراف معیار و در بخش سمت چپ، از میانگین منهای دو برابر انحراف معیار استفاده شده است که در هر دو مورد، شناسایی داده‌های نویزی به درستی انجام گرفته است. لازم به ذکر است که ضریب یک و دو که بر روی انحراف معیار اعمال شده، بستگی به پراکندگی داده‌ها دارد و باید بررسی شود که اکثریت داده‌ها در کدام یک از بازه‌های نرمال که در بخش روش پیشنهادی معرفی شد قرار دارند. ضمناً طبق تجارب به دست آمده، ضریب پنجاه درصد که در این پژوهش برای ایجاد داده نویزی استفاده شده به مراتب بیشتر از آن چیزی است که در عمل روی داده‌های واقعی اتفاق می‌افتد و این بیانگر این مطلب است که در نویزهایی که کاهش محسوس تری روی داده ایجاد می‌کنند، این روش برای یافتن داده نویزی همچنان کارا خواهد بود. در ادامه، یک نویز ۷۰ روی ده درصد از داده‌های همین مجموعه داده، اعمال شد. نتایج نشان داد که همچنان روش پیشنهادی کارا است. لازم به ذکر است که اگر تعداد داده‌های

| shdate | count(*) | sum(like | sum(com | sum(copi | shdate | count(*) | sum(likes | sum(comments | sum(copies) |
|------------|----------|----------|---------|----------|------------|----------|-----------|--------------|-------------|
| 1399-07-01 | 627835 | 7641542 | 677278 | 682071 | 1399-08-01 | 868055 | 7391655 | 1368460.72 | 691789 |
| 1399-07-02 | 598443 | 6206703 | 638478 | 574105 | 1399-08-02 | 780309 | 6370673 | 644346 | 474896 |
| 1399-07-03 | 570534 | 6195023 | 445897 | 455125 | 1399-08-03 | 915292 | 6333758 | 784818 | 455705 |
| 1399-07-04 | 626976 | 7024088 | 689473 | 744683 | 1399-08-04 | 936144 | 6342516 | 791850 | 598823 |
| 1399-07-05 | 618325 | 6177121 | 650874 | 370499 | 1399-08-05 | 959373 | 6367915 | 880607 | 402816 |
| 1399-07-06 | 641865 | 9562344 | 566189 | 1340446 | 1399-08-06 | 949754 | 6378741 | 854941 | 434746 |
| 1399-07-07 | 526219 | 7276575 | 710957 | 815281 | 1399-08-07 | 947583 | 5983693 | 820995 | 344251 |
| 1399-07-08 | 648248 | 6868799 | 696674 | 680869 | 1399-08-08 | 889461 | 6382950 | 824411 | 593416 |
| 1399-07-09 | 625236 | 6255763 | 659553 | 445237 | 1399-08-09 | 949311 | 6024357 | 1335492.22 | 540918 |
| 1399-07-10 | 636944 | 7347858 | 716346 | 715309 | 1399-08-10 | 971298 | 6238731 | 845321 | 516733 |
| 1399-07-11 | 615528 | 8792936 | 878943 | 1044204 | 1399-08-11 | 753427 | 6810216 | 889109 | 592451 |
| 1399-07-12 | 659217 | 9317304 | 866396 | 879730 | 1399-08-12 | 988663 | 6410562 | 833442 | 596025 |
| 1399-07-13 | 651665 | 6880148 | 782858 | 482993 | 1399-08-13 | 980563 | 6564640 | 829135 | 403334 |
| 1399-07-14 | 696065 | 7337649 | 739806 | 635748 | 1399-08-14 | 1103120 | 8212112 | 936133 | 475949 |
| 1399-07-15 | 528994 | 7953498 | 829496 | 773184 | 1399-08-15 | 943760 | 7371177 | 839755 | 771296.86 |
| 1399-07-16 | 738398 | 1.34E+07 | 753428 | 3283430 | 1399-08-16 | 1018486 | 7623558 | 884186 | 543726 |
| 1399-07-17 | 658697 | 1.53E+07 | 1329164 | 2910053 | 1399-08-17 | 993523 | 7157405 | 894313 | 409951 |
| 1399-07-18 | 625532 | 7457920 | 675548 | 545539 | 1399-08-18 | 959347 | 5479251 | 753358 | 362575 |
| 1399-07-19 | 619436 | 1.54E+07 | 1216929 | 2855663 | 1399-08-19 | 738398 | 5197907 | 816933 | 390504 |
| 1399-07-20 | 671312 | 3.76E+06 | 1067032 | 120814 | 1399-08-20 | 1009438 | 5960716 | 828097 | 450282 |
| 1399-07-21 | 663739 | 2.41E+07 | 1711460 | 5637413 | 1399-08-21 | 942125 | 5952368 | 795123 | 392040 |
| 1399-07-22 | 662170 | 1.05E+07 | 854167 | 1474092 | 1399-08-22 | 923169 | 6848392 | 1697474.26 | 677195.56 |
| 1399-07-23 | 651952 | 8504318 | 838455 | 813544 | 1399-08-23 | 959226 | 6347570 | 786535 | 355037 |
| 1399-07-24 | 638390 | 3.14E+06 | 1031598 | 1684202 | 1399-08-24 | 939518 | 8210904 | 868070 | 619166 |
| 1399-07-25 | 596779 | 7332689 | 663417 | 873225 | 1399-08-25 | 738398 | 7289412 | 867362 | 478111 |
| 1399-07-26 | 711513 | 1.29E+07 | 1118875 | 1955421 | 1399-08-26 | 944048 | 6913326 | 849862 | 484271 |
| 1399-07-27 | 738398 | 9448569 | 843738 | 1005405 | 1399-08-27 | 936620 | 8517277 | 922928 | 833761 |
| 1399-07-28 | 747862 | 1.98E+07 | 1301373 | 3331408 | 1399-08-28 | 970427 | 8452914 | 957453 | 806700 |
| 1399-07-29 | 892420 | 2.63E+07 | 1756449 | 4185377 | 1399-08-29 | 920701 | 7923406 | 883409 | 736119 |
| 1399-07-30 | 964481 | 2.49E+07 | 1692083 | 5312499 | 1399-08-30 | 954370 | 8458948 | 936657 | 794518 |

