

# پرکردن داده‌های گمشده در داده‌های

## سری زمانی چندمتغیره

نگین دانشپور\* و سیده فاطمه میرابولقاسمی

دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجایی، تهران، ایران



### چکیده

داده‌های سری زمانی چندمتغیره در زمینه‌های مختلف مانند بیوانفورماتیک، زیست‌شناسی، ژنتیک، نجوم، علوم جغرافیایی و امور مالی یافت می‌شوند. بسیاری از این مجموعه‌داده‌ها دارای داده گمشده هستند. جایگذاری داده‌های گمشده سری زمانی چندمتغیره، یکی از مباحث چالش برانگیز است و قبل از فرایند یادگیری یا پیش‌بینی سری‌های زمانی باید با دقت مورد توجه و بررسی قرار گیرد. تحقیقات فراوانی در استفاده از روش‌های مختلف برای جایگذاری داده‌های گمشده سری زمانی انجام شده است که به‌طور معمول شامل روش‌های تجزیه و تحلیل و مدل‌سازی‌های ساده در کاربردهای خاص و یا سری‌های زمانی تک‌متغیره هستند. در این مقاله یک نسخه بهبودیافته از درون‌یابی معکوس فاصله وزن‌دار برای جایگذاری داده‌های گمشده پیشنهاد شده است. روش درون‌یابی معکوس فاصله وزن‌دار دو محدودیت اساسی دارد: (۱) یافتن بهترین نقاط نزدیک‌تر به داده‌های گمشده (۲) انتخاب توان تأثیر بهینه برای همسایگان داده گمشده. برای بهبود روش درون‌یابی، از خوشه‌بندی k-means استفاده شده است، تا همسایه‌های با بیشترین شباهت به الگوی داده‌ای انتخاب شوند. از آنجا که میزان تأثیر هر یک از همسایه‌ها بر روی داده گمشده متفاوت است، از الگوریتم جستجوی فاخته برای تعیین توان تأثیر همسایگی استفاده می‌شود. برای ارزیابی عملکرد روش پیشنهادی، از پنج معیار ارزیابی شناخته‌شده استفاده می‌شود. نتایج تجربی بر روی چهار مجموعه‌داده UCI با درصدهای مختلف گمشدگی مورد بررسی قرار گرفته و در مجموع الگوریتم پیشنهادی نسبت به سه روش مقایسه‌ای دیگر عملکرد بهتر و به‌طور میانگین حدود ۰/۰۵ خطای RMSE، ۰/۰۴ خطای MAE، ۰/۰۰۳ خطای MSE و ۵ درصد خطای MAPE داشته است. میزان همبستگی داده‌های واقعی و مقدار برآوردشده در روش پیشنهادی بسیار مطلوب و در حدود ۹۹ درصد است.

واژگان کلیدی: جایگذاری داده‌های گمشده، درون‌یابی IDW، الگوریتم جستجوی فاخته، خوشه‌بندی k-means، سری‌های زمانی چندمتغیره

## Missing Data Imputation in Multivariate Time Series Data

Negin Daneshpour\* & Seyede Fateme Mirabolghasemi

Faculty of Computer Engineering, Shahid Rajaei Teacher Training University, Tehran, Iran

### Abstract

Multivariate time series data are found in a variety of fields such as bioinformatics, biology, genetics, astronomy, geography and finance. Many time series datasets contain missing data. Multivariate time series missing data imputation is a challenging topic and needs to be carefully considered before learning or predicting time series. Frequent researches have been done on the use of different techniques for time series missing data imputation, which usually include simple analytic methods and modeling in specific applications or univariate time series.

\* Corresponding author

\* نویسنده عهده‌دار مکاتبات



In this paper, a hybrid approach to obtain missing data is proposed. An improved version of inverse distance weighting (IDW) interpolation is used to missing data imputation. The IDW interpolation method has two major limitations: 1) finding closest points to missing data 2) Choosing the optimal effect power for missing data neighbors. Clustering has been used to remove the first constraint and find closest points to the missing data. With the help of clustering, the search radius and the number of input points that are supposed to be used in interpolation calculations are limited and controlled, and it is possible to determine which points are used to determine the value of a missing data. Therefore, most similar data to the missing data are found. In this paper, the k-means clustering method is used to find similar data. This method has been more accurate than other clustering methods in multivariate time series.

Evolutionary algorithms are used to find the optimal effect power of each data point to remove the second constraint. Considering that each sample within each cluster has a different effect on the estimation of missing data, cuckoo search is used to find the effect on missing data. The cuckoo search algorithm is applied to the data of each cluster, and each data sample that has more similarity with the missing data has more influence, and each data sample that has less similarity has less influence and has less influence in determining the amount of missing data. Among evolutionary algorithms, evolutionary cuckoo search algorithm is used due to high convergence speed, much less probability of being trapped in local optimal points, and ability to quickly solve high dimensional optimization problems in multivariate time series problems.

To evaluate the performance of the proposed method, RMS, MAE,  $R^2$ , MSE and MAPE criteria are used. Experimental results are investigated on four UCI datasets with different percentages of missingness and in general, the proposed algorithm performs better than the other three comparative methods with an average RMSE error of 0.05, MAE error of 0.04, MSE error of 0.003, and MAPE error of 5. The correlation between the actual data and the estimated value in the proposed method is about 99%.

**Keywords:** Missing Data imputation, IDW Interpolation, Cuckoo Search Algorithm, k-means Clustering, Multivariate Time Series.

اغلب به دلایل مختلف مانند اتفاقات پزشکی، صرفه‌جویی در هزینه‌ها، ناهنجاری‌ها، عوامل محیطی، عدم پاسخ در آزمایش‌های علمی، خطاهای انسانی در اندازه‌گیری، مشکلات انتقال داده‌ها در سیستم‌های دیجیتال و کیفیت پایین حس‌گرها ممکن است از دست بروند [4-6]؛ از این رو وجود مقادیر گمشده در مجموعه داده‌هایی که جمع‌آوری شده‌اند تا حدودی اجتناب‌ناپذیر است. بسیاری از روش‌های تحلیل بر روی سری‌های زمانی نیاز به مجموعه اطلاعات کامل در هر گام زمانی دارند؛ به طوری که ممکن است داده‌های گمشده دارای مهم‌ترین ویژگی‌های سری زمانی باشند [7]. حضور داده‌های گمشده در مجموعه داده به‌طور چشم‌گیری نتایج تفسیر داده‌های حاصل از فرایندهای داده‌کاوی و یادگیری ماشین را تحت تأثیر قرار می‌دهد و حتی ممکن است، باعث ایجاد نتایج اشتباه و گمراهی پژوهش‌گران شود [8]. بنابراین رسیدگی به داده‌های گمشده یک مسأله مهم قبل از به کارگیری از روش‌های داده‌کاوی و یادگیری ماشین می‌باشد.

لیتل و روبین یک طبقه‌بندی سه‌گانه برای داده‌های گمشده پیشنهاد داده‌اند که عبارتند از: گمشده به‌صورت به‌طور کامل تصادفی ( $MCAR^1$ )، گمشده به‌صورت تصادفی ( $MAR^2$ ) و به‌طور تصادفی گم‌نشده<sup>3</sup>

<sup>1</sup> Missing Completely at Random

<sup>2</sup> Missing at Random

<sup>3</sup> Missing Not at Random

## ۱- مقدمه

داده‌های سری زمانی، مجموعه‌ای از مشاهدات تصادفی هستند که به تدریج در طول زمان جمع‌آوری شده‌اند. در این نوع داده‌ها نمونه‌های مجاور دارای همبستگی قوی و تغییرات اندک هستند. در سری‌های زمانی چندمتغیره از چندین ویژگی وابسته به زمان برای ایجاد مدل سری زمانی استفاده می‌شود. هر متغیر نه تنها به مقادیر گذشته خود بلکه به برخی از متغیرهای دیگر نیز وابستگی دارد و از این وابستگی برای پیش‌بینی مقادیر آینده استفاده می‌شود [1]. با افزایش قدرت ذخیره‌سازی داده‌ها و پردازنده‌ها، برنامه‌های دنیای واقعی فرصتی برای ذخیره و نگهداری داده‌ها برای مدت زمان طولانی پیدا کرده‌اند؛ از این رو داده‌ها در بسیاری از کاربردهای عملی به شکل داده‌های سری زمانی ذخیره می‌شوند. برای مثال داده‌های فروش، قیمت سهام، نرخ ارز در امور مالی، اطلاعات آب و هوا و اندازه‌گیری‌های پزشکی (مانند فشار خون و اندازه‌گیری الکتروکاردیوگرام) از این قبیل هستند. بر این اساس، داده‌های سری زمانی چندمتغیره در زمینه‌های مختلف مانند بیوانفورماتیک و زیست‌شناسی، ژنتیک، نجوم، علوم جغرافیایی و امور مالی یافت می‌شوند [2]، [3]. این مقدار داده‌های سری زمانی فرصت تجزیه و تحلیل سری زمانی برای بسیاری از پژوهش‌گران در جوامع داده‌کاوی را در دهه گذشته فراهم کرده است. این داده‌ها

پیش‌بینی جدا از هم هستند. با انجام این کار اغلب الگوهای گمشده‌گی در مدل‌های پیش‌بینی در نظر گرفته نمی‌شوند و منجر به نتایج و تحلیل ضعیف می‌شود [20]. بیشتر روش‌های جایگزینی داده‌های گمشده دارای الزامات و محدودیت‌هایی هستند که ممکن است در دنیای واقعی برقرار نباشد. برای مثال، بسیاری از این روش‌ها تنها روی داده‌ها با نرخ گمشده‌گی کوچک کار می‌کنند، یا نمی‌توانند گمشده‌گی در داده‌هایی را که به‌طور تصادفی یا به‌طور کامل تصادفی در سری‌های زمانی با طول‌های مختلف موجودند، رسیدگی کنند.

با بررسی انجام‌شده بر روی مقالات موجود، برخی از الگوریتم‌ها فقط متناسب با برخی از سری‌های زمانی خاص مانند داده‌های پزشکی [21]، ترافیک [22]، هواشناسی و یا داده‌های جغرافیایی [23] طراحی شده‌اند و قابلیت مقیاس‌پذیری بر روی سایر سری‌های زمانی را ندارند. بسیاری از الگوریتم‌ها تنها بر روی سری‌های زمانی تک‌متغیره قابل اجرا هستند [24]، [25]. برخی از الگوریتم‌ها بر روی سری‌های زمانی غیرمتناوب دقیق عمل نمی‌کنند [26]. کاستی‌های روش‌های موجود باعث شد تا درصد ارائه روشی بهتر برای جای‌گذاری داده‌های گمشده باشیم.

درون‌یابی، فرایند استفاده از نقاط با مقادیر معلوم یا نمونه‌برداری از نقاط معلوم برای برآورد مقادیر در سایر نقاط نامعلوم است. می‌توان از این روش برای پیش‌بینی مقادیر گمشده در هر گونه داده‌ای استفاده کرد. در این مقاله از روش درون‌یابی فضایی معکوس فاصله وزن‌دار ( $IDW^{15}$ )، برای تخمین داده‌های گمشده استفاده شده است [28]. درون‌یابی  $IDW$  به‌صراحت این فرضیه را پیاده‌سازی می‌کند که نقاطی که به یکدیگر نزدیکتر هستند، نسبت به مواردی که از هم دورتر هستند، شبیه‌تر هستند.  $IDW$  برای پیش‌بینی مقدار برای هر موقعیت غیرقابل اندازه‌گیری از مقادیر اندازه‌گیری‌شده در اطراف مکان پیش‌بینی استفاده می‌کند. زمانی که داده‌ها به‌اندازه کافی متراکم هستند، دقت اجرای  $IDW$  بالاتر است. روش‌های درون‌یابی به‌طور معمول دقیق، ساده و دارای محاسبات سریع هستند البته نیاز به مقداردهی اولیه پارامترها دارند. از جمله پارامترهای درون‌یابی  $IDW$ ، شعاع جستجو و تعیین توان تأثیر نقاط است. روش  $IDW$  با بررسی نقاط اطراف داده در شعاع جستجوی تعریف‌شده مقدار هر داده گمشده را محاسبه می‌کند. مقدار داده گمشده را می‌توان با میانگین‌گیری مجموع وزنی نقاط شعاع جستجو محاسبه کرد. نقاطی که از داده گمشده

(MNAR) [9]. وقتی مقدار گمشده از نوع MCAR است، به این معنی است که مقادیر گمشده مستقل از سایر متغیرها هستند. در حالت MAR، مقادیر گمشده را می‌توان با استفاده از مقادیر دیگر برآورد کرد. اگر مقادیر گمشده از نوع MNAR باشند، این مقادیر به سایر متغیرهای گمشده بستگی دارند و سازوکار لازم برای پرکردن داده‌های گمشده، نیاز به مدل‌سازی برای پرکردن داده‌ها دارد. از این رو مقادیر گمشده را نمی‌توان از متغیرهای موجود برآورد کرد. برای بیش‌تر رویکردها، مقادیر گمشده از نوع MAR محسوب می‌شوند [10].

در دهه‌های گذشته، بسیاری از روش‌ها برای محاسبه مقادیر گمشده توسعه یافته‌اند. روش‌های پرکردن داده‌های گمشده در داده‌های غیرزمانی نسبت به داده‌های سری زمانی به جهت وجود الگوهای زمانی مانند روند، فصل<sup>۱</sup>، تناوب<sup>۲</sup> و تغییرات نامنظم<sup>۳</sup> متفاوت هستند. پژوهش‌های متعددی برای نشان‌دادن اهمیت محاسبه مقادیر گمشده در سری زمانی انجام شده‌است [11]. یک راه‌حل ساده این است که داده‌های گمشده را حذف کرده و تحلیل، تنها بر روی داده‌های مشاهده شده انجام دهیم؛ اما هنگامی که تعداد داده‌ها کم باشد، این روش عملکرد خوبی ندارد. روش‌های هموارسازی<sup>۴</sup>، درون‌یابی<sup>۵</sup> [12] و اسپلاین<sup>۶</sup> [13] در جای‌گذاری داده‌های گمشده سری زمانی، ساده و کارآمد هستند؛ بنابراین به‌طور گسترده در عمل استفاده می‌شوند، اما در الگوهای پیچیده عملکرد قوی ندارند. برای تخمین بهتر داده‌های گمشده روش‌های پیچیده‌تری در جای‌گذاری داده‌های گمشده در سری زمانی نیز طراحی شده‌اند. این روش‌ها شامل تجزیه و تحلیل طیفی<sup>۸</sup> [14]، تابع‌های کرنل<sup>۹</sup> [15]، الگوریتم  $EM^{10}$  [16]، تکمیل ماتریس<sup>۱۱</sup> [17] و فاکتورسازی ماتریس<sup>۱۲</sup> [18] هستند. روش جای‌گذاری چندگانه<sup>۱۳</sup> [19] با اجرای روش‌های جای‌گذاری، عدم قطعیت<sup>۱۴</sup> را کاهش می‌دهد. ترکیبی از روش‌های جای‌گذاری به همراه مدل‌های پیش‌بینی در سری‌های زمانی اغلب منجر به یک فرآیند دومرحله‌ای می‌شود که در آن مدل‌های جای‌گذاری و

<sup>1</sup> Trend

<sup>2</sup> Seasonal

<sup>3</sup> Cyclic

<sup>4</sup> Irregular

<sup>5</sup> Smoothing

<sup>6</sup> Interpolation

<sup>7</sup> Spline

<sup>8</sup> Spectral Analysis

<sup>9</sup> Kernel Methods

<sup>10</sup> Expectationmaximization

<sup>11</sup> Matrix Completion

<sup>12</sup> Matrix Factorization

<sup>13</sup> Multiple Imputation

<sup>14</sup> Uncertainty

<sup>15</sup> Inverse Distance Weighting

دورتر هستند، نسبت به نقاط نزدیک‌تر تأثیر کمتری در مقدار گمشده دارند. یکی از معایب روش‌های درون‌یابی این است که همبستگی داده‌ها را در نظر نمی‌گیرند [29]، [30].

