استفاده از خوشه‌بندی و رویکردی ترکیبی برای پرکردن مقادیر جاافتاده عددی

امیرمسعود سفیدیان و نگین دانشپور

چکیده: تخمین مقادیر جاافتاده یک گام مهم در پیش‌پردازش داده‌ها است. در این مقاله یک رویکرد دومرحله‌ای برای پرکردن مقادیر جاافتاده عددی ارائه شده است. در مرحله اول داده‌ها خوشه‌بندی می‌شوند و در مرحله دوم داده‌های جاافتاده درون هر خوشه با استفاده از یک روش ترکیبی از k نزدیک‌ترین همسایه وزن‌دار و رگرسیون خطی تخمین زده می‌شوند. از معیار همبستگی بین صفات در هر خوشه برای تعیین روش پرکردن داده‌های جاافتاده استفاده می‌شود. کیفیت پرکردن مقادیر جاافتاده با استفاده از معیار میانگین مربعات خطا سنجیده می‌شود. تأثیر پارامترهای مختلف بر میزان خطای داده‌های تخمین زده شده بررسی می‌گردد. عملکرد روش ارائه‌شده برای تخمین داده‌های جاافتاده بر روی پنج مجموعه داده نیز بررسی می‌شود. در نهایت عملکرد روش ارائه‌شده با چهار روش پرکردن با مقدار میانگین، روش تخمین با شبکه عصبی پرسپترون چندلایه (MLP)، روش پرکردن با خوشه‌بندی c-means فازی و روش k خوشه و نزدیک‌ترین همسایه مبتنی بر دسته (CKNNI) مقایسه می‌شود. نتایج به دست آمده نشان داده که خطای تخمین مقادیر جاافتاده در روش ارائه‌شده کمتر از خطا در دیگر روش‌های مقایسه‌شده است.

کلیدواژه: **رگرسیون، مقادیر جاافتاده، نزدیک‌ترین همسایگان، همبستگی.**

۱- مقدمه

مجموعه داده‌های دنیای واقعی معمولاً کامل نیستند. ممکن است مقدار یک یا چند صفت از صفات چندین نمونه به دلایل مختلف جاافتاده و نامعلوم باشند. وجود مقادیر جاافتاده اجتناب‌ناپذیر است. پیش‌پردازش یک گام مهم پیش از داده‌کاوی و یادگیری ماشین و یکی از مراحل پیش‌پردازش رسیدگی به مقادیر جاافتاده است. وجود مقادیر جاافتاده در مجموعه داده می‌تواند دقت الگوریتم‌های یادگیری مانند خوشه‌بندی و دسته‌بندی را به شدت کاهش دهد. مقدار جاافتاده می‌تواند به دلایل مختلفی مثل خطای انسانی، در دسترس نبودن مقادیر و یا خطای تجهیزات ثبت اطلاعات رخ دهند.

سه دسته‌بندی معمولاً برای داده‌های جاافتاده انجام می‌شود [۱] تا [۴]:

۱) جاافتاده به طور کاملاً تصادفی (MCAR)[1]:

در این حالت احتمال این که برای هر نمونه مقدار یک صفت مثل A_j معلوم یا جاافتاده باشد (مقدار تهی یا NULL داشته باشد) وابسته به هیچ یک از دیگر صفات نیست و کاملاً مستقل از دیگر صفات است یعنی

$$P(A_j = \text{NULL} \mid A_1, \ldots, A_M) = P(A_j = \text{NULL}) \qquad (۱)$$

که در آن M تعداد صفات هر نمونه است.

۲) جاافتاده به طور تصادفی (MAR)[2]:

در این حالت احتمال این که برای هر نمونه مقدار یک صفت مثل A_j معلوم یا جاافتاده باشد ممکن است به دیگر صفات وابسته باشد، ولی به مقدار خود آن صفت وابسته نیست. بنابراین در این حالت مقدار جاافتاده از دیگر صفات قابل محاسبه است یعنی

$$P(A_j = \text{NULL} \mid A_1, \ldots, A_M) = P(A_j = \text{NULL} \mid A_1, \ldots, A_{j-1}, A_{j+1}, \ldots, A_M) \qquad (۲)$$

که در آن M تعداد صفات هر نمونه است.

۳) جاافتاده به طور غیر تصادفی (MNAR)[3]:

در این حالت احتمال این که برای هر نمونه مقدار یک صفت مثل A_j معلوم یا جاافتاده باشد ممکن است به مقدار خود آن صفت وابسته باشد. بنابراین در این حالت یک صفت جاافتاده به طور مستقیم از روی صفات دیگر قابل محاسبه نیست. در عمل، این که داده جاافتاده در کدام دسته‌بندی ذکرشده قرار می‌گیرد معمولاً غیر ممکن است زیرا نحوه ایجادشدن مقدار جاافتاده نامعلوم است [۳].