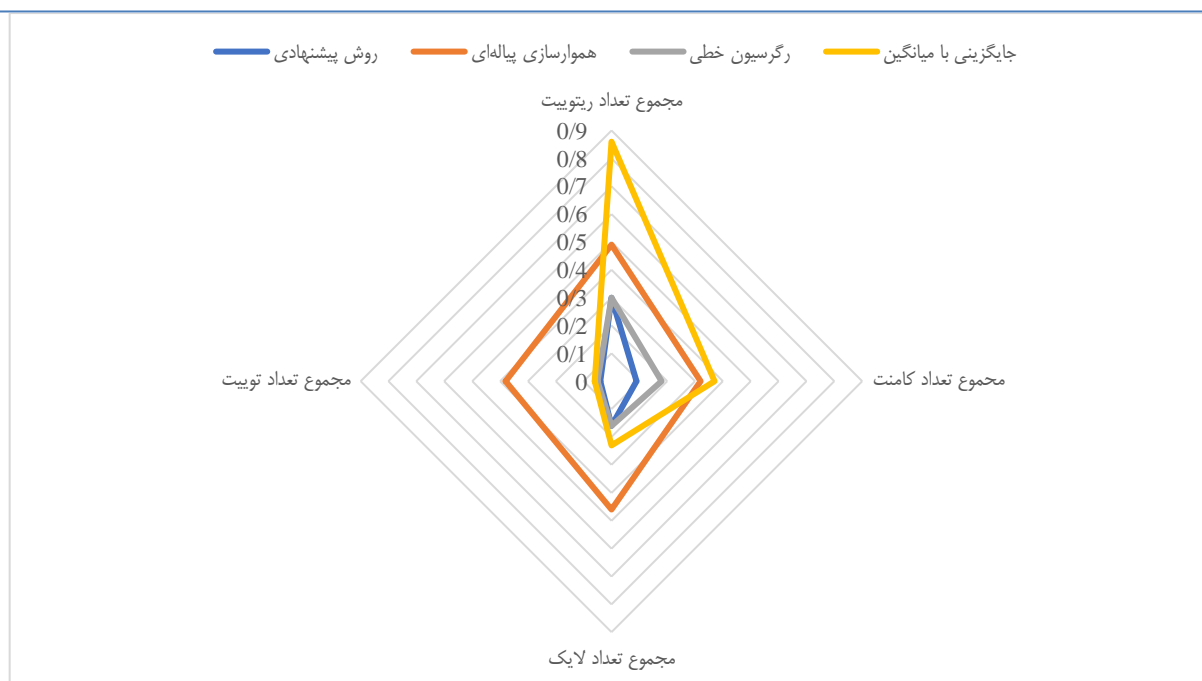
شکل ۷. جایگزینی داده‌های نویزی با استفاده از روش پیشنهادی

مربعات خطای به دست آمده در این چهار روش را نشان می‌دهند. همان طوری که مشاهده می‌شود، روش پیشنهادی در این مقاله، همواره عملکرد بهتری نسبت به سایر روش‌ها داشته است.