هدف از این مقاله، پرکردن داده‌های گمشده در سری‌های زمانی چندمتغیره است. این مقاله به دنبال ارائه روشی است که با وجود پیچیدگی‌های توزیع، غیرمتناوب بودن و یا غیرفصلی بودن سری‌های زمانی بتواند داده‌های مشابه و هم‌ساختار با داده گمشده را بیابد. همچنین دارای قدرت تخمین داده‌های بسیار متغیر نیز باشد و همبستگی متقابل دنباله داده‌ها را نیز در نظر بگیرد. با توجه به اینکه داده‌های سری زمانی داده‌های حجیم و پرتراکم هستند، یکی از روش‌های مناسب برای تخمین داده‌های گمشده در این داده‌ها درون‌یابی IDW است. درون‌یابی IDW قادر به پیش‌بینی در سری‌های زمانی چندمتغیره است. روش درون‌یابی IDW دو محدودیت اساسی دارد: (۱) یافتن بهترین نقاط نزدیک‌تر به داده‌های گمشده (۲) انتخاب بهترین توان تأثیر برای همسایگان داده گمشده. در این مقاله برای رفع محدودیت نخست و در واقع پیدا کردن نزدیک‌ترین نقاط به داده‌های گمشده، به الگوییابی داده‌ها پرداخته شده است تا داده‌هایی پیدا شوند که بیشترین شباهت و ساختار مشابه با داده گمشده را دارند. برای پیدا کردن داده‌های مشابه، در این مقاله از روش خوشه‌بندی k-means استفاده شده است. با پژوهش‌ها و بررسی‌های انجام‌شده، این روش دقت بالاتری نسبت به سایر روش‌های خوشه‌بندی در سری‌های زمانی چندمتغیره داشته است [31]، [32]. در روش خوشه‌بندی پیشنهادی از معیار اندازه‌گیری پیچش زمانی پویا<sup>۱</sup> (DTW) استفاده شده است. روش DTW روشی است که تطبیق بهینه بین دو دنباله زمانی با محدودیت‌های معین را پیدا می‌کند. با DTW می‌توان نقاط فضای n بعدی را به هم متصل و آن‌ها را با یکدیگر مقایسه کرد. این روش، زمانی مفید است که همه متغیرها استفاده شوند؛ بنابراین این روش برای سری‌های زمانی چندمتغیره مناسب است. با DTW می‌توان دنباله‌هایی از سری زمانی را پیدا کرد که دارای همبستگی متقابل هستند [33]. روش DTW حتی همبستگی اندک بین دنباله‌ها را نشان می‌دهد. برای بررسی میزان همبستگی دنباله‌ها می‌توان از یک روش تکاملی استفاده کرد. روش‌های تکاملی از نظر کاوش در فضای راه‌حل بسیار انعطاف‌پذیر و به‌نسبه سریع هستند؛ بنابراین برای رفع محدودیت دوم از الگوریتم‌های تکاملی برای جستجوی بهترین همسایگی‌ها و پیدا کردن توان

تأثیر بهینه هر نقطه داده استفاده شده است. بین الگوریتم‌های تکاملی، الگوریتم تکاملی جستجوی فاخته به جهت مواردی که در ادامه بیان می‌شوند، مناسب‌تر است: دقت و سرعت بالا، توانایی جستجوی محلی در کنار جستجوی کلی، احتمال بسیار کمتر گیرافتادن در بهینه‌های محلی، حرکت کلی جمعیت به سمت نقاط بهتر با نابود شدن جواب‌های نامناسب‌تر و توانایی حل سریع مسائل بهینه‌سازی در ابعاد بالا، هم‌گرایی سریع و داشتن پارامترهای کم جهت مقداردهی اولیه برای پیدا کردن نقاط بهینه همسایگی و تعیین توان تأثیر همسایگی [34]، [35]. روش پیشنهادی، مناسب برای پرکردن داده‌های گمشده در سری‌های زمانی چندمتغیره است.

در ادامه مروری بر ادبیات موضوع صورت می‌گیرد. در بخش سوم توصیف زمینه‌های نظری لازم برای درک روش پیشنهادی ارائه شده است، سپس الگوریتم ترکیبی پیشنهادی معرفی می‌شود. این روش‌ها روی داده‌های معیار اعمال می‌شوند و در بخش پنجم و ششم آزمایش‌ها بر روی مجموعه داده‌های معیار انجام شده و نتایج با استفاده از تجزیه و تحلیل آماری غیرپارامتری، ارزیابی می‌شود. در بخش آخر به نتیجه‌گیری می‌پردازیم.

## ۲- پیشینه پژوهش

روش‌های فراوانی در مقالات برای پرکردن داده‌های گمشده در سری‌های زمانی ارائه شده است. اگرچه الگوریتم‌های زیادی وجود دارند، اما آن‌ها اغلب در حوزه‌های خاص یا حتی برای مجموعه داده‌های خاص استفاده شده‌اند، برای مثال حمل و نقل [36]، هواشناسی [37] و دیگر کاربردها [38]. از این‌رو، از آن‌ها به‌طور عمومی استفاده نمی‌شود. در ادامه به بررسی مقالات عمومی تخمین داده‌های گمشده در سری‌های زمانی می‌پردازیم.

در مقاله [27] مدل یادگیری عمیق GRU-D بر روی سری‌های زمانی چندمتغیره براساس واحدهای بازگشتی شبکه عصبی بازگشتی عمل می‌کند. در این روش از یک معماری مدل عمیق استفاده شده است به طوری که علاوه بر اینکه وابستگی‌های زمانی طولانی مدت را در سری‌های زمانی حفظ می‌کند، از الگوهای گمشده نیز برای پیش‌بینی بهتر استفاده می‌کند. این مدل با رشد اندازه داده‌ها مقیاس‌پذیر است. فرایند یادگیری این الگوریتم بسیار زمان‌بر است و برای نخستین آموزش داده‌ها ۴۸ ساعت زمان لازم است. اگر داده گمشده حاوی اطلاعات مفید نباشد و یا ارتباط بین الگوهای داده‌های گمشده و پیش‌بینی‌ها مشخص نباشد این مدل ممکن است پیشرفت محدودی داشته باشد و یا حتی شکست

<sup>۱</sup> Dynamic Time Warping

مناسبی ارائه می‌دهد. این الگوریتم برای سری‌های زمانی تصادفی‌تر که دارای اختلال<sup>۳</sup> هستند و یا گام‌های تصادفی دارند، مناسب نیست و هم‌چنین در مجموعه‌داده‌هایی که دارای داده‌های گمشده به‌طورکامل تصادفی هستند و غیرمتناوب می‌باشند، کاربردی نیست. نویسنده پیشنهاد می‌کند در این شرایط از روش‌های درون‌یابی برای پیدا کردن داده‌های گمشده استفاده شود. تعداد محاسبات الگوریتم پیشنهادی در مجموعه‌داده‌های خیلی بزرگ زیاد می‌شود.

روش جایگذاری در [41] مبتنی بر شناسایی و مقایسه شباهت زیردنباله‌های موجود است. به‌منظور مقایسه زیردنباله‌ها یک معیار اندازه‌گیری شباهت جدید با استفاده از درون‌یابی قوانین فازی چندگانه ایجاد شده است. نکته مهم دیگر این رویکرد این است که هر سیگنال ناقص به‌عنوان دو سری زمانی جدا از هم که عبارتند از یک سری زمانی قبل از داده گمشده  $Q_a$  و یک سری زمانی بعد از این شکاف  $Q_b$  پردازش می‌شود. این دو سری به‌عنوان مرجع برای جستجو در نظر گرفته می‌شوند. مشابه‌ترین زیردنباله‌ها به زیردنباله‌های  $Q_a$  و  $Q_b$  جستجو و پیدا می‌شوند. برای اندازه‌گیری شباهت زیردنباله‌ها از امتیازات فازی براساس قوانین منطق فازی استفاده شده است. سرانجام داده گمشده با مقدار میانگین زیردنباله‌ها پر می‌شود. رویکرد پیشنهادی در پر کردن سری‌های زمانی با شکاف‌های بزرگ و با وجود عدم همبستگی بین متغیرها دقت مطلوب دارد. این روش در مجموعه‌داده‌های بزرگ پاسخ‌های بهتری را ارائه می‌کند. در این روش از مجموعه‌های فازی ساده استفاده شده است که نیاز است با مجموعه‌های فازی پیچیده هم امتحان شود.

یکی از روش‌های دیگر [42] یک رویکرد ترکیبی است که از روش فازی C-Means با الگوریتم ژنتیک برای تخمین داده‌های گمشده ترافیک استفاده کرده است. با استفاده از شباهت هفتگی بین داده‌ها، ابتدا ساختار داده‌های مبتنی بر بردار به الگوی داده‌ای مبتنی بر ماتریس تبدیل می‌شود؛ سپس از  $GA^4$  برای بهینه‌سازی کارکرد تابع عضویت و انتخاب مرکزها در مدل FCM استفاده شده است. در این روش یک ساختار جدید ماتریس با چند ویژگی ساخته شده است تا بتواند شباهت‌های هفتگی را روزه‌روز و بهتر تفسیر کند. هر داده گمشده به‌احتمال به بیش از یک مرکز خوشه متعلق است. این رویکرد، نتایج پایدارتری از تخمین داده‌های گمشده نسبت به C-means ساده ایجاد می‌کند.

بخورد. این مدل نمی‌تواند در روش‌های بدون نظارت و بدون برچسب مورد استفاده قرار گیرد.

در روشی دیگر [39]، از معیار پیچش زمانی پویا برای پر کردن سری‌های زمانی تک‌متغیره استفاده شده‌است. برای به‌دست‌آوردن مقدار گمشده، بزرگ‌ترین زیردنباله مشابه با زیردنباله قبلی از داده گمشده را می‌یابد؛ سپس داده گمشده را با مشابه‌ترین زیردنباله کامل می‌کند. در این مقاله از الگوریتم پیچش زمانی پویا برای مقایسه زیردنباله‌ها و از الگوریتم استخراج shape-feature برای کاهش محاسبات استفاده شده است. روش موجود، در سری‌های زمانی چندمتغیره قابل اجرا نیست روش پیشنهادی در سری‌های زمانی با همبستگی متقابل<sup>۱</sup> بالا، خودهمبستگی، توزیع‌های پیچیده و سری‌های فصلی قوی عمل می‌کند. از محدودیت‌های این الگوریتم فرض اولیه آن است که سری‌های زمانی دارای همبستگی متقابل بالا و دارای داده‌های تکراری باشند. این الگوریتم برای شکاف‌های بزرگ به زمان محاسبه زیاد نیاز دارد.

در [40] یک روش اصلاح الگوی پیش‌بینی مبتنی بر دنباله (PSF<sup>۲</sup>) ارائه شده است. در ابتدا الگوریتم PSF برچسب‌گذاری داده‌های سری زمانی را با استفاده از خوشه‌بندی انجام می‌دهد و سپس پیش‌بینی مبتنی بر دنباله انجام می‌شود. در الگوریتم PSF از خوشه‌بندی k-means برای تقسیم‌بندی سری زمانی استفاده شده است. دنباله‌ها با برچسب‌های تولیدشده حاصل از خوشه‌بندی به‌عنوان ورودی به الگوریتم PSF داده می‌شوند. الگوریتم PSF دارای سه مرحله است: انتخاب بهینه پنجره، تطابق الگوی دنباله‌ها و تخمین. الگوریتم imputPSF نیز براساس اصول الگوریتم PSF برای سری‌های زمانی طراحی شده است. این الگوریتم با تجزیه سری‌های زمانی ورودی با مقادیر گمشده در یک ماتریس به‌عنوان یک مرحله مقدماتی، برای تعیین مقدار ویژگی‌های گمشده آغاز می‌شود. یکی از دنباله‌ها به اندازه W برای جستجو انتخاب می‌شود. این دنباله در مجموعه ورودی جستجو می‌شود. اگر این الگوی پنجره در مجموعه‌داده‌های دارای برچسب تکرار شود، مقدار بعدی آن در یک بردار برای مقدار پیش‌بینی‌شده<sup>۲</sup> بعدی استفاده می‌شود. اگر هیچ تکراری برای دنباله پیدا نشود، اندازه پنجره یک واحد کاهش می‌یابد. این روند تا زمانی که الگوی موجود در یک پنجره انتخاب شده دست‌کم یک‌بار در مجموعه‌داده ورودی پیدا شود، تکرار می‌شود. روش بالا زمانی که داده‌های گمشده در پنجره‌های پشت سر هم رخ می‌دهند، پاسخ‌های

<sup>3</sup> Noise

<sup>4</sup> Genetic Algorithm

<sup>1</sup> Cross correlation

<sup>2</sup> Pattern Sequence Forecasting

(Table-1): Comparison of previous methods

شماره مرجع روش	نوآوری روش	سری زمانی	کار آیی	مقیاس پذیری	دقت
[27]	استفاده از مدل یادگیری عمیق GRU-D برای پرکردن داده‌های گمشده، حفظ وابستگی‌های زمانی طولانی مدت در ضمن پرکردن داده‌های گمشده، استفاده از الگوهای داده گمشده	چندمتغیره	متوسط	متوسط	خوب
[39]	استفاده از معیار پیش‌پیش زمانی پویا برای پیدا کردن بزرگترین زیر دنباله مشابه، استفاده از الگوریتم استخراج shape-feature برای کاهش محاسبات	تک‌متغیره	متوسط	خوب	ضعیف
[40]	اصلاح الگوریتم PSF برای پرکردن داده‌های گمشده، پیدا کردن زیر دنباله‌های مشابه و استفاده از خوشه‌بندی K-means برای برچسب‌گذاری دنباله‌ها	تک‌متغیره	ضعیف	متوسط	خوب
[41]	معرفی یک معیار شباهت جدید بر اساس قوانین فازی، پیدا کردن زیر دنباله‌های قبل و بعد از داده گمشده	چندمتغیره	متوسط	خوب	ضعیف
[42]	استفاده از ماتریس برای نمایش داده‌های گمشده، تطبیق الگوریتم فازی C-means برای پرکردن داده‌های گمشده، استفاده از الگوریتم ژنتیک برای بهینه‌سازی الگوریتم فازی C-means	تک‌متغیره	متوسط	متوسط	خوب
[43]	تطبیق الگوریتم اتورگرسیو برای تخمین داده‌های گمشده، پیدا کردن ضرایب اتو رگرسیو	چندمتغیره	متوسط	متوسط	خوب
[44]	ترکیب برنامه‌نویسی ژنتیک و درون‌یابی لاگرانژ برای پرکردن داده‌های گمشده، حفظ خصوصیات آماری در ضمن پرکردن داده‌های گمشده	چندمتغیره	متوسط	متوسط	ضعیف
[45]	اصلاح درون‌یابی IDW، استفاده از مجموعه راف برای پیدا کردن متغیرهای وابسته، استفاده از کلاس تولرانس برای پیدا کردن داده‌های مشابه	چندمتغیره	خوب	متوسط	متوسط