رویکردهای ساده متفاوتی برای رسیدگی به مقادیر جاافتاده وجود دارد:

- حذف رکوردهایی با مقادیر جاافتاده
- پرکردن دستی مقادیر جاافتاده
- پرکردن با مقدار ثابت
- پرکردن با مقدار میانگین (یا مد)

در رویکرد اول اگر تعداد مقادیر جاافتاده بالا باشد بسیاری از داده‌های موجود را بیهوده هدر می‌دهد. رویکرد دوم در عمل نشدنی و یا بسیار زمان‌بر است. در رویکرد سوم فرض می‌شود که همه مقادیر جاافتاده مقدار یکسانی دارند که اصولاً فرض غلطی است. روش چهارم استفاده از مقدار میانگین برای صفات عددی و مقدار مُد برای داده‌های نمادین است. مقدار میانگین بسیار حساس به داده‌های پَرت است. مشکل دیگر این روش‌ها این است که در این روش‌ها از روابط احتمالی موجود بین صفات استفاده نمی‌شود. بنابراین این رویکردهای ساده نمی‌توانند دقت کافی برای پرکردن مقادیر جاافتاده را فراهم کنند و بنابراین روش‌های پیچیده‌تری برای رفع مقادیر جاافتاده ارائه شده است. توجه به ارتباطات موجود احتمالی در زیرمجموعه‌های یک مجموعه داده و بهره‌گیری از روشی مناسب در صورت عدم وجود ارتباطات قوی به صورت ترکیبی برای تخمین مقادیر جاافتاده می‌تواند پیشنهاد مناسبی برای استفاده از اطلاعات داده‌های موجود برای تخمین مقادیر جاافتاده باشد.

در این پژوهش رویکردی جدید مبتنی بر خوشه‌بندی داده‌ها و استفاده

این مقاله در تاریخ ۱۵ آذر ماه ۱۳۹۵ دریافت و در تاریخ ۲ مرداد ماه ۱۳۹۶ بازنگری شد.

امیرمسعود سفیدیان، دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهیدرجایی، تهران، (email: amirmasoud.sefidian@srttu.edu).

نگین دانشپور، دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهیدرجایی، تهران، (email: ndaneshpour@srttu.edu).

1. Missing Completely at Random
2. Missing at Random
3. Missing Not at Random

از ترکیب روش k نزدیک‌ترین همسایه وزن‌دار (WKNN)[1] و رگرسیون خطی[2] برای پرکردن مقادیر جاافتاده عددی ارائه شده است. ویژگی اصلی روش ارائه‌شده استفاده ترکیبی و توأم از همبستگی موجود بین داده‌ها و فاصله بین نمونه‌ها برای پرکردن مقادیر جاافتاده است.

در روش‌های موجود برای پرکردن مقادیر جاافتاده اصولاً دو نوع رویکرد وجود دارد. رویکرد اول روش‌هایی هستند که فرض می‌کنند روابط قوی بین صفات در مجموعه داده وجود دارد و بنابراین تلاش می‌کنند تا روابط موجود را مدل‌سازی کنند. از جمله این روش‌ها می‌توان به روش‌های آماری مانند رگرسیون اشاره کرد. مشکل اصلی استفاده از این روش‌ها به تنهایی این است که ممکن است در داده‌های ورودی روابط قوی وجود نداشته باشد و بنابراین مدل ساخته‌شده دقت پایینی را ارائه خواهد داد. به علاوه در اکثر روش‌های ارائه‌شده مبتنی بر روابط قوی از معیار مشخص و دقیقی برای سنجش روابط بین صفات، مانند ضریب همبستگی[3] استفاده نشده است.

دسته دیگری از روش‌ها مانند روش‌های مبتنی بر k نزدیک‌ترین همسایگان، کاملاً از روابط بین صفات چشم‌پوشی می‌کنند و فقط از میزان نزدیکی نمونه‌ها استفاده می‌کنند. در این روش‌ها اطلاعات بسیار مفیدی مانند قدرت استنتاج یک مقدار جاافتاده از طریق روابط موجود با دیگر صفات نادیده گرفته می‌شود. بنابراین در این مقاله ترکیبی از دو دسته روش ذکرشده به منظور بهره‌وری از مزایای آنها ارائه می‌شود. با توجه به مطالعات انجام‌شده، پیش از این ترکیبی از این دو دسته روش ارائه نشده است. در روش ارائه‌شده ابتدا سعی می‌شود تا با استفاده از خوشه‌بندی بتوان روابط قوی موجود بین نمونه‌ها را پیدا کرد. علت این کار این است که ممکن است روابط در زیرمجموعه‌ای از مجموعه داده ورودی قوی‌تر از کل مجموعه داده باشد. در کارهای پیشین از خوشه‌بندی برای گروه‌بندی نمونه‌های نزدیک به هم استفاده شده است. در این مقاله از خوشه‌بندی برای یافتن نمونه‌هایی با همبستگی بالا استفاده شده است. پس از خوشه‌بندی، میزان ارتباط بین صفات با معیاری دقیق و کمّی به نام ضریب همبستگی درون هر خوشه اندازه‌گیری می‌شود. لازم به ذکر است که این معیار پیش از این برای این منظور استفاده نشده است. در نهایت بسته به این که آیا روابط به حد کافی قوی هستند یا خیر، از یکی از دو دسته روش بالا استفاده می‌شود.

نتایج بررسی‌شده بر روی پنج مجموعه داده نشان داده است که رویکرد پیشنهادی، دقت بیشتری در مقایسه با چهار روش پرکردن با مقدار میانگین، روش تخمین با شبکه عصبی پرسپترون چندلایه (MLP)[4] [۵]، روش تخمین با خوشه‌بندی c-means فازی [۶] و روش k خوشه و نزدیک‌ترین همسایه مبتنی بر دسته (CKNNI)[5] [۷] دارد.