به منظور مقایسه عملکرد روش‌های هموارسازی پیاپله‌ای، میانگین داده‌ها، رگرسیون خطی و همچنین روش پیشنهادی، از میانگین مربعات خطا استفاده شد. برای این منظور، میانگین مربعات خطای داده‌های اصلی قبلی از نویزی شدن و داده‌های جایگزین شده با استفاده از هر روش، محاسبه شد. جدول ۱ و شکل ۸، میانگین

جدول ۱. مقایسه عملکرد روش‌های هموارسازی پیاپله‌ای، میانگین داده‌ها، رگرسیون خطی و روش پیشنهادی، در تشخیص و جایگزینی داده‌های نویزی بر اساس معیار میانگین مربعات خطا

| روش | مجموع تعداد ری توییت | مجموع تعداد کامنت | مجموع تعداد لایک | مجموع تعداد توییت |
|---------------------|----------------------|-------------------|------------------|-------------------|
| روش پیشنهادی | ۰.۳۰ | ۰.۰۹ | ۰.۱۶ | ۰.۰۴ |
| هموارسازی پیاپله‌ای | ۰.۴۹ | ۰.۳۲ | ۰.۴۶ | ۰.۳۸ |
| رگرسیون خطی | ۰.۳۰ | ۰.۱۸ | ۰.۱۶ | ۰.۰۵ |
| جایگزینی با میانگین | ۰.۸۶ | ۰.۳۷ | ۰.۲۳ | ۰.۰۶ |



شکل ۸. نمودار مقایسه عملکرد روش‌های هموارسازی پیاله‌ای، میانگین داده‌ها، رگرسیون خطی و روش پیشنهادی، در تشخیص و جایگزینی داده‌های نویزی بر اساس معیار میانگین مربعات خطا

نتیجه‌گیری

از بین بردن خط روند، ممکن است روی داده‌های واقعی نیز تأثیر بگذارد.

در این پژوهش، روشی جامع برای شناسایی و جایگزینی داده‌های پرت با بهره‌گیری از قابلیت‌های روش‌های موجود در موقعیت‌های مختلف ارائه شد. روش پیشنهادی، در برگزیده شناسایی داده‌های از دست‌رفته و داده‌های پرت و همچنین هموارسازی داده‌ها است. در بخش شناسایی داده‌های پرت، ابتدا میانگین همه داده‌ها محاسبه شده و از انحراف معیار آن‌ها کم می‌شود. اگر این مقدار مثبت بود تمام داده‌های کمتر از آن، به‌عنوان داده پرت در نظر گرفته می‌شود. اما در حالتی که این مقدار باز هم صفر یا منفی باشد برای شناسایی داده پرت، داده‌ها به چندین پیاله تقسیم می‌شوند و پیاله پایینی، داده پرت در نظر گرفته می‌شود. در بخش هموارسازی، از برازش داده‌ها با رگرسیون خطی استفاده می‌شود. اما در شرایط مختلف که این روش کارایی کمتری دارد به‌تناسب از روش‌های دیگری با خط مبتنی بر میانگین و هموارسازی پیاله‌ای استفاده می‌شود.

به‌منظور ارزیابی روش پیشنهادی، از بستر ذکاوت که با همکاری نویسندگان این مقاله و جمعی از پژوهشگران دیگر به‌منظور ذخیره‌سازی و تحلیل کاربران شبکه‌های اجتماعی توییتر، تلگرام و اینستاگرام، توسعه داده شده است و داده‌های واقعی جمع‌آوری شده از شبکه‌ها توسط این بستر، استفاده گردید. نتایج به‌دست‌آمده نشان‌دهنده عملکرد مناسب روش پیشنهادی است. همچنین مقایسه میانگین مربعات خطای هموارسازی داده‌های نویزی با