داده‌های گمشده است. این روش بر اساس دو مفهوم اساسی ساخته شده است: (۱) سری‌های زمانی فرایندهای تصادفی هستند که دارای ویژگی‌هایی از جمله میانگین، واریانس و خودهمبستگی هستند که باید در هنگام تخمین داده‌های گمشده، حفظ این خصوصیات اصلی مورد توجه قرار گیرد، (۲) یک الگوریتم برنامه‌نویسی ژنتیک قادر است یک مدل رگرسیون جامع ایجاد کند و امکان درک بهتر از الگوی داده‌های گمشده را فراهم می‌کند و برای استخراج دانش از آن استفاده می‌شود. الگوریتم پیشنهادی از لحاظ دقت عملکرد و حفظ خصوصیات اصلی توزیع داده‌ها عملکرد مطلوبی دارد؛ اما به دلیل ویژگی تصادفی، فرایند برنامه‌نویسی ژنتیک، الگوریتم باید بیش از یک‌بار اجرا شود تا نتایج خوبی را تضمین کند. بهتر است این الگوریتم بر روی زیر دنباله‌هایی از سری زمانی که دارای توزیع یکسان هستند و هر قطعه رفتار متفاوتی ندارند، اجرا شود. این تغییر ممکن است، عملکرد الگوریتم را در سری‌های زمانی غیرایستا افزایش دهد و تابع رگرسیون دقیق‌تری پیدا شود.

در مقاله [45] از روش اصلاح‌شده درون‌یابی IDW برای پرکردن داده‌های گمشده استفاده شده است. در این مقاله از روش مجموعه‌های راف برای پیدا کردن متغیرهای وابسته با یکدیگر استفاده شده است و از کلاس تولرانس برای پیدا کردن داده‌های مشابه با داده‌های گمشده استفاده

مقاله [43] برای تخمین داده‌های گمشده از یک مدل مبتنی بر اتورگرسیو<sup>۱</sup> استفاده می‌کند. این روش در شرایطی که یک نقطه خاص از زمان شامل داده گمشده باشد یا کل نقاط زمانی گمشده باشند، مؤثر است. خروجی پیش‌پردازش داده‌ها به پیش‌بینی‌کننده خطی و درجه دوم داده می‌شود. ضرایب اتورگرسیو بر اساس داده‌های مشابه داده گمشده تخمین زده و سپس مقدار داده گمشده تخمین زده می‌شود. در این الگوریتم، روش اتورگرسیو برای پیش‌بینی داده‌های گمشده توسعه داده شده است. این روش در شرایطی که یک ستون داده‌های گمشده زیادی داشته باشد، نتایج مطلوبی دارد؛ اما این الگوریتم برای شرایطی که بیش از یک ستون داده گمشده داشته باشد، طراحی نشده و عملکرد مطلوب ندارد.

در مقاله [44] از برنامه‌نویسی ژنتیک<sup>۲</sup> و درون‌یابی لاگرانژ<sup>۳</sup> برای پرکردن داده‌های گمشده استفاده شده است. روش ابتکاری پیشنهادی یک مدل رگرسیون قابل تفسیر است که ویژگی‌های آماری سری زمانی مانند میانگین، واریانس و خودهمبستگی<sup>۴</sup> را بررسی می‌کند. همچنین برای برآورد داده‌های گمشده از ایجاد ارتباط بین سری‌های چندمتغیره استفاده می‌کند. روش پیشنهادی بدون از بین بردن خصوصیات آماری قادر به تخمین

<sup>1</sup> Autoregression

<sup>2</sup> Genetic Programming

<sup>3</sup> Lagrange Interpolation

<sup>4</sup> Autocorrelation

به دست آوردن بینش نسبت به داده‌ها استفاده می‌شود [46]. همان‌طور که در پژوهش‌ها نشان داده شده است، روش‌های خوشه‌بندی عمومی، مانند k-means، برای داده‌های سری زمانی طراحی نشده و بنابراین ممکن است در سری‌های زمانی عملکرد خوبی نداشته باشند. این به‌طور عمده ناشی از این واقعیت است که بیش‌تر روش‌های خوشه‌بندی عمومی با معیار اندازه‌گیری فاصله اقلیدسی ساخته شده‌اند که به نظر می‌رسد اندازه‌گیری خوبی برای داده‌های سری زمانی نباشد [39]، [47]. معیارهای اندازه‌گیری مختلفی برای تشابه و عدم تشابه سری‌های

زمانی پیشنهاد شده است. معیار اندازه‌گیری پیچش زمانی پویا (DTW) مناسب‌ترین روش برای اندازه‌گیری فاصله در سری‌های زمانی است که استفاده از آن در خوشه‌بندی باعث بهبود عملکرد خوشه‌بندی در سری‌های زمانی می‌شود [47]. برای تعیین فاصله DTW بین دو سری زمانی به ترتیب با طول‌های  $n$  و  $m$ ، ابتدا یک ماتریس هزینه  $(n \times m)$  محاسبه می‌شود. عنصر  $(i, j)$  در این ماتریس فاصله بین دو نقطه  $x_i$  و  $y_j$  در دو سری زمانی  $i$  و  $j$  است. این فاصله به‌طور معمول به‌عنوان اختلاف درجه دوم تعریف می‌شود:

$$d(x_i, y_j) = (x_i - y_j)^2$$

سپس مسیر پیچش  $W = w_1, w_2, \dots, w_k$  به دست می‌آید که هر عنصر  $W$

نشان‌دهنده فاصله بین یک نقطه  $i$  در سری زمانی  $Q$  و نقطه  $j$  در سری زمانی  $C$  است.  $\max(m, n) \leq k \leq m + n - 1$  تحت سه شرط به دست می‌آید: شرط مرزی، شرط پیوستگی و شرط مونوتونیک.

شرط مرزی: نقطه شروع و پایان مسیر در گوشه پایین سمت چپ و بالا سمت راست ماتریس است:

$$w_k = (n, m) \text{ و } w_1 = (1, 1)$$

شرط پیوستگی: فقط عناصر مجاور همسایگی (از جمله مورب) در ماتریس در مسیر مجاز به انتخاب هستند که شامل عناصر مجاور قطری است. برای دو دنباله  $T_1$  و  $T_2$  فاصله پیچش زمانی پویا در ماتریس از طریق رابطه بازگشتی (۱) به دست می‌آید.

$$d(i, j) = D_{i,j} + \min \begin{cases} d(i, j-1) \\ d(i-1, j) \\ d(i-1, j-1) \end{cases} \quad (1)$$

$$i = 1, \dots, n \quad j = 1, \dots, m$$

که  $D_{i,j}$  فاصله دو نقطه  $i$  و  $j$  می‌باشد.

شده است، سپس از الگوریتم بهینه‌سازی ذرات برای پیدا کردن توان تأثیر هر یک از همسایگان استفاده شده است و در نهایت با پیدا کردن نزدیکترین همسایگان و توان تأثیر هر همسایه از درون‌یابی IDW برای پرکردن داده‌های گمشده استفاده شده است. این روش در راستای برطرف کردن محدودیت‌های درون‌یابی IDW بوده است و برای داده‌های پزشکی که سری‌های زمانی در فاصله‌های زمانی غیرمنظم اندازه‌گیری شده‌اند، مفید است. با افزایش میزان گمشدگی دقت عملکرد الگوریتم در این روش کاهش پیدا می‌کند.

در جدول (۱) مقایسه‌ای بر روی کارهای ارائه شده در این بخش صورت گرفته است. در این جدول نوآوری هر کار به صورت خلاصه آورده شده و دستاوردهای هر کدام مورد بررسی قرار گرفته است. ویژگی تک‌متغیره و یا چندمتغیره بودن سری‌های زمانی و سه پارامتر کارایی، مقیاس‌پذیری و دقت در این بررسی مورد توجه قرار گرفته است. کارایی، بیان‌گر نرخ تعداد مراحل لازم برای اجرای الگوریتم، برحسب طول داده ورودی است. مقیاس‌پذیری، مناسب بودن روش برای اجرا بر روی مجموعه داده‌های بزرگ را نشان می‌دهد. دقت نیز صحت تخمین داده‌های گمشده را برای هر روش مشخص می‌کند.

### ۳- مبانی نظری

هدف از این مقاله، پرکردن داده‌های گمشده در سری‌های زمانی چندمتغیره است. در روش پیشنهادی از درون‌یابی IDW برای جایگذاری داده‌های گمشده استفاده شده است. روش IDW محدودیت‌هایی دارد که هدف برطرف کردن این محدودیت‌ها در جهت ارتقای الگوریتم و افزایش دقت جایگذاری داده‌های گمشده است. در ادامه، پیش از پرداختن به جزئیات روش پیشنهادی، یک مرور کلی از برخی از مفاهیم پایه‌ای در روش پیشنهادی خواهیم داشت. در بخش ۳-۱ معیار فاصله DTW که در خوشه‌بندی‌های سری زمانی استفاده شده، مورد بررسی قرار می‌گیرد. در بخش ۳-۲ خوشه‌بندی k-means و در بخش ۳-۳ الگوریتم تکاملی جستجوی فاخته و ۳-۴ درون‌یابی معکوس فاصله وزن‌دار (IDW) شرح داده شده است.

#### ۳-۱- معیار فاصله DTW

خوشه‌بندی، عمل یافتن الگوهای پنهان یا گروه‌های مشابه در داده‌ها است. خوشه‌بندی در سری زمانی نیز برای

شرط مونوتونیک: مراحل بعدی در مسیر باید به صورت یک نواخت در زمان قرار بگیرند. به عنوان مثال در گام دوم باید مقادیر غیرکاهشی باشند. برای اجرای این الگوریتم از روش برنامه نویسی پویا برای پیدا کردن کمینه مسیر  $W$  با توجه به رابطه (۱) استفاده می شود؛ سپس فاصله DTW با مجذور حداقل فاصله تجمعی عناصر مسیر مطابق رابطه (۲) به دست می آید.

$$d_{DTW}(x, y) = \min \sqrt{\sum_{k=1}^K w_k} \quad (2)$$

که در این رابطه  $w_k$  فاصله مربوط به عنصر  $k$  ام در مسیر  $W$  است [47-51]. از این معیار فاصله DTW می توان در سایر خوشه بندی های سری های زمانی استفاده کرد.

### ۳-۲ - خوشه بندی k-means

روش های خوشه بندی، داده ها را به گروه هایی تقسیم می کنند که بیشترین شباهت را به همدیگر داخل گروه و بیشترین تفاوت را با داده های گروه های دیگر داشته باشند. خوشه بندی k-means که یکی از روش های خوشه بندی است، نمونه های  $n$  را به  $k$  خوشه ( $k$  پارامتر ورودی) گروه بندی می کند؛ به طوری که شباهت درون خوشه ای بالا، اما شباهت بین خوشه ای کم باشد. شباهت خوشه ها با توجه به میانگین نمونه های داخل خوشه سنجیده می شود. خوشه بندی k-means شامل مراحل زیر است [52]، [53]:

مجموعه  $X = \{x_1, x_2, x_3, \dots, x_n\}$ ، مجموعه  $n$  نمونه داده است.

(۱) به طور تصادفی  $k$  مرکز خوشه از داده ها انتخاب می شود مانند  $\{v_1, v_2, v_3, \dots, v_k\}$ .

(۲) فاصله هر یک از نمونه ها با مرکز خوشه ها محاسبه می شود.

(۳) هر یک از نمونه ها به یکی از مراکز خوشه ها نزدیک تر هستند و کمترین فاصله را دارند. نمونه داده در گروه آن خوشه قرار می گیرد.

(۴) دوباره مرکز خوشه های جدید  $v_i$ ،  $1 \leq i \leq k$  محاسبه می شود.

برای  $i$  امین خوشه داریم:

$$v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_j \quad (3)$$

که  $c_i$  نشان دهنده تعداد نقاط نمونه  $i$  امین خوشه و  $x_j$  نقطه داده ها در  $i$  امین خوشه  $1 \leq j \leq c_i$  است. (۵) فاصله بین هر نمونه و مراکز خوشه ای به دست آمده دوباره محاسبه می شود. (۶) در صورتی که هیچ نمونه ای تغییر خوشه نداشته باشد، الگوریتم متوقف، در غیر این صورت از مرحله سوم تکرار می شود.

### ۳-۳ - الگوریتم جستجوی فاخته (Cuckoo search)

الگوریتم بهینه سازی فاخته یکی از قوی ترین روش های بهینه سازی تکاملی است. در این مقاله از این الگوریتم برای پیدا کردن ضریب اهمیت هر یک از نقاط استفاده می شود. این الگوریتم دارای سرعت و دقت بالا است و هم گرایی سریعی دارد؛ هم چنین توانایی جستجوی محلی در کنار جستجوی کلی دارد. احتمال گیرافتادن در نقاط بهینه محلی کمتر است. حرکت کلی جمعیت به سمت نقاط بهتر با نابود شدن جواب های نامناسب اتفاق می افتد. الگوریتم فاخته با الهام از روش زندگی پرنده ای به نام فاخته است که در سال ۲۰۰۹ توسط شین او یانگ و دب ساوش توسعه یافته است [54]. الگوریتم فاخته بعدها در سال ۲۰۱۱ توسط رامین رجبیون با جزئیات بیشتر مورد بررسی قرار گرفت که از این نسخه در این مقاله استفاده شده است. مانند سایر الگوریتم های تکاملی، این الگوریتم نیز با جمعیتی از فاخته ها کار خود را شروع می کند. فاخته ها تعدادی تخم در لانه بعضی پرندگان میزبان می گذارند. بعضی از این تخم ها که شباهت بیشتری به تخم های پرنده میزبان دارند، برای رشد و پرند بالغ شدن، فرصت بیشتری دارند. سایر تخم ها توسط پرندگان میزبان شناسایی و از بین می روند. تخم های رشد یافته در یک منطقه، شایستگی لانه ها را در آنجا نشان می دهند. پس از چند تکرار تمام جمعیت فاخته ها به یک نقطه بهینه با بیشینه شباهت تخم ها به تخم های پرندگان میزبان و همچنین به محل بیشترین منابع غذایی می رسد. این محل بیشترین سود کلی را خواهد داشت و در آن کمترین تعداد تخم ها از بین خواهند رفت.