بخش‌های بعدی این مقاله بدین شرح است. در بخش دوم مروری بر روش‌های گذشته بیان می‌شود. در بخش سوم جزئیات رویکرد پیشنهادی ارائه می‌گردد و پیچیدگی زمانی رویکرد پیشنهادی تحلیل می‌شود. در بخش چهارم نتایج و آزمایش‌ها مطرح شده است و در نهایت در بخش پنجم نتیجه‌گیری بیان می‌شود.

۲- پیشینه تحقیق

در این بخش خلاصه‌ای از رویکردهای موجود برای پرکردن مقادیر جاافتاده بیان می‌شود. روش‌های مختلفی مثل روش‌های آماری، استفاده از یادگیری ماشین و داده‌کاوی برای پرکردن مقادیر جاافتاده ارائه شده است.

دسته‌ای از روش‌ها با استفاده از قوانین انجمنی[6] داده‌های جاافتاده را پر می‌کنند. محدودیت اصلی این روش‌ها این است که تنها قابل استفاده برای داده‌های غیر عددی هستند. در [۸] از یک رویکرد مبتنی بر قوانین انجمنی برای پرکردن مقادیر جاافتاده استفاده شده است. ابتدا مجموعه داده رابطه‌ای به مجموعه داده تراکنشی تبدیل می‌شود تا بتوان از آن قوانین انجمنی را استخراج نمود. پس از یافتن قوانین یافت‌شده از مجموعه داده تراکنشی برای هر مقدار جاافتاده، مجموعه قوانینی که در سمت راست آنها همان صفت جاافتاده حضور دارند، انتخاب می‌شوند. برای هر یک از قوانین انتخاب‌شده امتیازی محاسبه می‌گردد. قانونی با بیشترین امتیاز به عنوان قانون برگزیده انتخاب شده و از مقدار سمت راست آن قانون به عنوان مقدار جایگزین صفت مقدار صفت جاافتاده استفاده می‌شود. از نقاط قوت این روش می‌توان به دخیل‌کردن معیار وابستگی دو طرف قوانین انجمنی و طول قوانین و حمایت از مقادیر جاافتاده چندگانه اشاره کرد.

در [۹] قوانین یافت‌شده رتبه‌بندی می‌شوند و قانونی با بهترین رتبه به عنوان قانون انتخاب‌شده برای پرکردن مقدار جاافتاده استفاده خواهد شد. قوانین گزینش‌شده به ترتیب بر اساس طول قانون، درصد اعتماد[7] و درصد درصد پشتیبانی[8] رتبه‌بندی می‌شوند. از نقاط قوت این روش می‌توان به دخیل‌کردن طول قوانین و رتبه‌بندی قوانین اشاره کرد. در نظر گرفتن تنها بهترین قانون هنگام پرکردن مقادیر از نقاط ضعف این روش می‌باشد.

از شبکه‌های عصبی نیز برای پرکردن مقادیر جاافتاده استفاده شده است. در [۵] با استفاده از شبکه عصبی پرسپترون چندلایه (MLP) داده‌های جاافتاده برای هجده مجموعه داده تخمین زده شده‌اند. همچنین تأثیر روش‌های مختلف یادگیری و پارامترهای مختلف شبکه عصبی بر روی دقت نهایی نیز بررسی شده‌اند. در [۱۰] با استفاده از طراحی و آموزش یک شبکه عصبی نگاشت خود سازمان‌ده (SOM)[9] مقادیر جاافتاده تخمین زده شده‌اند. مشکل اصلی شبکه‌های عصبی زمان‌بر بودن فرایند یادگیری و ساخت مدل بهینه برای یک مجموعه داده نسبت به دیگر روش‌ها و همچنین قابلیت تفسیر[10] پایین است [۱۱]. به علاوه، اصولاً آموزش مناسب شبکه‌های عصبی نیاز به تعداد نمونه‌های ورودی بالایی دارد و بنابراین هنگامی که تعداد نمونه‌های جاافتاده بالا است، دقت تخمین این روش‌ها کاهش می‌یابد.

روش k نزدیک‌ترین همسایه یکی از معروف‌ترین و پایه‌ای‌ترین روش‌ها برای تخمین مقادیر جاافتاده است. سادگی و نتایج قابل قبول آن نسبت به دیگر روش‌ها از ویژگی‌های آن است. از دیگر مزایای این روش قابلیت استفاده برای هم صفات عددی و هم صفات غیر عددی با تغییر تابع سنجش فاصله است. همچنین در این روش زمانی صرف ساخت مدل برای تخمین داده‌های جاافتاده نمی‌شود [۷]. این ویژگی‌ها سبب شده که در این مقاله نیز از حالت وزن‌دار این روش به عنوان یک روش تخمین

1. Weighted K-Nearest Neighbor
2. Linear Regression
3. Correlation Coefficient
4. Multi-Layer Perceptron
5. Class-Based K-Clusters Nearest Neighbor Imputation
6. Association Rules
7. Confidence
8. Support
9. Self Organizing Map
10. Interpretability

استفاده شود.