یکی از مشکلات مهم و رایج در تحلیل داده‌ها که باعث ایجاد اشتباهات و انحرافات زیاد در نتایج تحلیل و در نتیجه، تصمیم‌سازی‌های غلط می‌گردد فقدان بخشی از داده‌ها و وجود داده‌های نویزی به دلیل ایجاد مشکلات فنی در فرایند جمع‌آوری داده‌ها است. روش‌های مختلفی برای حذف یا جایگزینی چنین داده‌هایی وجود دارد لیکن هرکدام از آن‌ها دارای مشکلات و محدودیت‌های خاص خود است. یک روش مرسوم، جایگزین کردن داده‌های از دست‌رفته با میانگین سایر داده‌ها است. این روش با این که در بسیاری از مواقع، کارآمد است؛ اما موجب اختلال در توزیع و کم‌برآورد شدن واریانس می‌شود. یکی دیگر از راه‌های جایگزینی مقادیر از دست‌رفته، وقتی که چند دسته داده با ارتباط معنادار میان آن‌ها وجود داشته باشد، استفاده از برازش رگرسیون خطی است. اشکال این روش این است که همه داده‌ها را روی یک خط راست برازش می‌کند و لذا برای حالتی که داده‌ها

نوسان زیادی دارند، تخمین مناسبی به حساب نمی‌آید. روش SVM نیز یکی دیگر از راه‌های جایگزینی داده‌های از دست‌رفته است. یکی دیگر از روش‌های مرسوم، استفاده از هموارسازی پیاله‌ای است که در بسیاری مواقع نسبت به جایگزینی با میانگین، عملکرد بهتری دارد. اشکال این روش این است که علاوه بر

موجود در خوشه استفاده شود. استفاده از مدل رگرسیون در خوشه که سبب می‌شود در محاسبه داده‌های ازدست‌رفته، فیلدهای مربوط در صفات (ستون‌ها) دیگر نیز در نظر گرفته شود. همچنین، استفاده از جهش و ترکیب کروموزوم‌ها در الگوریتم ژنتیک که منجر به استفاده تلفیقی از میانگین، میانه، مد و مدل رگرسیون می‌شود، سبب دستیابی به نتایج قابل قبول تری خواهد شد.

استفاده از روش‌های هموارسازی پیمانه‌ای، میانگین داده‌ها، رگرسیون خطی و راهکار پیشنهادی نشان داد که روش پیشنهادی همواره دارای دقت بالاتری نسبت به سایر روش‌ها است. به‌عنوان پژوهش‌های آتی می‌توان به استفاده از تکنیک‌های داده‌کاوی شامل خوشه‌بندی و رگرسیون، و همچنین الگوریتم‌های هیوریستیک شامل الگوریتم ژنتیک اشاره کرد. بدین صورت که از خوشه‌بندی به‌منظور شناسایی رکوردهای مشابه، و محاسبه داده ازدست‌رفته بر اساس رکوردهای مشابه

References

- Aggarwal, C. C., & Yu, P. S. (2005). An effective and efficient algorithm for high-dimensional outlier detection. *The VLDB journal*, 14, 211-221.
- Arning, A., Agrawal, R., & Raghavan, P. (1996, August). A Linear Method for Deviation Detection in Large Databases. In *KDD* (Vol. 1141, No. 50, pp. 972-981).
- Biessmann, F., Rukat, T., Schmidt, P., Naidu, P., Schelter, S., Taptunov, A., ... & Salinas, D. (2019). DataWig: Missing Value Imputation for Tables. *J. Mach. Learn. Res.*, 20(175), 1-6.
- Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques*, 2nd. University of Illinois at Urbana Champaign: Morgan Kaufmann.
- Honghai, F., Guoshun, C., Cheng, Y., Bingru, Y., & Yumei, C. (2005, September). A SVM regression based approach to filling in missing values. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 581-587). Springer, Berlin, Heidelberg.
- Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.
- Kiani, R., & Montazeri, M. (2015). A review of outlier detection methods. *International Conference on Research in Science and Technology*, Kuala Lumpur, Malaysia. (Persian)
- 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.
- 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.
- Sadik, M., & Gruenwald, L. (2010, August). DBOD-DS: Distance based outlier detection for data streams. In *International Conference on Database and Expert Systems Applications* (pp. 122-136). Springer, Berlin, Heidelberg.
- Tada, M., Suzuki, N., & Okada, Y. (2022). Missing Value Imputation Method for Multiclass Matrix Data Based on Closed Itemset. *Entropy*, 24(2), 286.

- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520-525.
- 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.
- Zhou, X., Wang, X., & Dougherty, E. R. (2003). Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics*, 19(17), 2302-2307.