فاخته ها در جستجوی یافتن مناسب ترین منطقه برای تخم گذاری، به منظور افزایش نرخ بقای تخم هایشان هستند. بعد از اینکه تخم های باقی مانده رشد کردند و تبدیل به یک پرند بالغ شدند، اجتماعاتی را تشکیل می دهند. هر گروه محدوده سکونت مختص خودش را برای زندگی دارد و بهترین منطقه سکونت تمام گروه ها

مقصد بعدی فاخته‌ها در سایر گروه‌ها خواهد بود؛ بنابراین آن‌ها به سمت بهترین زیستگاه، مهاجرت خواهند کرد. آن‌ها در جایی نزدیک به بهترین زیستگاه، ساکن خواهند شد. با توجه به تعداد تخم‌هایی که هر فاخته دارد و نیز فاصله فاخته تا نقطه هدف (بهترین زیستگاه)، تعدادی شعاع تخم‌گذاری نسبت به آن مشخص می‌شود؛ سپس پرنده شروع به تخم‌گذاری تصادفی در لانه‌های درون شعاع تخم‌گذاری خود می‌کند. این فرآیند تا زمانی که بهترین موقعیت با ارزش سود بیشینه کسب شود ادامه می‌یابد و بیشتر جمعیت فاخته‌ها اطراف بهترین موقعیت جمع می‌شوند.

خوشه‌بندی در الگوریتم جستجوی فاخته کمک می‌کند تا محیط را به سرعت به چندین بخش تقسیم کرده و بهترین ناحیه را به صورت تخمینی مشخص کند. این ناحیه به احتمال زیاد شامل نقطه بهینه کلی است؛ سپس تمام فاخته‌ها به سمت این ناحیه مهاجرت و داخل آن ناحیه را به صورت بهتری جستجو می‌کنند. این امر موجب هم‌گرایی بسیار سریع‌تر الگوریتم فاخته می‌شود [54].

در این مقاله هر چقدر تعداد فاخته‌ها در یک لانه بیشتر باشد، یعنی شباهت بیشتری با تخم داخل لانه داشته است و در نتیجه این لانه توان اهمیت بیشتری دارد. این لانه‌ها همان نزدیکترین همسایگی به داده گمشده هستند. با استفاده از الگوریتم فاخته می‌توان توان بهینه اهمیت همسایگان داده گمشده را پیدا کرد.

همبستگی بین مشاهدات به عنوان تابعی از زمان است. خودهمبستگی فضایی برای به دست آوردن درجه شباهت یک شیء به اشیای نزدیک به خود است. یکی از دلایل اصلی این که چرا خودهمبستگی سری زمانی اهمیت دارد، این است که در آمار به مشاهدات مستقل از یکدیگر می‌پردازد. اگر خودهمبستگی در یک مجموعه داده وجود داشته باشد، مشاهدات مستقل از یکدیگر را نقض می‌کند. یکی دیگر از کاربردهای آن تجزیه و تحلیل خوشه‌ها و پراکندگی در داده‌های محیط زیست و بیماری است.

درون‌یابی فقط از تعداد نقاط شناخته شده در شعاع جستجوی خود استفاده می‌کند. به طور معمول از این نقاط یک میانگین وزن دار گرفته می‌شود و نتیجه برای نقطه مجهول ثبت می‌شود. در اینجا وزنی که هر نقطه معلوم با آن در میانگین مشارکت می‌کند، اهمیت پیدا می‌کند. در روش درون‌یابی IDW فرض بر این است که تأثیر هر پدیده متناسب با توانی از معکوس فاصله آن است؛ بنابراین تأثیر پدیده مورد نظر با افزایش فاصله، کاهش می‌یابد.

بر اساس این روش، ارتباط با نقاط گمشده، با افزایش فاصله کاهش می‌یابد. به عنوان مثال، ما سه نقطه داریم که مختصاتشان را می‌دانیم. با توجه به این روش درون‌یابی می‌توانیم به این نتیجه برسیم که نقطه مجهولی که مختصاتش را نمی‌دانیم، بیشترین شباهت را به نزدیکترین نقطه دارد؛ بنابراین برای تخمین نقاط مجهول، نمونه‌های اطراف باید مشارکت بیشتری نسبت به آن‌ها که دورترند، داشته باشند.

#### ۴-۳- درون‌یابی معکوس فاصله وزن دار (IDW)

روش درون‌یابی IDW یا همان معکوس فاصله وزن دار یکی از روش‌های معمول و پرکاربرد درون‌یابی است. هدف اصلی در درون‌یابی، مشخص کردن اندازه یک پارامتر گمشده در مناطقی است که در آن‌ها نمونه برداری انجام نشده و یا داده گمشده است. درون‌یابی فاصله معکوس وزن دار (IDW) یک تخمین ریاضیاتی (قطعی) است و فرض می‌کند که مقادیر نزدیک‌تر نسبت به مقادیر دورتر با عملکرد داده‌مان مرتبط‌تر هستند. در صورتی که اطلاعات متراکم و در فاصله‌های زمانی مساوی باشند، IDW بهتر عمل می‌کند [28].

یکی از پیش‌فرض‌های استفاده از IDW وجود خودهمبستگی فضایی بین داده‌ها است و دنبال پیدا کردن الگوهای تکراری در داده‌های سری زمانی است. از خودهمبستگی فضایی در آمار و تحلیل سری‌های زمانی استفاده می‌شود. خودهمبستگی فضایی به معنی

زمانی که نزدیک‌ترین نقاط مشخص شدند، حالا باید توان‌ها را در IDW یاد بگیریم. نقاط درون‌یابی شده بر اساس فاصله آن‌ها از مقدار نقاط شناخته شده برآورد می‌شوند. نقاطی که به مقادیر شناخته شده نزدیکتر هستند نسبت به نقاطی که دورتر هستند بیشتر باید تأثیر بگذارند. در درون‌یابی IDW با استفاده از رابطه (۴) می‌توان با استفاده از همسایگی‌ها و میزان اهمیتشان نقاط گمشده را تخمین زد [33].

$$z_p = \frac{\sum_{i=1}^n \left( \frac{z_i}{(d_i)^p} \right)}{\sum_{i=1}^n \left( \frac{1}{(d_i)^p} \right)} \quad (4)$$

در این رابطه  $z_p$  ویژگی داده گمشده،  $i = 1, \dots, N$  همسایگان معلوم داده گمشده،  $z_i$  ویژگی نظیر داده گمشده در همسایگان،  $d_i$  فاصله اقلیدسی همسایه  $i$ ام از داده گمشده و  $p_i$  توان تأثیر همسایه  $i$ ام داده گمشده است.

روش پیشنهادی چندین مرحله اصلی دارد. در این روش نسخه‌ای بهبودیافته از درون‌یابی IDW برای تخمین داده‌های گمشده پیشنهاد شده است. در روش درون‌یابی IDW فرض اصلی این است که میزان همبستگی و تشابه بین همسایه‌ها با فاصله بین آن‌ها متناسب است. IDW دو محدودیت اصلی دارد: (۱) در این روش درون‌یابی، باید از روشی مناسب برای انتخاب بهترین نقاط نزدیک‌تر به داده گمشده استفاده شود، (۲) انتخاب توان تأثیر برای همسایگان داده گمشده نیز باید بهینه باشد. مجموعه داده‌های ورودی، سری‌های زمانی چندمتغیره هستند. این روش درون‌یابی مناسب برای سری‌های زمانی ای هست که علاوه بر زمان به ویژگی‌های دیگر نیز وابسته و چندمتغیره هستند.

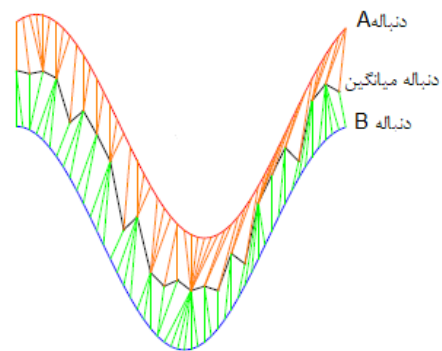
در روش‌های ابتدایی از یک شعاع جستجوی ثابت با فاصله مشخص و یا کمینه تعداد نقاط مورد نیاز برای ورودی درون‌یابی IDW استفاده شده است. اما در روش پیشنهادی، برای رفع محدودیت نخست، برای پیدا کردن نزدیکترین نقاط، از خوشه‌بندی داده‌های سری زمانی استفاده شده است. به کمک خوشه‌بندی، شعاع جستجو و تعداد نقاط ورودی که قرار است از آن‌ها در محاسبات مربوط به درون‌یابی استفاده شود، محدود و کنترل می‌شوند و می‌توان تعیین کرد که کدام نقاط برای تعیین ارزش یک داده گمشده به کار گرفته شود. خوشه‌بندی سری‌های زمانی یک مسأله چالش برانگیز است. در وهله نخست سری‌های زمانی اغلب بسیار بزرگ هستند. دومین چالش این است که سری‌های زمانی اغلب با ابعاد بالا هستند و در نهایت معیار شباهت که برای ایجاد خوشه‌ها استفاده می‌شود، مورد اهمیت است. از طرفی دیگر روش‌های عمومی خوشه‌بندی برای داده‌های سری‌های زمانی طراحی نشده‌اند و ممکن است عملکرد خوبی نداشته باشند. این امر هم اغلب از این واقعیت ناشی می‌شود که بیشتر روش‌های خوشه‌بندی عمومی از فاصله اقلیدسی استفاده می‌کنند که روش اندازه‌گیری خوبی برای داده‌های سری زمانی نیست. وجود این ویژگی‌ها نشان می‌دهد که از روش‌های موجود در خوشه‌بندی نمی‌توان به‌سادگی در دسته‌بندی سری‌های زمانی استفاده کرد. در روش پیشنهادی از روش خوشه‌بندی k-means برای پیدا کردن داده‌های مشابه استفاده شده است. با پژوهش‌ها و بررسی‌های انجام شده خوشه‌بندی k-means دقت بالاتری نسبت به سایر روش‌های خوشه‌بندی در سری‌های زمانی چندمتغیره داشته است [31]، [32].

معیارهای اندازه‌گیری شباهت Lock-step برای داده‌های سری زمانی مناسب نیستند. دلیل آن نیز محدودیت این معیارها در مدیریت اختلال، تأخیرهای زمانی و پرش‌های زمانی است. از طرفی سری‌های زمانی ممکن است در سرعت و زمان متفاوت باشند. با استفاده از تابع فاصله DTW می‌توان سری‌های زمانی را که ساختارهای مشابه دارند ولی در دوره‌های زمانی متفاوت هستند، مشابه قلمداد کرد. با استفاده از تابع DTW ماتریس فاصله بین مرکز خوشه‌ها و داده‌ها محاسبه می‌شود. مسأله بعدی در خوشه‌بندی سری‌های زمانی، خوشه‌بندی براساس نقطه داده‌ها، کل سری زمانی و یا زیر دنباله‌ای از داده‌ها است. سری‌های زمانی به‌طور طبیعی دارای اختلال و داده‌های پرت و تغییر مکان (شیفت) هستند و از طرفی طول سری‌های زمانی متفاوت است و کارکردن با سری‌های زمانی با طول زیاد فرایندی پیچیده است. برای اینکه بهبود در دقت عملکرد خوشه‌بندی و افزایش سرعت حاصل شود، به دنبال پیدا کردن زیردنباله‌های مشابه از طریق DTW هستیم. مجموعه‌ای از زیردنباله‌ها از سری زمانی با استفاده از sliding window استخراج می‌شوند. برای تعیین بهترین تعداد خوشه از تحلیل حساسیت بر روی خطای RMSE استفاده شده است. به‌عنوان مثال مطابق شکل (۲) با توجه به نتیجه تحلیل حساسیت در درصد گمشدگی ۳۰٪ بر روی مجموعه داده Energy، بهترین تعداد خوشه هشت است. برخلاف حالت معمول که چندین نقطه داده به‌عنوان مرکز خوشه انتخاب می‌شوند در اینجا چندین زیردنباله از سری‌های زمانی به‌عنوان مراکز خوشه انتخاب می‌شوند. بعد از انتخاب مراکز اولیه، نزدیک‌ترین زیردنباله‌ها به مراکز اولیه یافته می‌شوند. در مرحله بعدی برای پیدا کردن مراکز جدید خوشه‌ها نمی‌توان از میانگین‌گیری معمول استفاده کرد؛ زیرا به دنبال پیدا کردن دنباله‌هایی هستیم که ساختار مشابه دارند و ممکن است زیردنباله‌ها در سرعت و زمان متفاوت باشند. در هنگام استفاده از معیار DTW نقاط اتصالی بین دو دنباله در یک نمودار مشابه شکل (۱) ایجاد می‌شوند [55]. برای پیدا کردن دنباله مرکز خوشه جدید از میانگین مؤلفه‌های متصل به هم استفاده می‌شود. استفاده از این روش باعث می‌شود ساختار زیردنباله‌ها حفظ شود و دقت بالاتری در میانگین‌گیری زیردنباله‌ها داشته باشیم. این عملیات تکرار می‌شوند تا زمانی که هیچ دنباله‌ای از داده‌ها جابه‌جا نشوند. در نهایت بررسی می‌شود که دنباله داده گمشده به کدام خوشه متعلق می‌باشد. بنابراین نزدیکترین همسایگان نسبت به داده گمشده که بیشترین شباهت الگویی را دارند، پیدا می‌شوند.