در [۱۲] از یک رویکرد مبتنی بر روش k نزدیک‌ترین همسایه برای پرکردن مقادیر جاافتاده استفاده شده است. در این رویکرد ابتدا درجه جاافتادگی هر نمونه که برابر است با تعداد صفات جاافتاده در هر نمونه، محاسبه می‌گردد. نمونه‌هایی که درجه جاافتادگی آنها بزرگ‌تر از دو است از ادامه پردازش خارج می‌شوند و سپس مقادیر جاافتاده با مقادیر نزدیک‌ترین نمونه‌های کامل پر می‌شوند. از نقاط قوت این روش می‌توان به سادگی رویکرد ارائه‌شده و قدرت تعمیم بالا برای کاربردهای مختلف با تغییر معیار شباهت اشاره کرد. عدم توانایی در پرکردن مقادیر جاافتاده چندگانه، حساسیت به داده‌های پرت و حساسیت بالا به تابع شباهت از نقاط ضعف این روش می‌باشد. همچنین لزوماً بررسی فقط نزدیک‌ترین نمونه، همواره بهترین گزینه نیست.

در [۱۳] از یک رویکرد پرکردن k نزدیک‌ترین همسایه تکراری مبتنی بر گری[1] به نام GBKII[2] استفاده شده است. رویکرد مطرح‌شده در یک فرایند تکراری بر روی مقادیر جاافتاده، تک به تک مقادیر جاافتاده را پر می‌نماید. از مزیت‌های این روش امکان پرکردن مقادیر جاافتاده چندگانه است. به‌علاوه چون سنجش شباهت فقط بین نمونه‌های با برچسب یکسان صورت می‌گیرد، پیچیدگی زمانی این روش پایین است..

در [۷] از یک نسخه بهبودیافته روش k نزدیک‌ترین همسایه و استفاده از خوشه‌بندی داده‌های قرارگرفته در یک دسته (کلاس) به نام CKNNI برای تخمین مقادیر جاافتاده استفاده شده است. مشکل اصلی این روش نیاز به دانستن برچسب کلاس تمام نمونه‌ها است. در واقع اگر کلاس نمونه‌ها مشخص نباشند نمی‌توان از روش ذکرشده استفاده نمود.

در [۱۴] دقت رویکرد نزدیک‌ترین همسایه بر روی چهار مجموعه داده با دو معیار فاصله مَنهَتن[3] و اقلیدسی برای پرکردن مقادیر جاافتاده بررسی و نتایج با روش پرکردن با میانگین مقایسه شده است. رویکرد نزدیک‌ترین همسایه دقت بیشتری در تخمین مقادیر جاافتاده در برابر روش میانگین داشته است.

در [۱۵] به کمک جمع‌سپاری[4] مقادیر جاافتاده پر می‌شوند. مشکل اصلی این روش به طور کامل خودکار نبودن و تأثیرگذاری عامل انسانی در دقت نهایی داده‌های پرشده است.

در [۱۶] ابتدا داده‌ها خوشه‌بندی شده و سپس هر مقدار جاافتاده با میانگین دو مقدار جایگزین می‌شود. مقدار اول فاصله نمونه جاافتاده تا مرکز خوشه‌ای است که نمونه جاافتاده به آن تعلق دارد. مقدار دوم برابر با مقدار صفتی از مرکز خوشه است که متناظراً در نمونه جاافتاده، نامعلوم است. چشم‌پوشی از روابط احتمالی بین نمونه‌های موجود در یک مجموعه داده مشکل اصلی این روش است.

در [۱۷] با استفاده از یک خوشه‌بندی به نام خوشه‌بندی همسایگان مشترک برای هدف پرکردن داده‌های جاافتاده استفاده شده است. از نقاط قوت این روش می‌توان به استفاده از معیار شباهت برای صفات ترکیبی و دخیل‌کردن معیار جدیدی برای سنجش شباهت نمونه‌ها اشاره کرد. نحوه توزیع داده‌ها بر تعداد همسایگان مشترک در این روش تأثیرگذار است و ممکن است ماتریس همسایگان مشترک تنک[5] شود.

در [۱۸] از خوشه‌بندی‌های k-means و c-means فازی برای پرکردن مقادیر جاافتاده استفاده شده است. در [۱۹] یک رویکرد دومرحله‌ای برای پرکردن مقادیر جاافتاده ارائه شده است. این رویکرد شامل یک تخمین اولیه از مقادیر جاافتاده توسط نزدیک‌ترین مرکز خوشه به روش خوشه‌بندی k-means و یک گام پالایش و بهبود مقدار تخمین زده شده از گام قبلی، توسط شبکه MLP می‌باشد. در [۲۰] در گام اول به جای استفاده از خوشه‌بندی k-means از خوشه‌بندی c-means فازی استفاده شده که در آن هر نمونه با یک احتمال متعلق به هر خوشه است. این روش نتایج بهتری را به همراه داشته است. از نقاط قوت این روش‌ها می‌توان به بهبود k-means برای پرکردن مقادیر جاافتاده اشاره کرد. نیاز به آموزش یک شبکه MLP برای هر مقدار جاافتاده و آموزش مجدد روی کل مجموعه آموزشی، تعیین ساختار (به هم بندی[6] و وزن‌ها) بهینه شبکه شبکه MLP برای هر بار آموزش برای پرکردن هر مقدار جاافتاده از نقاط ضعف این روش می‌باشد.

در [۲۱] یک رویکرد ترکیبی از قوانین انجمنی و k نزدیک‌ترین همسایه برای پرکردن مقادیر جاافتاده ارائه شده است. از نقاط قوت این روش می‌توان به بهره‌گیری از قدرت ترکیب دو روش متفاوت اشاره کرد. در واقع در این روش سعی در استفاده از ارتباطات و فواصل موجود برای پرکردن مقادیر جاافتاده شده و این روش قابلیت استفاده برای داده‌های عددی را به طور مستقیم ندارد (به دلیل استفاده از قوانین انجمنی).

در [۲۲] روشی مبتنی بر الگوریتم EM[7] و درخت تصمیم به نام DMI[8] DMI[8] برای پرکردن مقادیر جاافتاده طراحی شده است. ایده اصلی این روش این است که وابستگی بین افرازهای مجموعه داده قوی‌تر از وابستگی بر کل مجموعه داده است و بنابراین الگوریتم EM بر روی بخش‌های متفاوت مجموعه داده بهتر عمل خواهد کرد و در نهایت نتیجه به دست آمده بهتر خواهد شد. افرازبندی مجموعه داده توسط درخت تصمیم صورت می‌پذیرد. از نقاط قوت این روش توجه به این است که وابستگی بین صفات لزوماً در کل مجموعه داده برقرار نیست. مناسب‌بودن برای مجموعه داده‌هایی که وابستگی بین صفات آنها وجود دارد از نقاط ضعف این روش است.

رگرسیون نیز به عنوان یک روش پایه برای تخمین مقادیر جاافتاده می‌تواند مورد استفاده قرار گیرد. زمانی که رابطه بین دو متغیر قوی‌تر باشد، دقت رگرسیون نیز بیشتر خواهد بود. با استفاده از محاسبه معادله رگرسور خطی می‌توان داده‌های جاافتاده را از طریق داده‌های موجود تخمین زد. در [۳] از یک رویکرد افزایشی و با استفاده از تولید دنباله‌ای از رگرسورها تخمین داده‌های جاافتاده انجام می‌شود. مشکل اصلی روش‌های مبتنی بر رگرسیون، ضعف عملکرد آنها در صورت عدم وجود روابط قوی بین صفات پیش‌بینی کننده و صفت پیش‌بینی شده است.

در [۲۳] دو راهبرد برای بهبود دقت پرکردن مقادیر جاافتاده ارائه شده است. ایده اصلی در هر دو راهبرد، انجام عمل انتخاب نمونه[9] است. در این این عمل سعی می‌شود تا نمونه‌های شامل داده‌های پَرت و داده‌های مغشوش[10] از مجموعه داده حذف شوند تا کیفیت داده‌ها پیش از عملیات پرکردن مقادیر جاافتاده افزایش یابد. در واقع مجموعه داده از یک فیلتر حذف داده‌های پرت و مغشوش عبور داده می‌شود. در هر دو راهبرد

1. Grey
2. Grey-Based K-NN Iteration Imputation
3. Manhattan
4. Crowdsourcing
5. Sparse
6. Topology
7. Expectation Maximization
8. Decision Tree Based Missing Value Imputation Technique
9. Instance Selection
10. Noisy Data

جدول ۱: بررسی کلی مشکلات روش‌های موجود.

رویکرد پرکردن	مشکل اصلی
روش‌های مبتنی بر شبکه عصبی	کاهش چشم‌گیر دقت تخمین هنگامی که تعداد نمونه‌های جاافتاده بالا است
روش‌های مبتنی بر قوانین انجمنی	غیر قابل اعمال مستقیم برای تخمین مقادیر جاافتاده عددی
روش‌های مبتنی بر رگرسیون و الگوریتم EM	عملکرد ضعیف وقتی ارتباطات قوی بین صفات وجود ندارد
روش‌های مبتنی بر k نزدیک‌ترین همسایگان	عدم توجه و استفاده از روابط موجود بین صفات

عملیات انتخاب نمونه از طریق الگوریتم DROP[1] [۲۴] انجام شده است. تفاوت دو راهبرد پیشنهادی در ترتیب انجام عملیات انتخاب نمونه و پرکردن داده‌ها است. انجام عملیات انتخاب نمونه منجر به بهبود دقت حاصل از دسته‌بندی داده‌های تخمین زده شده است.

در [۲۵] با استفاده از ترکیب نظریه مجموعه‌های رافِ فازی[2] و k نزدیک‌ترین همسایگان، الگوریتمی به نام پرکردن نزدیک‌ترین همسایگان راف فازی (FRNNI)[3] ارائه شده است. در این الگوریتم از تقریب‌های پایین و بالا[4] در نزدیک‌ترین نمونه‌ها، نسبت به نمونه جاافتاده برای پرکردن مقادیر نامعلوم استفاده می‌شود. علت اصلی استفاده از نظریه مجموعه‌های راف فازی در این روش مناسب‌بودن این رویکردها در تعامل با محیط‌هایی با حضور عدم قطعیت[5] است و بنابراین انتخاب خوبی برای مسئله پرکردن مقادیر جاافتاده است.