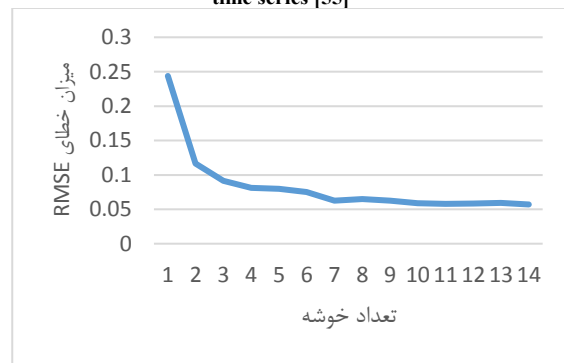
توان تأثیر هر یک از داده‌ها بر روی داده گمشده استفاده شده است. ابتدا متغیرهای مسأله به فرم یک آرایه شکل می‌گیرند و از یک آرایه  $habitat = [x_1, x_2, \dots, x_{N_{var}}]$  برای ذخیره موقعیت فعلی زندگی فاخته‌ها استفاده می‌شود. تعداد پارامترهایی که لازم است بهینه شوند نیز به صورت یک آرایه است که برابر تعداد داده‌های داخل خوشه است. این پارامترها همان توان تأثیر هر یک از داده‌ها هستند و مقداری بین صفر و یک دارند. میزان مناسب بودن یا مقدار سود با ارزیابی تابع سود  $maxprofit$  به دست می‌آید که میزان شباهت به داده گمشده را نشان می‌دهد. برخلاف ساختار عمومی جستجوی فاخته، این مقاله به دنبال بیشینه‌سازی سود است؛ بنابراین یک منفی در کنار  $maxprofit$  در نظر گرفته می‌شود. برای شروع الگوریتم بهینه‌سازی یک ماتریس  $habitat$  به ابعاد  $N_{population} \times N_{variable}$  تولید می‌شود و به هر کدام از این  $habitat$ ها تعدادی تصادفی تخم تخصیص داده می‌شود. جمعیت اولیه داده‌ها به صورت تصادفی بین صفر و یک تولید می‌شود. در هر دوره تخم‌گذاری تعداد تصادفی دو تا سه بردار ریاضی (فاخته) تولید می‌شود. حرکت فاخته‌ها به صورت دایره‌ای و در یک دامنه خاص قادر به تخم‌گذاری است. بعد از هر تخم‌گذاری ۱۰٪ از فاخته‌ها که مقدار تابع سود آن‌ها کمتر است، نابود می‌شوند. فاخته‌هایی که شباهت بیشتری به محل زندگی خود دارند به  $habitat$ های بهتر مطابق رابطه (۵) مهاجرت می‌کنند.

$$ELR = \alpha \times \frac{N}{T} \times (Var_{hi} - Var_{low}) \quad (5)$$

در رابطه (۵) آلفا متغیری است که بیشینه مقدار  $ELR$  با آن تنظیم می‌شود.  $N$  تعداد تخم‌های فعلی فاخته،  $T$  تعداد کل تخم‌ها و دو متغیر  $Var_{hi}$  و  $Var_{low}$  حد بالا و پایین متغیرهای مسأله هستند. هر فاخته فقط ۲٪ از کل مسیر را به سمت هدف ایده‌آل فعلی طی می‌کند و یک انحراف  $(-\frac{\pi}{6}, \frac{\pi}{6})$  نیز دارد. این دو پارامتر به فاخته‌ها کمک می‌کند تا محیط بیشتری را جستجو کنند. از خوشه‌بندی  $k$ -means در الگوریتم جستجوی فاخته استفاده شده است. این خوشه‌بندی کمک می‌کند تا محیط را به سرعت به چندین بخش تقسیم کرده و بهترین ناحیه را به صورت تخمینی مشخص کنند. این ناحیه به احتمال زیاد شامل نقطه بهینه کلی است. پس از چند تکرار تمام جمعیت فاخته‌ها به یک نقطه بهینه با بیشینه شباهت تخم‌ها به محل میزبان می‌رسند. در واقع محل بهینه جایی است که بیشتر فاخته‌ها در آن گرد می‌آیند و



(شکل-۱): میانگین‌گیری دو زیر دنباله سری زمانی [55]  
(Figure-1): The averaging two subsequence time series [55]



(شکل-۲): تحلیل حساسیت تعداد خوشه بر روی مجموعه داده Energy در درصد گمشده‌گی ۳۰٪

(Figure-2): Sensitivity analysis of the number of clusters on the Energy data set at a missing rate of 30%

برای رفع محدودیت دوم، از الگوریتم تکاملی جستجوی فاخته برای پیدا کردن توان تأثیر استفاده می‌شود. با توجه به اینکه هر کدام از نمونه‌های داخل هر خوشه تأثیر متفاوتی در برآورد داده گمشده دارند از جستجوی فاخته برای پیدا کردن توان تأثیر بر روی داده گمشده استفاده می‌شود. الگوریتم جستجوی فاخته بر روی داده‌های هر خوشه اعمال می‌شود و هر نمونه داده که شباهت بیشتر را با داده گمشده داشت توان تأثیر بیشتر و هر نمونه داده که شباهت کمتر داشت توان تأثیر کمتر می‌گیرد و تأثیر کمتر در تعیین مقدار داده گمشده می‌گذارد. در الگوریتم فاخته تخم‌هایی که شباهت بیشتری به تخم‌های پرنده میزبان دارند، شانس بیشتری برای رشد و تبدیل شدن به فاخته بالغ خواهند داشت. سایر تخم‌ها توسط پرنده میزبان شناسایی شده و از بین می‌روند؛ بنابراین موقعیتی که در آن بیشترین تعداد تخم‌ها نجات یابند پارامترهایی خواهد بود که الگوریتم جستجوی فاخته قصد بهینه‌سازی آن را دارد. تخم‌گذاری در منطقه با بیشترین سود ادامه می‌یابد. این رویه تا رسیدن به بهترین محل با بیشترین تعداد فاخته‌ها ادامه می‌یابد. از همین ویژگی الگوریتم فاخته برای پیدا کردن

<sup>1</sup> Egg Laying Area

```

for  $i = 1$  to  $m$  do loop
   $DTW[0.i] = \infty$ 
end
for  $i=1$  to  $n$  do loop
   $DTW[i.0] = \infty$ 
end
//Using pairwise method, incrementally fill in the
time similarity matrix with the difference of the two
series
for  $i = 1$  to  $n$  do loop
  for  $j = 1$  to  $m$  do loop
     $cost = d(v_1[i], v_2[j])$ 
     $DTW [i.j] = cost + MIN (DTW [i - 1.j],$ 
       $DTW [i.j - 1],$ 
       $DTW [i - 1.j - 1])$ 
  end
end
Return  $DTW [n.m]$ 

```

(الگوریتم-۳): شبه کد میانگین سری‌های زمانی با استفاده از

معیار اندازه‌گیری DTW

**Input:**  $a = a_1.a_2 \dots a_n$ , the first time series with length  $n$

$x_1 = x_{11}.x_{12} \dots x_{1m_1}$ , the first time series with length  $m_1$

$x_2 = x_{21}.x_{22} \dots x_{2m_2}$ , the first time series with length  $m_2$

.

.

$x_N = x_{N1}.x_{N1} \dots x_{Nm_N}$ , the first time series with length  $m_N$

**output:**  $y = y_1.y_2 \dots y_n$ , average of time series  $x_1.x_2 \dots x_N$ ;

$DTW_{avg}(a, x_1.x_2 \dots x_N)$ :

**Let**  $T = [\varphi.\varphi \dots \varphi]$ , an empty vector with  $n$  element;

**for**  $k = 1.2 \dots N$  **do**

$i = n$

$j = m_k$

**while**  $i \geq 1$  and  $j \geq 1$  **do**

$T_i = T_i \cup x_{kj}$

$(i.j) \leftarrow path_{i,j}$

**end**

**end**

**for**  $i = 1.2 \dots n$  **do**

$y_i =$  average of elements located in  $T_i$

**end**

(الگوریتم-۴): شبه کد الگوریتم تکاملی جستجوی فاخته

**Input:** Define fitness function  $F(x)$ ,  $x =$

$[x_1.x_2 \dots x_d]^T$

Generate initial habitats with some random points the profit function  $x_i$  ( $i = 1.2.3 \dots n$ )

Generate random keys for each  $x_i$

**While** ( $n <$  Maximum Generation) or (stop criterion)

Get a cuckoo (say,  $x_i$ ) randomly

Define ELR for each cuckoo

Let cuckoo to lay eggs inside their corresponding ELR

Generate a new solution  $x'_i$

جایی است که بیشترین شباهت را با داده گمشده دارد. زمانی که بردار ریاضی تمام فاخته‌ها که تولید می‌کنند مشابه هم شدند و یا اینکه انحراف معیار داده‌ها از  $0.000001$  کمتر شد، الگوریتم متوقف می‌شود. بدین ترتیب مقادیر توان تأثیر بهینه هر یک از داده‌ها نسبت به داده گمشده مشخص می‌شود. مقداردهی اولیه متغیرهای الگوریتم جستجوی فاخته در جدول (۲) نشان داده شده است.

(جدول-۲): مقداردهی اولیه متغیرهای الگوریتم

جستجوی فاخته

(Table-2): initialization of cuckoo search algorithm variables

مقدار متغیر	نام متغیر
۳۰	بیشینه تعداد فاخته‌های زنده
۱۰	تعداد اولیه فاخته‌ها
۳	بیشینه تخم‌گذاری هر فاخته
۲	کمینه تخم‌گذاری هر فاخته
۱	ضریب $\alpha$
۲	درصد حرکت در کل مسیر
$(-\frac{\pi}{6}, \frac{\pi}{6})$	زاویه انحراف
۱۰	درصد نابودی فاخته‌های نامناسب
۲	تعداد دسته‌های خوشه‌بندی k-means
۴	حداکثر تعداد تکرار

درنهایت در مرحله آخر از درون‌یابی IDW برای

تخمین داده‌های گمشده استفاده شده است. با استفاده از رابطه (۴) و نزدیکترین همسایگان داده گمشده و توان تأثیر هر یک از همسایگان، مقدار نقاط گمشده برآورد می‌شود.

روند اجرای الگوریتم روش پیشنهادی مطابق شبه

کد الگوریتم (۱ تا ۵) است.

(الگوریتم-۱): شبه کد خوشه‌بندی k-means در سری زمانی

**Input:**  $X$  (time series set),  $K$  (number of cluster)

**Output:** Clusters  $C = \{C_1.C_2 \dots C_k \dots C_K\}$

Generate the subsequence by sliding window

**While** termination is not satisfied **do**

**foreach** subsequence  $x_p$  **do**

Calculate the distance of  $x_p$  to each cluster by DTW distance

Assign  $x_p$  to the closest centroid

**end**

Update centroids by averaging

subsequence by DTW within each cluster

**end**

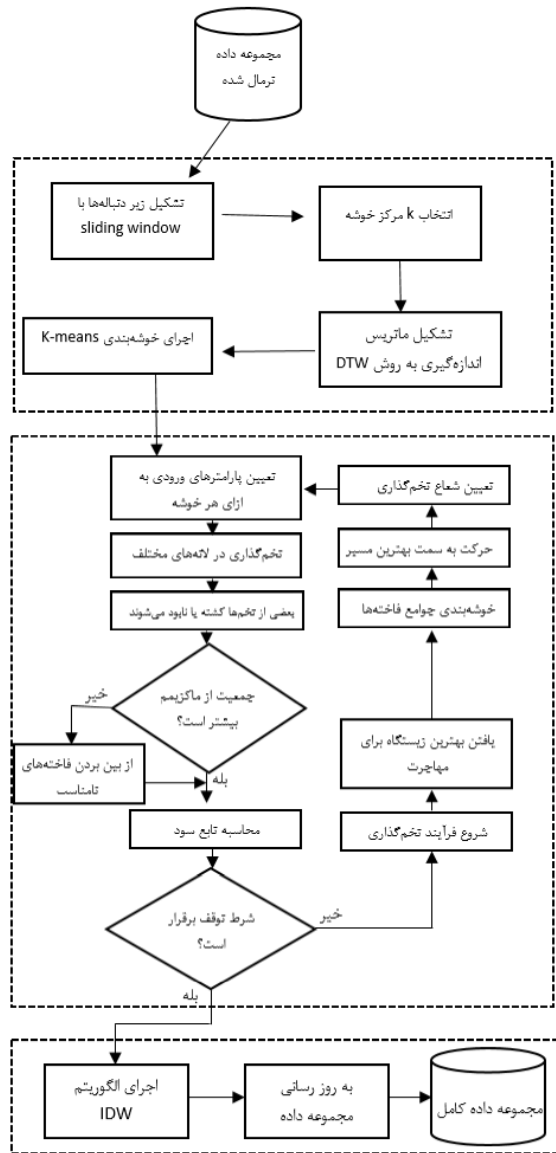
(الگوریتم-۲): شبه کد معیار اندازه‌گیری فاصله DTW

**Input:** vectors  $v_1 = (a_1 \dots a_n)$ ,  $v_2 = (b_1 \dots b_m)$  are the time series with  $n$  and  $m$  time points

**Output:**  $DTW [n.m]$

$DTW [0.0] = 0$  // Let a two dimensional data matrix DTW be the store of similarity measures such that  $DTW [0 \dots n.0 \dots m]$ , and  $i, j$ , are loop index, cost is an integer. // initialize the data matrix

الگوریتم DTW با پیچیدگی زمانی  $O(n^2)$  اجرا می‌شود. این پیچیدگی زمانی درجه دوم است. در DTW چون دو سری زمانی با یکدیگر مقایسه می‌شوند و نیاز به تشکیل ماتریس فاصله‌ها است یک فرایند زمان‌بر است. مرتبه زمانی الگوریتم k-means با توجه به معیار فاصله DTW مرتبه زمانی  $O(n^2ikm)$  است که در آن  $i$  تعداد تکرارها،  $k$  تعداد خوشه‌ها و  $m$  تعداد زیر دنباله‌ها است.



(شکل-۳): نمودار فعالیت روش پیشنهادی  
(Figure-3): Flowchart of the proposed method

در بخش بعدی روش پیشنهادی، از الگوریتم جستجوی فاخته استفاده شده است. مرحله نخست ایجاد تعداد اولیه لانه‌های میزبان است. پیچیدگی این بخش  $O(h \times z)$  است که  $h$  تعداد فاخته است که هر کدام راه‌حل به اندازه  $z$  را نگهداری می‌کنند. در مرحله تخم‌گذاری  $p$  تعداد از خانه‌ها که دارای سود بیشتر هستند انتخاب می‌شوند و به‌ازای هر یک،  $h$  تعداد فاخته با

Evaluate the habitat of each newly grown cuckoo

If  $F(x_i) > F(x_j)$

Replace  $x_j$  by the new solution  $x_i$

end

Cluster cuckoo and find best group and select goal habitat

From the available  $m$  host nest, poor quality nests ( $p_a$ ) are left

Keep the best solution (or nest)

Arrange solutions according to their

fitness value and select the solution with highest fitness value

end while

Post processing of the generated results

end

(الگوریتم-۵): شبه کد درون‌یابی IDW در سری‌های زمانی

**Input:** Clusters =  $\{C_1, C_2, \dots, C_k, \dots, C_K\}$ , keys for each  $x_i$ , The missing data is displayed with  $Z$ ,  $z_p$  is a missing attribute

**Output:** complete data set

**foreach** missing value  $M$  in dataset

Evaluate Eq 4

Replace Update the  $z_p$  value in  $Z$

end

end

در شکل (۳) نمودار فرایند روش پیشنهادی نشان داده شده است. همان‌طور که در روندنا نشان داده شده، روی مجموعه داده سری زمانی چندمتغیره، الگوریتم خوشه بندی k-means اجرا؛ سپس بررسی می‌شود که هر داده گمشده به کدام خوشه تعلق دارد. داده‌های آن خوشه نزدیکترین همسایگان داده گمشده هستند؛ سپس بر روی داده‌های آن خوشه، الگوریتم جستجوی فاخته را نسبت به داده گمشده اجرا کرده و توان تأثیر بهینه را که میزان اهمیت هر داده نسبت به داده گمشده را نشان می‌دهد به‌دست می‌آوریم؛ سپس با توجه به نزدیک‌ترین همسایگان به‌دست‌آمده برای داده گمشده و توان تأثیر، از درون‌یابی IDW برای تخمین داده گمشده استفاده می‌شود.

#### ۴-۱- پیچیدگی زمانی روش پیشنهادی

در روش پیشنهادی از الگوریتم k-means برای خوشه‌بندی سری‌های زمانی استفاده شده است. از معیار فاصله DTW نیز برای پیدا کردن فاصله بین زیردنباله‌ها در خوشه‌بندی k-means استفاده شده است. برای به‌دست‌آوردن مقدار DTW نیاز است ماتریس فاصله‌ها تشکیل شود تا مسیر کمینه فاصله بین دو زیردنباله به طول  $n$  و  $m$  به‌دست آورده شود. براساس برنامه‌نویسی پویا این الگوریتم با  $O(n \times m)$  به‌دست می‌آید. با توجه به اینکه زیردنباله‌ها دارای طول یکسان هستند، بنابراین

روی مجموعه داده‌های مختلف معنی‌دار باشد، مقادیر هر ویژگی از این مجموعه داده‌ها نرمال‌سازی شده‌اند [56]. مقیاس هر کدام از ویژگی‌ها در محدوده [0,1] قرار می‌گیرد. به این منظور، برای هر ورودی یک مجموعه داده، به‌عنوان مثال  $x_i^j$  برای تبدیل مقیاس از رابطه (۶) استفاده می‌شود.

$$(normalized) x_i^j = \frac{x_i^j - x^{j,min}}{x^{j,max} - x^{j,min}} \quad (6)$$

برای نرمال‌کردن یک عنصر آن عنصر را منهای مینیمم ( $x^{j,min}$ ) کرده و بر دامنه تغییرات ( $x^{j,max} - x^{j,min}$ ) تقسیم می‌نماییم. با این کار داده‌هایی که در محدوده‌های متفاوتی قرار دارند در یک دامنه مشابه قرار می‌گیرند تا اهمیت داده‌ها به واحد اندازه‌گیری‌شان بستگی نداشته باشد و در تحلیل نتایج اثر نامطلوبی ایجاد نکند.

(جدول-۳): مجموعه داده‌های استفاده شده در آزمایش‌ها

(Table-3): The dataset used in the experiments

مجموعه داده	تعداد نمونه‌ها	تعداد ویژگی
Electricity	۱۴۰۲۵۶	۳۷۰
Energy	۱۹۷۳۵	۲۹
Exchange	۵۳۶	۸
Traffic	۱۳۵	۱۸

## ۲-۵- روش‌های مقایسه

روش پیشنهادشده با سه روش پرکردن داده‌های گمشده به نام‌های روش بیشینه احتمال EM برای سری زمانی چندمتغیره [57]، روش FLK-NN [58] و روش TRiBS مقایسه شده است [45].

در روش بیشینه احتمال EM برای سری زمانی چندمتغیره فرض می‌کند داده‌ها توزیع گاوسی دارند. پارامترهای توزیع را برای داده‌های بدون داده گمشده به‌دست می‌آورد. در یک روند تکراری پارامترهای توزیع را اصلاح می‌کند. در این مقاله الگوهای مختلف داده‌های گمشده سری زمانی را نیز در نظر می‌گیرد [57]. در روش FLK-NN از ترکیب فوریه و نزدیک‌ترین همسایگی برای پرکردن داده‌های گمشده سری زمانی استفاده می‌کند [58]. روش TRiBS از ترکیب درون‌یابی IDW، مجموعه راف، کلاس تولرانس و الگوریتم تکاملی بهینه‌سازی ذرات برای پرکردن داده‌های گمشده سری زمانی استفاده می‌کند [45].

راه‌حلهایی به اندازه  $n$  ایجاد می‌شود که پیچیدگی این بخش  $O(h \times z \times p)$  است. راه‌حل‌ها رتبه‌بندی می‌شوند و بهترین راه‌حل فعلی پیدا می‌شود. این کار مستلزم مرتب‌سازی است. این مرتب‌سازی وابسته به تعداد راه‌حل‌ها است. اگر  $s$  راه حل داشته باشیم آن را می‌توان با پیچیدگی زمانی  $O(s \log s)$  انجام داد. در مرحله بعد خوشه‌بندی  $k$ -means بر روی فاخته‌ها انجام می‌شود تا محیط به چندین بخش تقسیم شود. مرتبه زمانی اجرای الگوریتم  $k$ -means از مرتبه  $O(i'k'm')$  است که در آن  $i'$  تعداد تکرارها،  $k'$  تعداد خوشه‌ها و  $m'$  تعداد نقاط است. الگوریتم فاخته  $g$  بار تکرار می‌شود؛ بنابراین پیچیدگی زمانی الگوریتم جستجوی فاخته از مرتبه  $O(g(phz + i'k'm'))$  است.

الگوریتم درون‌یابی IDW به‌ازای  $D$  داده گمشده اجرا می‌شود؛ بنابراین از مرتبه زمانی  $O(D)$  برخوردار است.

بدین ترتیب پیچیدگی زمانی الگوریتم پیشنهادی از مرتبه زمانی  $O(D + n^2ikm + g(phz + slogs + i'k'm'))$  می‌باشد.

## ۵- طراحی آزمایش‌ها

در این بخش جزئیات آزمایش‌های تجربی نشان داده شده است که شامل معرفی مجموعه داده‌های مورد استفاده، معیارهایی که در مقایسه‌ها استفاده شده‌اند و نتیجه آزمایش‌ها است. کدهای روش پیشنهادی در متلب اجرا شده است.

### ۱-۵- مجموعه داده‌ها و پیش‌پردازش داده‌ها

آزمایش‌ها بر روی چهار مجموعه داده واقعی از مخزن یادگیری ماشین UCI انجام شد [56]. این مجموعه داده‌ها شامل مجموعه داده‌های مصرف برق، انرژی خورشیدی، سهام و ترافیک هستند. این مجموعه داده‌ها سری‌های زمانی چندمتغیره هستند. این مجموعه داده‌ها در تعداد داده‌ها و ویژگی‌ها متفاوت هستند و سعی شده است مجموعه داده‌هایی با حجم، تعداد ویژگی و یا کاربرد متفاوت برای انجام آزمایش‌ها انتخاب شوند.

شرحی مختصر از هر مجموعه داده در جدول (۳) ارائه شده است. هیچ کدام از این مجموعه داده‌ها مقادیر گمشده ندارند. مجموعه داده‌های کاملی هستند که برای مقایسه نتایج مناسب هستند. برای اینکه مقایسه نتایج بر

### ۳-۵- معیارهای ارزیابی

از ابزارهای آماری برای یافتن دقت پیش‌بینی انجام‌شده در روش پیشنهادی استفاده می‌شود. میزان دقت و کارایی روش پیشنهادی با استفاده از پنج معیار ارزیابی شناخته شده بررسی می‌شود. معیارهای ارزیابی عبارتند از: جذر میانگین مربعات خطا (RMSE<sup>۱</sup>)، میانگین مطلق خطا (MAE<sup>۲</sup>)، ضریب تعیین (R<sup>۲</sup>)، میانگین مربعات خطا (MSE<sup>۳</sup>) و درصد میانگین مطلق خطا (MAPE<sup>۴</sup>).

جذر میانگین مربعات خطا (RMSE) به‌طورمعمول برای بررسی کارایی در مقادیر کمی استفاده می‌شود. این معیار اختلاف میانگین بین مقادیر واقعی و مقادیر برآورد شده در یک مدل را محاسبه می‌کند [58]. میانگین مطلق خطا (MAE) فاصله بین مقدار پیش‌بینی و واقعی را به‌عنوان معیار استفاده کرده ولی جهت این تفاضل را در نظر نمی‌گیرد؛ بنابراین در محاسبه خطا MAE فقط میزان فاصله و نه جهت فاصله به کار می‌رود [60]. هرچقدر مقدار RMSE و MAE کمتر باشد، میزان خطا کمتر بوده و مدل خوب پیش‌بینی انجام داده است. ضریب تعیین (R<sup>۲</sup>) میزان احتمال همبستگی میان دو دسته داده را مشخص می‌کند و به ما این امکان را می‌دهد که تعیین کنیم چقدر می‌توانیم به پیش‌بینی یک مدل مطمئن باشیم. این معیار تفاوت میانگین و واریانس مدل‌ها را بهتر نشان می‌دهد [61]. اگر این همبستگی زیاد باشد مدل داده‌ها را برازش کرده است درحالی‌که اگر همبستگی پایین (نزدیک به صفر) باشد، مدل برازش خوبی از داده‌ها ارائه نداده است. میانگین مربعات خطا (MSE) مقدار میانگین مربعات فاصله بین مقدار پیش‌بینی و واقعی را نشان می‌دهد. هرچقدر مقدار آن به صفر نزدیکتر باشد، نشان‌دهنده میزان کمتر خطاست. معیار درصد میانگین مطلق خطا (MAPE) میانگین فاصله مقدار پیش‌بینی شده و مقدار واقعی تقسیم بر مقدار واقعی را در نظر می‌گیرد. مزیت استفاده از این شاخص این است که می‌توان برای پیش‌بینی سری‌هایی که دارای مقیاس متفاوت هستند از آن استفاده کرد؛ زیرا این معیار وابسته به مقیاس نیست و بیشتر برای مقایسه چند سری زمانی مختلف کاربرد دارد [62]. در رابطه‌های (۷) تا (۱۱) نحوه محاسبه این معیارهای ارزیابی نشان داده شده است.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (7)$$

<sup>1</sup> Root Mean Square Error

<sup>2</sup> Mean Absolute Error

<sup>3</sup> Mean Squared Error

<sup>4</sup> Mean Absolute Percentage Error

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (8)$$

$$R^2 = \left( \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} \right)^2 \quad (9)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (10)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (11)$$

در این رابطه‌ها  $y_i$  مقدار واقعی و  $\hat{y}_i$  مقدار پیش‌بینی شده است.  $\bar{y}$  میانگین داده‌های مشاهده شده و  $n$  تعداد کل نمونه‌ها در مجموعه داده است.

برای ارزیابی دقت خوشه‌بندی انجام شده بر روی داده‌های سری زمانی نیاز به استفاده از یک معیار ارزیابی است. یکی از شاخص‌های ارزیابی درونی در خوشه‌بندی، شاخص دان<sup>۵</sup> است. شاخص دان با دو معیار فاصله و قطر، میزان فشردگی و تفکیک‌پذیری را محاسبه می‌کند. در رابطه (۱۲) نحوه محاسبه این معیار ارزیابی نشان داده شده است [63].

$$V_D = \left[ \frac{\min_{1 \leq i \leq k} D(C_i, C_j)}{\max_{1 \leq l \leq k} \text{diam}(C_l)} \right] \quad (12)$$

در صورت این کسر، فاصله بین دو خوشه به‌عنوان معیاری برای تفکیک‌پذیری و در مخرج نیز قطر هر خوشه دیده می‌شود. نسبت این دو، مقیاسی برای سنجش فاصله بین دو خوشه خواهد بود. هر چه مقدار این شاخص بزرگتر باشد، بیان‌گر تفکیک‌پذیری بهتر و در نتیجه خوشه‌بندی مؤثرتر است.

برای بررسی معنی‌دار بودن بهبود روش پیشنهادی نسبت به سایر روش‌ها از آزمون آماری t-test استفاده شده است. ستون‌های T در جدول‌های (۵ تا ۸) نشان‌دهنده آزمایش‌های معنی‌داری ستون‌های قبل نسبت به روش پیشنهادی است. مقدار احتمال در سطح اطمینان (p-value < 0.1)، ۹۰٪، (p-value < 0.05)، ۹۵٪ و (p-value < 0.01)، ۹۹٪ است که به ترتیب با "\*\*\*\*"، "\*\*\*"، "\*\*\*\*" و "\*\*\*\*" نشان داده شده است. اگر هیچ نمادی روی روشی قرار نگیرد بدین معنا هست که هیچ تفاوت معناداری با روش پیشنهادی ندارد.

<sup>5</sup> Dunn's Index

## ۶- تحلیل و بررسی نتایج آزمایش‌ها

این بخش نتایج آزمایش‌ها انجام شده برای ارزیابی عملکرد روش ترکیبی پیشنهادی ارائه می‌دهد. آزمایش‌ها شامل بررسی معیارهای ارزیابی بر روی پرکردن داده‌های گمشده با استفاده از روش درون‌یابی در ترکیب با روش خوشه‌بندی k-means و جستجوی فاخته است. در نهایت نتایج با دو روش دیگر پرکردن داده‌های گمشده مقایسه شده است. روش‌ها در درصد‌های مختلف گمشدگی ۳۰، ۴۰، ۵۰ و ۶۰ درصد بررسی می‌شوند تا حالت‌های مختلف مورد سنجش قرار گیرد و تأثیر آن مورد بررسی قرار گیرد.

در بخش نخست ابتدا خوشه‌بندی k-means بر روی مجموعه‌داده اعمال می‌شود. بعد از پیاده‌کردن خوشه‌ها بررسی می‌شود هر داده گمشده به کدام خوشه تعلق دارد. سپس الگوریتم جستجوی فاخته برای پیدا کردن توان بهینه تأثیر هر یک از همسایگان بر داده گمشده اجرا می‌شود و در نهایت درون‌یابی IDW انجام می‌شود. هر یک از این روش‌ها بر روی چهار مجموعه‌داده اعمال شده و معیارهای اندازه‌گیری خطا بر روی آن اعمال شده و نتایج در جدول‌های (۵ تا ۸) نمایش داده شده است.

برای ارزیابی خوشه‌بندی k-means بر روی چهار مجموعه‌داده از شاخص دان استفاده شده است. نتایج حاصل از این ارزیابی در جدول (۴) نشان داده شده است. به‌طور متوسط میزان این شاخص در هر چهار مجموعه حدود ۱/۴ بوده است که نشان از دقت مطلوب خوشه‌بندی k-means پیشنهادی بر روی داده‌های سری زمانی دارد.

همان‌طور که پیش‌بینی می‌شد با افزایش میزان گمشدگی از سی تا شصت درصد میزان خطا افزایش پیدا کرده است. میزان خطای روش پیشنهادی در دو مجموعه‌داده Electricity و Energy نسبت به دو مجموعه‌داده Exchange و Traffic کمتر بوده و میزان خطای RMSE آن به‌طور میانگین در حدود ۰/۰۵ در دو مجموعه نخست و حدود ۰/۰۶ در دو مجموعه دوم بوده است. با توجه به اینکه دو مجموعه‌داده Electricity و Energy دارای تعداد نمونه و ویژگی بیشتری هستند، پیش‌بینی می‌شد که دقت عملکرد بالاتری داشته باشند. با توجه به معیارهای MAE و MSE و MAPE بهترین پاسخ مربوط به دو مجموعه‌داده Electricity و Energy با تعداد نمونه بیشتر نسبت به سایر مجموعه‌داده‌ها است و روش خوشه‌بندی k-means در ترکیب با روش درون‌یابی پاسخ‌های بهتری برای پرکردن داده‌های گمشده نسبت به سایر روش‌ها داشته است. میزان خطا MAE در این روش

در حدود ۰/۰۴ بوده است. مجموعه‌داده Exchange و Traffic که تعداد نمونه‌های کمتری دارند، میزان خطای بیشتری نسبت به مجموعه‌داده‌های دیگر در هر سه روش دارند.