با مطالعات انجام‌گرفته نیاز به روشی برای پرکردن مقادیر جاافتاده با توجه به همبستگی احتمالی موجود بین صفات و همچنین استفاده از روش جایگزینی مناسب برای حالتی که همبستگی کافی بین صفات وجود ندارد ضروری می‌رسد. بنابراین رویکردی در این مقاله پیشنهاد می‌شود که ابتدا با استفاده از رگرسیون خطی، سعی در استفاده از روابط موجود بین داده‌ها برای پرکردن مقادیر جاافتاده می‌نماید. تلاش می‌شود که روابط در زیرمجموعه‌هایی از مجموعه داده اصلی یافت شوند. دلیل این امر این است که شاید روابط قوی بر روی کل مجموعه داده یافت نشود در حالی که بین زیرمجموعه‌ای از نمونه‌ها، ارتباطات قوی‌تری وجود داشته باشد. در صورت یافت‌نشدن میزان ارتباط کافی درون زیرمجموعه داده‌ها از روش پایه نزدیک‌ترین همسایگان وزن‌دار استفاده می‌شود. مرجع [۲۲] اگرچه با استفاده از درخت تصمیم سعی در افزایش همبستگی نمونه‌ها دارد با این حال از معیاری دقیق برای سنجش صحت فرضیه افزایش همبستگی درون هر افراز از نمونه‌ها استفاده نمی‌کند. روش‌های مبتنی بر رگرسیون به تنهایی این نکته که ممکن است ارتباط بین صفات در یک مجموعه داده قوی نباشند را در نظر نمی‌گیرند و بنابراین دقت عملکرد آنها در این شرایط به شدت کاهش می‌یابد. در روش‌های مبتنی بر k نزدیک‌ترین همسایه، اطلاعات راجع به همبستگی بین صفات به طور کلی نادیده گرفته می‌شوند. در روش ترکیبی ارائه‌شده سعی گردیده تا تمام این موارد به طور توأم برای بهبود کیفیت تخمین در نظر گرفته شوند. جدول ۱ مشکلات اصلی در روش‌های موجود را به اختصار نمایش می‌دهد. روش پیشنهادی سعی دارد تا به طور هم‌زمان مشکلات دو دسته از روش‌های مبتنی بر رگرسیون و روش‌های مبتنی بر k نزدیک‌ترین همسایگان را برطرف سازد.

۳- رویکرد پیشنهادی

در این بخش بیانی دقیق و رسمی از مراحل رویکرد پیشنهادی ارائه و سپس پیچیدگی رویکرد پیشنهادی تحلیل می‌شود.

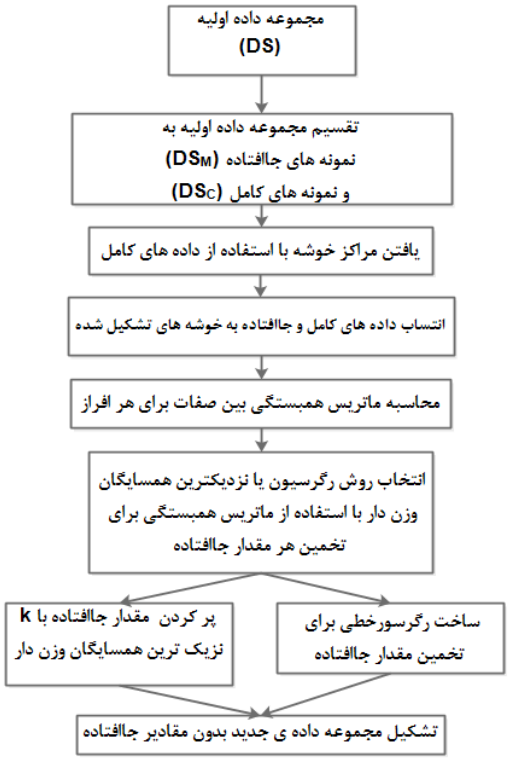
۳-۱ مراحل رویکرد پیشنهادی

روش ارائه‌شده برای پرکردن داده‌های جاافتاده شامل دو گام اصلی است. در گام اول داده‌های کامل (بدون مقادیر جاافتاده) خوشه‌بندی می‌شوند. هدف اصلی از خوشه‌بندی این است که ممکن است همبستگی قوی بر روی کل مجموعه داده برقرار نباشد در حالی که بین زیرمجموعه‌ای از نمونه‌ها، ارتباطات قوی وجود داشته باشد. بنابراین سعی می‌شود با استفاده از خوشه‌بندی، همبستگی‌های احتمالی موجود در زیرمجموعه‌های مجموعه داده اصلی را پیدا نمود. در گام بعدی مقادیر جاافتاده در هر خوشه با استفاده از روشی مرکب از k نزدیک‌ترین همسایه وزن‌دار و رگرسیون خطی پر خواهند شد. از یک حد آستانه ($0 \le t \le 1$) بر روی مقادیر ماتریس همبستگی بین صفات برای تعیین روش تخمین مقدار جاافتاده استفاده می‌شود. اگر صفتی با همبستگی کافی به مقدار صفت جاافتاده پیدا شود از روش رگرسیون خطی و در غیر این صورت از روش k نزدیک‌ترین همسایه وزن‌دار استفاده می‌گردد. در واقع ایده اصلی این است که در صورت وجود همبستگی بین داده‌ها از این همبستگی تا حد ممکن استفاده شود و در غیر این صورت از نزدیک‌ترین نمونه‌ها برای پرکردن مقادیر جاافتاده استفاده شود. خوشه‌بندی داده‌ها در این مقاله با روش خوشه‌بندی k-means انجام می‌شود. شکل ۱ روند کلی رویکرد پیشنهادی را نشان می‌دهد. شبه‌کد الگوریتم پیشنهادی در شکل ۲ نمایش داده شده است. در ادامه جزئیات مراحل مختلف رویکرد پیشنهادی شرح داده می‌شود.

فرض می‌شود مجموعه داده رابطه‌ای DS شامل N رکورد بر روی شمای $(A_1, A_2, \ldots, A_M)$ تعریف شده و هر نمونه شامل M صفت است. x_i^j مقدار صفت j ام نمونه i ام را نمایش می‌دهد و تابع $d(x,y)$ بیانگر فاصله دو نمونه x و y است.