مقدار  $R^2$  که میزان همبستگی بین مجموعه‌داده‌های واقعی و مقدار تخمین زده‌شده را در روش پیشنهادی را نشان می‌دهد در حدود ۹۹ درصد است که نشان از دقت بالای تخمین مقادیر گمشده است. روش پیشنهادی در دو مجموعه‌داده Electricity و Energy که تعداد نمونه بیشتری دارند، میزان دقت بالاتری دارد که این یک حسن است. با توجه به اینکه سری‌های زمانی به‌طور معمول داده‌های حجیم هستند، این روش برای پرکردن داده‌های گمشده در این مجموعه‌داده‌ها مناسب‌تر است. روش EM دقت عملکرد بهتری نسبت به روش FLK-NN دارد. میزان خطای RMSE در روش EM در دو مجموعه‌داده Electricity و Energy در حدود ۰/۰۶ و در دو مجموعه‌داده Exchange و Traffic در حدود ۰/۰۸ می‌باشد که در روش پیشنهادی این میزان خطا به ۱/۲ تا ۱/۳ کاهش یافته است. میزان خطای RMSE در FLK-NN در دو مجموعه‌داده نخست در حدود ۰/۰۸ است و در مجموعه‌داده سوم و چهارم در حدود ۰/۰۹ است. با توجه به معیارهای MAE و MSE و MAPE میزان خطا در روش پیشنهادی نسبت به سایر روش‌ها در هر چهار مجموعه‌داده کمتر بوده است. میزان همبستگی داده‌های واقعی و تخمین زده‌شده در روش‌های مقایسه‌ای نسبت به روش پیشنهادی کمتر و در حدود ۸۵ درصد بوده است. همان‌طور که پیش‌بینی می‌شد میزان خطا در تمام روش‌ها با افزایش درصد گمشدگی داده‌ها افزایش می‌یابد. میزان خطای روش TRiBS در مجموعه‌داده Energy مشابه با میزان خطای روش پیشنهادی است و در درصد‌های گمشدگی ۳۰٪ و ۴۰٪ نیز از روش پیشنهادی بهتر عمل کرده است.

با توجه به نتایج حاصل از آزمایش‌ها، روش پیشنهادی خطای کمتری نسبت به سایر روش‌ها دارد. هر چقدر تعداد نمونه‌ها و ویژگی‌ها در مجموعه‌داده‌ها بیشتر باشد میزان خطا کمتر می‌شود. میزان همبستگی داده‌های واقعی و مقدار برآوردشده در روش پیشنهادی بسیار مطلوب و در حدود ۹۹ درصد است.

تحلیل آماری از نتایج t-test نشان می‌دهد که تفاوت معناداری در میانگین مقدار RMSE،  $R^2$  و MSE در روش پیشنهادی نسبت به سایر روش‌ها است که نشان‌دهنده عملکرد قابل اطمینان‌تر روش پیشنهادی

(جدول-۴): اجرای شاخص ارزیابی دان برای بررسی کیفیت خوشه‌بندی K-means بر روی چهار مجموعه داده

(Table-4): Implementation of Dunn Evaluation Index to evaluate the quality of K-means clustering on four data sets

معیار اندازه‌گیری دان	Electricity	Energy	Exchange	Traffic
روش پیشنهادی	۱/۱۱۳	۱/۲۳۱	۱/۶۸۵۸	۱/۵۸۰۲

(جدول-۵): مقایسه روش پیشنهادی با درصدهای گمشدگی متفاوت در مجموعه داده Electricity در مقایسه با سایر روش‌ها

(Table-5): Comparison of Proposed Method with Different Missing Percentages in Electricity Data Set Compared to Other Methods

روش	درصد گمشده‌گی	RMSE	T	MAE	T	$R^2$	T	MSE	T	MAPE	T
روش پیشنهادی	۳۰٪	۰.۰۴۷۳۲۴		۰.۰۳۳۴۰۸		۰.۹۹۴۶۹		۰.۰۲۲۳۹۵		۵.۶۴۳۸	
	۴۰٪	۰.۰۴۷۳۷۲		۰.۰۳۳۴۹۵		۰.۹۹۴۶۸		۰.۰۲۲۴۴۱		۵.۶۵۹۳	
	۵۰٪	۰.۰۴۷۳۵۲		۰.۰۳۳۴۹۶		۰.۹۹۴۶۶		۰.۰۲۲۴۵۱		۵.۶۵۸۹	
	۶۰٪	۰.۰۴۷۳۶۱		۰.۰۳۳۴۹۸		۰.۹۹۴۶۸		۰.۰۲۲۴۳۱		۵.۶۶۱۵	
EM	۳۰٪	۰.۰۶۳۶۵۱	***	۰.۰۵۵۰۵۲	**	۰.۸۸۵۷۲	***	۰.۰۴۰۵۱۴	***	۱۶.۵۲۳	
	۴۰٪	۰.۰۶۳۲۹۸	***	۰.۰۵۷۱۸۵	**	۰.۸۸۶۹۳	***	۰.۰۴۰۶۶	***	۱۶.۰۲۳	**
	۵۰٪	۰.۰۶۴۷۸	***	۰.۰۵۸۹۷۴	***	۰.۸۸۹۷۶	***	۰.۰۴۱۹۶۴	***	۱۶.۵۲۱	***
	۶۰٪	۰.۰۶۴۳۳۱	***	۰.۰۵۸۱۲۴	***	۰.۸۹۶۳	***	۰.۰۴۱۲۸۴	***	۱۶.۸۵۶	***
FLK-NN	۳۰٪	۰.۰۸۹۷۶۵	**	۰.۰۶۴۹۰۷	***	۰.۸۷۱۱۱	***	۰.۰۸۰۵۷۷	***	۱۶.۹۶۵	**
	۴۰٪	۰.۰۸۹۷۸۵	***	۰.۰۶۴۴۶	**	۰.۸۶۲۹۸	***	۰.۰۸۰۶۱۳	**	۱۶.۹۸۹	*
	۵۰٪	۰.۰۸۴۱۱۹	***	۰.۰۶۳۹۴۴	*	۰.۸۶۰۹۷	***	۰.۰۷۰۷۶۰	***	۱۶.۵۸۹	
	۶۰٪	۰.۰۸۴۷۰۲	***	۰.۰۶۳۸۷۷	***	۰.۸۶۱۷۹	***	۰.۰۷۱۷۴۴	***	۱۶.۴۵۶	**
TRiBS	۳۰٪	۰.۰۸۲۲۲۰	***	۰.۰۶۷۸۵۴	**	۰.۸۶۵۱۱	***	۰.۰۶۷۶۰۱	***	۱۵.۱۲۵	***
	۴۰٪	۰.۰۸۹۳۳۹	***	۰.۰۶۸۹۵۷	***	۰.۸۸۹۶۸	***	۰.۰۷۹۸۴۵	***	۱۵.۴۸۳	**
	۵۰٪	۰.۰۹۵۸۵۳	***	۰.۰۸۵۳۶۶	**	۰.۸۶۲۵۷	**	۰.۰۹۱۸۷۷	***	۱۶.۸۶۷	**
	۶۰٪	۰.۰۹۲۹۲۶	***	۰.۰۸۲۵۶۹	**	۰.۸۶۳۳۲	**	۰.۰۸۶۳۵۲	***	۱۶.۴۳۶	**

(جدول-۶): مقایسه روش پیشنهادی با درصدهای گمشدگی متفاوت در مجموعه داده Energy در مقایسه با سایر روش‌ها

(Table-6): Comparison of the proposed method with the percentage of different missing data in the Energy dataset compared to other methods

روش	درصد گمشده‌گی	RMSE	T	MAE	T	$R^2$	T	MSE	T	MAPE	T
روش پیشنهادی	۳۰٪	۰.۰۵۷۳۷۹		۰.۰۴۳۴۸۷		۰.۹۹۴۶۸		۰.۰۳۲۹۲۳۴		۵.۶۶۲۴	
	۴۰٪	۰.۰۵۷۳۶۱		۰.۰۴۳۴۶۷		۰.۹۹۴۶۸		۰.۰۳۲۹۰۲۸		۵.۶۵۸	
	۵۰٪	۰.۰۶۷۴۰۹		۰.۰۴۳۵۰۷		۰.۹۹۴۶۷		۰.۰۴۵۴۳۹۷		۵.۶۶۵۳	
	۶۰٪	۰.۰۵۷۳۵۹		۰.۰۴۳۴۶۳		۰.۹۹۴۶۸		۰.۰۳۲۹۰۰		۵.۶۵۸۲	
EM	۳۰٪	۰.۰۶۶۳۷۸	***	۰.۰۵۸۸۵۲	**	۰.۹۶۹۴۳	***	۰.۰۴۴۰۶۰	***	۱۳.۸۵۹	
	۴۰٪	۰.۰۶۸۴۵۲	***	۰.۰۵۹۳۶۷	**	۰.۹۶۵۷۲	***	۰.۰۴۶۸۵۶	***	۱۳.۴۵۲	**
	۵۰٪	۰.۰۶۵۵۲۶	***	۰.۰۵۹۸۸۳	***	۰.۹۵۰۲۷	***	۰.۰۴۲۹۳۶	***	۱۳.۴۵۶	**
	۶۰٪	۰.۰۶۷۷۸۴	**	۰.۰۵۹۵۸۹	***	۰.۹۵۶۰۸	***	۰.۰۴۵۹۴۶	*	۱۳.۴۸۹	***
FLK-NN	۳۰٪	۰.۰۷۵۳۲۱	***	۰.۰۳۸۸۰۹	***	۰.۸۵۰۸۵	***	۰.۰۵۶۷۳۲	***	۱۳.۸۵۶	
	۴۰٪	۰.۰۷۷۶۷۳	***	۰.۰۴۰۴۴۷	**	۰.۸۴۱۸۹	**	۰.۰۶۰۳۳۰	***	۱۳.۵۹۸	**
	۵۰٪	۰.۰۷۴۸۹۴	***	۰.۰۳۸۸۱۱	**	۰.۸۵۱۵۴	***	۰.۰۵۶۰۹۱	***	۱۲.۹۴۸	**
	۶۰٪	۰.۰۷۶۸۵۶	***	۰.۰۳۹۸۵۳	***	۰.۸۴۳۵۵	***	۰.۰۵۹۰۶۸	***	۱۳.۹۳۶	
TRiBS	۳۰٪	۰.۰۴۶۳۶۰		۰.۰۳۹۲۰۱	***	۰.۹۹۲۳۱	**	۰.۰۲۱۴۹۲		۵.۸۹۲۳۸	**
	۴۰٪	۰.۰۴۱۶۶۱	**	۰.۰۳۸۵۶۲	***	۰.۹۹۵۲۰	**	۰.۰۱۷۳۵۶	*	۵.۲۳۹۳	**
	۵۰٪	۰.۰۵۳۸۷۲	*	۰.۰۴۳۸۵۲	**	۰.۹۹۴۱۶	**	۰.۰۲۹۰۲۱	**	۵.۴۹۱۳	*
	۶۰٪	۰.۰۵۹۸۹۵	**	۰.۰۴۵۴۲۳	**	۰.۹۹۸۵۲	*	۰.۰۳۵۸۷۴	**	۵.۸۹۸۶	*

(جدول-۷): مقایسه روش پیشنهادی با درصدهای گمشدگی متفاوت در مجموعه داده Exchang در مقایسه با سایر روش‌ها

(Table-7): Comparison of the proposed method with the percentage of different missing data in the Exchange dataset compared to other methods

روش	درصد گمشده‌گی	RMSE	T	MAE	T	$R^2$	T	MSE	T	MAPE	T
روش پیشنهادی	۳۰٪	۰.۰۶۷۳۵۸		۰.۰۵۳۴۷		۰.۹۹۴۶۸		۰.۰۴۵۳۷۱		۷.۶۵۵۴	
	۴۰٪	۰.۰۶۷۳۹۵		۰.۰۵۳۵۰۱		۰.۹۹۴۶۷		۰.۰۴۵۴۲۰		۷.۶۵۹۹	
	۵۰٪	۰.۰۶۷۴۴		۰.۰۵۳۵۱۷		۰.۹۹۴۶۶		۰.۰۴۵۴۸۱		۷.۶۶۱۸	
	۶۰٪	۰.۰۷۴۱۸		۰.۰۵۳۵۲		۰.۹۹۴۶۷		۰.۰۵۵۰۲۶		۷.۶۶۶۱	
EM	۳۰٪	۰.۰۸۸۴۵	***	۰.۰۶۴۵۲۶	***	۰.۸۸۹۰۸	***	۰.۰۷۸۲۳۴	***	۱۴.۱۲۸	***

	۴۰٪	۰.۰۸۸۹۶۵	***	۰.۰۶۸۴۵۶	***	۰.۸۸۷۱۹	***	۰.۰۷۹۱۴۷	***	۱۴.۴۴۳	***
	۵۰٪	۰.۰۸۹۲۳۵	***	۰.۰۶۸۵۹۶	**	۰.۸۸۸۰۹	***	۰.۰۷۹۶۲۸	***	۱۶.۴۲۷	**
	۶۰٪	۰.۰۸۸۲۳۸	***	۰.۰۶۶۵۳	**	۰.۸۸۵۶۲	***	۰.۰۷۷۸۵۹	***	۱۸.۵۹۶	
<b>FLK-NN</b>	۳۰٪	۰.۰۹۶۷۸	***	۰.۰۸۳۵۷۶	**	۰.۸۶۷۱۷	***	۰.۰۹۳۶۶۳	***	۱۳.۵۱	**
	۴۰٪	۰.۰۹۶۷۶	***	۰.۰۸۳۵۸۹	**	۰.۸۶۷۱۶	***	۰.۰۹۳۶۲۴	*	۱۳.۳۱۲	
	۵۰٪	۰.۰۹۷۵۲	***	۰.۰۸۳۶۵۹	**	۰.۸۶۷۰۹	***	۰.۰۹۵۱۰	**	۱۳.۴۵۶	*
	۶۰٪	۰.۰۹۷۸۹	***	۰.۰۸۳۷۱۴	**	۰.۸۶۷۱۴	***	۰.۰۹۵۸۲۴	***	۱۳.۵۶۹	**
<b>TRiBS</b>	۳۰٪	۰.۰۹۵۳۱۴	***	۰.۰۷۸۹۵۶	***	۰.۸۵۳۳۶	**	۰.۰۹۰۸۴۷	***	۱۳.۲۸۹۴	**
	۴۰٪	۰.۰۹۲۸۵۵	***	۰.۰۷۳۶۳۸	***	۰.۸۵۵۶۲	***	۰.۰۸۶۲۰	***	۱۳.۲۶۹۸	**
	۵۰٪	۰.۰۹۴۱۲۷	***	۰.۰۸۲۵۸	***	۰.۸۱۲۰۳	**	۰.۰۸۸۵۹۸	***	۱۳.۶۵۸۹	***
	۶۰٪	۰.۰۹۹۸۵۹	***	۰.۰۸۸۹۶	***	۰.۸۲۰۵۵	***	۰.۰۹۹۷۱۸	***	۱۳.۸۹۵۶	***