در ابتدا مجموعه داده اصلی (DS) به دو مجموعه داده داده‌های کامل $(DS_C = \{c_1, c_2, \ldots, c_P\})$ (رکوردهایی بدون مقادیر جاافتاده) و مجموعه داده‌های با مقادیر جاافتاده $(DS_M = \{m_1, m_2, \ldots, m_Q\})$ تقسیم می‌شود (خط ۱ از شبه‌کد). سپس مجموعه DS_C با استفاده از خوشه‌بندی به c تا خوشه $S_1, S_2, \ldots, S_c$ تقسیم می‌شود و بنابراین c مرکز خوشه به دست می‌آید (خط ۲). هر یک از عناصر DS_C با توجه به میزان نزدیکی به مراکز خوشه به یک خوشه منتسب می‌شوند (خط ۳). هر یک از عناصر DS_M نیز با توجه به میزان نزدیکی به مراکز خوشه به یک خوشه منتسب می‌شود (خط ۴). قابل ذکر است که فاصله هر نمونه جاافتاده بدون در نظر گرفتن صفات جاافتاده محاسبه می‌گردد.

1. Decremental Reduction Optimization Procedure
2. Fuzzy Rough Sets Theory
3. Fuzzy-Rough Nearest Neighbor Imputation
4. Lower and Upper Approximations
5. Uncertainty



شکل ۱: روند کلی رویکرد پیشنهادی.

ضریب همبستگی پیرسون[1] برای جفت صفات (A_i, A_j) که $i \neq j$ است برای نمونه‌های هر خوشه محاسبه شده و نتایج در یک ماتریس همبستگی برای آن خوشه ذخیره می‌گردد (خط ۵). برای خوشه f ام، ماتریس همبستگی به صورت (۳) خواهد بود

$$COR(S_f) = \begin{bmatrix} ۱ & \rho_{۱,۲} & \cdots & \rho_{۱,M} \\ \rho_{۲,۱} & ۱ & \cdots & \rho_{۲,M} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{M,۱} & \rho_{M,۲} & \cdots & ۱ \end{bmatrix} \qquad (۳)$$

که در آن هر عنصر از (۴) محاسبه می‌گردد

$$\rho_{i,j} = \left| \frac{\mathrm{cov}(A_i, A_j)}{\sqrt{\sigma(A_i)\sigma(A_j)}} \right| \quad , \quad ۰ \le \rho_{i,j} \le ۱ \qquad (۴)$$

که $\mathrm{cov}(A_i, A_j)$ کواریانس بین مقادیر صفت A_i و صفت A_j و $\sigma(A_i)$ و $\sigma(A_j)$ به ترتیب برابر با واریانس مقادیر صفت A_i و صفت A_j هستند. $|.|$ بیانگر عملگر قدر مطلق است. چون میزان همبستگی بین صفات در این روش مهم است نه نوع همبستگی (همبستگی مثبت یا منفی) از قدر مطلق مقدار ضریب همبستگی در محاسبات استفاده می‌شود.

حال برای هر مقدار جاافتاده مثل $m_i^j \in S_f$ اندیس مرتبط‌ترین صفت نسبت به صفت جاافتاده طبق (۵) از طریق ماتریس همبستگی افراز مربوطه پیدا می‌شود (خط ۸)

$$w^* = \arg\max_w (\rho_{j,w}) \quad , \quad w^* \neq j \qquad (۵)$$

اگر $\rho_{j,w^*} \ge t$ باشد که در آن t حد آستانه از پیش تعریف شده است، آن گاه یک رگرسور خطی با صفت مستقل[2] A_{w^*} و صفت وابسته[3] A_j به صورت (۶) ساخته می‌شود (خط ۱۰)

1. Pearson Correlation Coefficient
2. Independent Variable

Input:
DS: A dataset with N instances defined on schema $(A_1, A_2, \ldots, A_M)$ with missing values
c: number of clusters
p: comparison percentage
t: threshold for method selection
Output:
DS: Imputed dataset
Imputation-Method (c, p, t):

1. Split DS to $DS_C = \{c_1, c_2, \ldots, c_P\}$ (contains complete instances) and $DS_M = \{m_1, m_2, \ldots, m_Q\}$ (contains instances with missing value $(P+Q=N)$
2. Cluster DS_C to c clusters using k-means algorithm $(S_1, S_2, \ldots, S_c)$
3. Assign each instance in DS_C to a cluster
4. Assign each instance in DS_M to a cluster
5. $COR(S_1), COR(S_2), \ldots, COR(S_c) \leftarrow$ Compute correlation matrix for each cluster
6. **foreach** cluster S_f
7. **foreach** $m_i^j \in S_f$
8. $w^* \leftarrow$ find index of most correlated attribute to A_j using $COR(S_f)$ $\rho \leftarrow COR(S_f)$
9. **if** $\rho_{j,w^*} \ge t$
10. Create a linear regressor using A_j as dependent variable and A_{w^*} as independent variable
11. $estimatedValue \leftarrow$ estimate m_i^j using created regressor
end
12. **else**
13. $c_1, c_2, \ldots, c_k \leftarrow$ Select $p\%(k)$ of nearest neighbors to m_i^j in S_f
14. $estimatedValue \leftarrow$ estimate m_i^j using weighted k nearest neighbor algorithm
end
15. $m_i^j \leftarrow estimatedValue$
16. **return** $DS_C \cup DS_M$