(جدول ۸-): مقایسه روش پیشنهادی با درصدهای گمشدگی متفاوت در مجموعه داده Traffic در مقایسه با سایر روش‌ها  
(Table-8): Comparison of the proposed method with the percentage of different missing data in the Traffic dataset compared to other methods

روش	درصد گمشده‌گی	RMSE	T	MAE	T	$R^2$	T	MSE	T	MAPE	T
روش پیشنهادی	۳۰٪	۰.۰۸۸۱۴۵		۰.۰۳۲۸۱۷		۰.۸۵۹۰۸		۰.۰۷۷۶۹		۱۴.۱۲۸	
	۴۰٪	۰.۰۸۸۸۵۶		۰.۰۳۳۴۹۵		۰.۸۴۷۱۹		۰.۰۷۹۰۱۴		۱۴.۴۴۳	
	۵۰٪	۰.۰۸۸۰۵۵		۰.۰۳۵۹۹۳		۰.۸۴۸۰۹		۰.۰۷۷۵۳۶		۱۶.۴۲۷	
	۶۰٪	۰.۰۸۹۲۳۱		۰.۰۳۹۵۶۲		۰.۸۹۶۶۲		۰.۰۷۹۶۳۱		۱۸.۵۹۶	
<b>EM</b>	۳۰٪	۰.۰۸۷۵۲۶	***	۰.۰۶۲۳۵۲	***	۰.۸۶۹۰۸	***	۰.۰۷۶۶۰۸	***	۱۵.۱۲۸	**
	۴۰٪	۰.۰۸۸۸۹۶	**	۰.۰۶۳۴۲۵	*	۰.۸۵۷۱۹	***	۰.۰۷۹۰۲۴	***	۱۵.۴۴۳	***
	۵۰٪	۰.۰۸۹۰۵۵	***	۰.۰۶۵۹۹۳	**	۰.۸۵۸۰۹	***	۰.۰۷۹۳۰۷	***	۱۸.۴۲۷	***
	۶۰٪	۰.۰۸۵۲۳۱	***	۰.۰۶۹۵۸۲	**	۰.۸۹۵۶۲	**	۰.۰۷۲۶۴۳	***	۱۶.۵۹۶	***
<b>FLK-NN</b>	۳۰٪	۰.۰۹۶۶	***	۰.۰۷۳۳۶۶	**	۰.۸۶۷۲۹	**	۰.۰۹۳۳۱۵	**	۱۳.۵۴۴	*
	۴۰٪	۰.۰۹۶۶۵	***	۰.۰۷۳۴۱۷	**	۰.۸۶۷۲۵	**	۰.۰۹۳۴۱۲	***	۱۳.۵۸۹	
	۵۰٪	۰.۰۹۶۴۵	***	۰.۰۷۱۳۱۲	***	۰.۸۷۳۵۵	***	۰.۰۹۳۰۲۶	***	۱۳.۶۲۳	**
	۶۰٪	۰.۰۹۸۵۶	***	۰.۰۷۳۵۶	**	۰.۸۵۲۶۳	***	۰.۰۹۷۱۴۰	***	۱۳.۶۴۵	**
<b>TRiBS</b>	۳۰٪	۰.۰۹۸۱۳	**	۰.۰۷۸۵۲	***	۰.۸۸۵۲	***	۰.۰۹۶۲۹۴	**	۱۵.۴۲۲	**
	۴۰٪	۰.۰۹۸۰۷۲	***	۰.۰۷۷۰۲۵	**	۰.۸۹۸۵۶	***	۰.۰۹۶۱۸۱	***	۱۵.۴۴۳	**
	۵۰٪	۰.۰۹۹۱۰۹	***	۰.۰۸۱۴۸۲	***	۰.۸۱۲۰۵	**	۰.۰۹۸۲۲۵	***	۱۶.۲۲۷	***
	۶۰٪	۰.۰۹۹۹۳۰	***	۰.۰۸۵۹۶	***	۰.۸۱۵۲۳	*	۰.۰۹۹۸۶۰	***	۱۶.۴۲۱	***

سایر روش‌ها از نظر معیارهای RMSE، MAE و  $R^2$  برتر بود و دقت عملکرد و کارایی بالاتری داشت. در مجموعه داده‌های سری زمانی با تعداد نمونه بیشتر دقت عملکرد افزایش می‌یابد. این روش در سری‌های زمانی با طول‌های مختلف قابل اجرا است. چون روش پیشنهادی به داده‌های متوالی وابسته نیست در شرایطی که شکاف‌های بزرگ متوالی از گمشدگی ایجاد شده باشد، عملکرد مطلوبی دارد. با توجه به اینکه به‌طور معمول سری‌های زمانی مجموعه‌های حجیم هستند این روش مناسب برای پرکردن داده‌های گمشده در این مجموعه داده‌ها است.

در روش پیشنهادی کمبودهایی وجود دارد. از معایب این روش این است که بیش از بیشینه و یا کمتر از کمینه مقادیر را نمی‌تواند تخمین بزند. اجرای الگوریتم‌ها نیازمند هزینه زمانی زیادی است. پیشنهاد می‌شود روش‌های دیگری نیز برای پیدا کردن بهترین همسایگی بررسی شوند. هم چنین بر روی ویژگی‌های استخراج شده از

## ۷- نتیجه‌گیری

پرکردن داده‌های گمشده یک گام مهم در پیش‌پردازش داده‌های سری زمانی است. با توجه به این‌که داده‌ها در بازه‌های زمانی جمع‌آوری می‌شوند، احتمال وجود داده گمشده در این نوع از داده‌ها نسبت به سایر مجموعه‌ها بیشتر است. در این مقاله یک روش ترکیبی برای پرکردن داده‌های گمشده پیشنهاد شد. ایده اصلی، بهبود عملکرد و رفع محدودیت‌های درون‌یابی IDW بود. از روش خوشه‌بندی k-means برای الگویابی داده‌های سری زمانی استفاده شد. سپس بررسی شد هر داده گمشده به کدام خوشه تعلق دارد. اعضای هر خوشه به‌عنوان نزدیک‌ترین همسایگان به داده گمشده در نظر گرفته شدند؛ سپس از الگوریتم تکاملی جستجوی فاخته برای پیدا کردن توان تأثیر بهینه هر یک از همسایگان استفاده شد. روش پیشنهادی با سه روش دیگر مورد مقایسه قرار گرفت. آزمایش‌های تجربی بر روی چهار مجموعه داده مختلف اعمال شد. در بیش‌تر موارد روش پیشنهادی نسبت به

- [14] D. Mondal, D. B. Percival, "Wavelet variance analysis for gappy time series", *Annals Inst. Stat. Math*, vol.62, pp. 943–966, 2010.
- [15] K. Rehfeld, N. Marwan, J. Heitzig, "Comparison of correlation analysis techniques for irregularly sampled time series", *Nonlinear Process. Geophys*, vol.18, 2011.
- [16] P.J. Garca-Laencina, J-L Sancho-Gómez. "Pattern classification with missing data: a review", *Neural Comput*, vol.19, 2010.
- [17] R. Mazumder, T. Hastie, R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices", *Machine learning research*, vol.11, pp. 2287–2322, 2010.
- [18] Y. Koren, R. Bell, C. Volinsky, "Matrix factorization techniques for recommender systems", *Comput*, vol.42, 2009.
- [19] I. R. White, P. Royston, A. M. Wood, "Multiple imputation using chained equations: issues and guidance for practice", *Stat. medicine*, vol.30, pp. 377–399, 2011.
- [20] B. J. Wells, K. M. Chagin, A. S. Nowacki, M. W. Kattan, "Strategies for handling missing data in electronic health record derived data," *EGEMS* 1, 2013.
- [21] C. Lipton, C. Kale," Modeling Missing Data in Clinical Time Series with RNNs", *Machine Learning for Healthcare*, pp.56, 2016.
- [22] Li. Li, J. Zhang, Y. Wang, "Missing Value Imputation for Traffic-Related Time Series Data Based on a Multi-View Learning Method", *IEEE Transactions on Intelligent Transportation Systems*, vol.20, pp. 2933 – 2943, 2019.
- [23] A. McLinden, V. Fioletov, W. Shephard, N. Krotkov, "Space-based detection of missing sulfur dioxide sources of global air pollution", *Nature Geoscience*, vol.9, pp. 496–500, 2016.
- [24] R. Mahmoudvand, P. Canas, "Missing value imputation in time series using Singular Spectrum Analysis", *International Journal of Energy and Statistics*, vol. 04,165005, 2016.
- [25] N. Bokde W. Beck, "A novel imputation methodology for time series based on pattern sequence forecasting", *Pattern Recognition Letters*, vol.116, pp.88-96, 2018.
- [26] T.T. Hong Phan, E. Poisson Caillault, A. Lefebvre, A. Bigand, "Dynamic Time Warping-based imputation for univariate time series data", *Pattern Recognition Letters*, 2017.
- [27] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu."Recurrent Neural Networks for Multivariate Time Series with Missing Values", *Scientific reports*, vol.6085, pp.85-99, 2018.

داده‌های سری‌های زمانی کار شود، زیرا داده‌های خام دارای اختلال زیاد هستند.

## 8- References

## ۸- مراجع

- [1] R. H. Shumway and D. S. Stoffer, "Time series analysis and its applications: with R examples", Springer Science & Business Media, Fourth edition, 2017.
- [2] Ratanamahatana C., "Multimedia retrieval using time series representation and relevance feedback", in: *Proceedings of 8<sup>th</sup> International Conference on Asian Digital Libraries*, 2005, pp. 400–405.
- [3] C.Ratanamahatana, V. Niennattrakul, "Clustering multimedia data using time series", in: *Proceedings of the International Conference on Hybrid InformationTechnology*, ICHIT '06, 2016, pp.372–379.
- [4] M.S. Mahmoud, M.F. Emzir, "State estimation with asynchronous multi-rate multi-smart sensors", *Information Sciences*, vol.196, pp.15-27, 2012.
- [5] S. Mohamed, T. Marwala, "Neural network based techniques for estimating missing data in databases", pp. 27-32, 2005.
- [6] W. Qiao, Z. GAO, R.G. Harley, "Continuous on-line identification of nonlinear plants in power systems with missing sensor measurements", *IEEE*, pp. 1729-1734, 2005.
- [7] J. Honaker, G. King, "What to do about missing values in time-series cross-section data", *American Journal of Political Science*, vol.54 (2), pp.561–581, 2010.
- [8] J. Lin, E. Keogh, S. Lonardi, J. Lankford, D. Nystrom, "Visually mining and monitoring massive time series", in: *Proceedings of 2004ACM SIGKDD International Conference on Knowledge Discovery and data Mining – KDD '04*, 2004, 460-475.
- [9] R.J.A. Little, D.B. Rubin, "Statistical analysis with missing data," 3rd Edition, 2014.
- [10] M. Amiri, R. Jensen, "Missing data imputation using fuzzy-rough methods", *Neurocomputing*, vol.196, pp.15-27, 2016.
- [11] C.K. Enders, "Applied Missing Data Analysis", Guilford Press. ISBN 978-1-60623-639-0 .2010.
- [12] D. M. Kreindler, C. J. Lumsden, "The effects of the irregular sample and missing data in time series analysis", *Nonlinear Dynamics Psychology and Life Sciences*, vol.10(2), pp.187-214, 2012.
- [13] C.De Boor, E. Mathématicien, "A practical guide to splines", *Mathematical Sciences*, vol. 27, 2005.



دانشگاه تربیت دبیر شریعتی و مدرک کارشناسی ارشد خود را نیز در رشته مهندسی کامپیوتر-نرم‌افزار، در سال ۱۳۹۹ از دانشگاه تربیت دبیر شهید رجایی دریافت کرد. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: پایگاه داده تحلیلی، داده‌کاوی و پیش‌پردازش داده. نشانی رایانامه ایشان عبارت است از:

**fmirabolghasemi@yahoo.com**



**نگین دانشپور** مدرک کارشناسی خود را در سال ۱۳۷۷ از دانشگاه شهید بهشتی و درجه کارشناسی ارشد و دکترای خود را در رشته مهندسی کامپیوتر-نرم‌افزار از دانشگاه صنعتی امیر کبیر در سال‌های ۱۳۸۰ و ۱۳۸۹ دریافت کرده است. در حال حاضر وی عضو هیئت علمی و دانشیار دانشکده مهندسی کامپیوتر دانشگاه تربیت دبیر شهید رجایی است. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: داده‌کاوی و پیش‌پردازش و مدیریت داده‌ها، پایگاه داده تحلیلی و سامانه‌های تصمیم‌یار. نشانی رایانامه ایشان عبارت است از:

**ndaneshpour@sru.ac.ir**

efficient k-means clustering algorithm: analysis and implementation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, pp.881-892.2002.

- [53] Celebi, M. E., H. A. Kingravi, and P. A. Vela “A comparative study of efficient initialization methods for the k-means clustering algorithm”. *Expert Syst, Appl.* 40(1), 200–210.2003.
- [54] R. Rajabioun, “Cuckoo Optimization Algorithm”, *Applied Soft Computing*, Vol.11, No.8, pp. 5508- 5518, 2011.
- [55] F. Petitjean, A. Ketterlin, P. Gancarski, “A global averaging method for dynamic time warping, with applications to clustering”, *Pattern Recognition*, vol. 44, no.3, pp. 678-693,2011.
- [56] M.Lichman, "UCI machine learning repository". <http://archive.ics.uci.edu/ml> .2013.
- [57] W.L. Junger, A.P. de Leon, "Imputation of missing data in time series for air pollutants", *Atmospheric Environment*, vol.102, pp. 96-104,2015.
- [58] S.A Rahman, Y. Huang, J. Claassen, “Combining Fourier and Lagged *k*-Nearest Neighbor Imputation for Biomedical Time Series Data”, *Nathaniel Heintzman, and Samantha Kleinberg*, *J Biomed Inform*, vol.58, pp.198–207, 2016.
- [59] M. G. Rahman, M. Z. Islam, "Missing value imputation using a fuzzy clustering-based EM approach", *Knowledge and Information Systems*, vol.46 (2), pp. 389–422, 2016.
- [60] R. Deb, A. Liew. "Missing value imputation for the analysis of incomplete traffic accident data," *Information Sciences*, vol.339, pp274–289 .2016.
- [61] M.E. Quinteros, S. Lu, C. BlazquezCárdenas-R, J.P., X. Ossa, J.-M. DelgadoSaborit, R.M. Harrison, P. Ruiz-Rudolph, "Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile", *Atmospheric Environment*, 2018.
- [62] B. Golden, B. Grand, F. Rossi. "Mean Absolute Percentage Error for regression models", *Neurocomputing*, vol.192, pp.38–48. 2016.
- [63] M. Misuraca, M. Spano, S. Balbi, “BMS: An improved Dunn index for Document Clustering validation”, *Communications in Statistics*, pp. 0361-0926, 2018.



**سیده فاطمه میرابوالقاسمی** مدرک کارشناسی خود را در رشته مهندسی کامپیوتر-نرم‌افزار در سال ۱۳۹۲ از