شکل ۲: شبه‌کد رویکرد پیشنهادی.

$$A_j = a.A_{w^*} + b \qquad (۶)$$

که در آن a و b ضرایب رگرسور خطی هستند. این ضرایب از طریق برازش[4] بهترین خط بر روی داده‌ها به کمک روش حداقل مربعات[5] محاسبه می‌شوند. در این حالت مقدار جاافتاده از (۷) محاسبه می‌گردد (خط ۱۱)

$$m_i^j = a.m_i^{w^*} + b \qquad (۷)$$

اگر $\rho_{j,w^*} \le t$ باشد آن گاه مقدار جاافتاده با استفاده از روش k نزدیک‌ترین همسایه وزن‌دار محاسبه می‌گردد. در این حالت، فاصله نمونه جاافتاده با تمام نمونه‌های کامل موجود در خوشه S_f محاسبه می‌شود و این فواصل به ترتیب صعودی مرتب می‌گردند (یعنی $d(c_1, m_i) < d(c_2, m_i) < \ldots < d(c_{|S_f|}, m_i)$ که در آن $|S_f|$ اندازه

3. Dependent Variable
4. Fitting
5. Least Square Method

جدول ۲: مشخصات مجموعه داده‌های استفاده‌شده در آزمایش‌ها.

مجموعه داده	تعداد نمونه	تعداد ویژگی
Iris	۱۵۰	۴
Wine	۱۷۸	۱۳
Glass	۲۱۴	۱۰
Haberman	۳۰۶	۳
Wholesale Customers	۴۴۰	۸

خوشه S_f است). p درصد نمونه‌های کامل در خوشه نمونه جاافتاده به عنوان k نزدیک‌ترین نمونه کامل به نمونه جاافتاده انتخاب می‌شوند ($\{c_1, c_2, \ldots, c_k\} \in S_f$) و در نهایت مقدار جاافتاده از (۸) محاسبه می‌گردد (خطوط ۱۳ و ۱۴)

$$m_i^j = \frac{D_1 \times c_1^j + D_2 \times c_2^j + \ldots + D_k \times c_k^j}{D_1 + D_2 + \ldots + D_k} \quad (۸)$$

که در آن D_e از (۹) به دست می‌آید

$$D_e = \frac{1}{d(c_e, m_i)} \quad , \quad e = 1, 2, \ldots, k \quad (۹)$$

در واقع (۸) رابطه میانگین وزن‌دار است که وزن‌ها در آن معکوس فاصله دو نمونه هستند. در نهایت مقدار محاسبه‌شده m_i^j به جای مقدار جاافتاده جایگزین می‌شود و پس از پرکردن تمام مقادیر جاافتاده مجموعه داده کامل بدون مقادیر جاافتاده به دست می‌آید (خط ۱۶).

۳-۲ تحلیل پیچیدگی زمانی

عملیات پیش‌پردازش تنها یک بار و پیش از عملیات اصلی استخراج دانش از داده‌ها انجام می‌شود و بنابراین پیچیدگی زمانی عملیات پیش‌پردازش در اولویت پایین‌تری نسبت به کیفیت این امر قرار می‌گیرد. با این حال برای توصیف دقیق‌تر، پیچیدگی زمانی رویکرد پیشنهادی در این بخش تحلیل می‌شود.

فرض کنید مجموعه داده ورودی شامل N نمونه است. از این N نمونه P تای آنها کامل و Q تای آنها شامل مقادیر جاافتاده است ($P+Q=N$). همچنین فرض کنید هر نمونه شامل M ویژگی است. روش مطرح‌شده شامل سه گام اصلی است. در گام اول یک خوشه‌بندی k-means فقط بر روی داده‌های کامل انجام می‌شود (خطوط ۲ تا ۴). مرتبه زمانی این خوشه‌بندی برابر با $O(c.I.P.M)$ است که در آن c تعداد خوشه‌ها و I برابر با حداکثر تعداد تکرار در الگوریتم k-means است. در مرحله دوم به تعداد خوشه‌ها، ماتریس همبستگی محاسبه می‌شود (خط ۵) و هر ماتریس شامل M^2 عنصر است. مرتبه زمانی محاسبه هر عنصر در بدترین حالت[1] برابر با $O(P.\log P)$ است. از طرفی به تعداد خوشه‌ها یعنی c بار، نیاز به محاسبه ماتریس همبستگی است. بنابراین هزینه کلی این گام در بدترین حالت از مرتبه $O(c.M^2.P\log P)$ خواهد بود. آخرین گام، تخمین مقادیر جاافتاده است (خطوط ۶ تا ۱۵). در این مرحله به تعداد مقادیر جاافتاده (یعنی Q) یکی از روش‌های رگرسیون خطی با پیچیدگی $O(M^2.P)$ و یا k نزدیک‌ترین همسایگان با پیچیدگی $O(M.P)$ انجام می‌شود. بنابراین پیچیدگی مرحله دوم در بدترین حالت (حالتی که تمام تخمین‌ها از طریق رگرسیون انجام شوند) برابر با $O(Q.M^2.P)$ خواهد بود. با توجه به این که در بدترین حالت $c = N = P + Q$ است، با در نظر گرفتن موارد گفته‌شده، پیچیدگی الگوریتم پیشنهادی در بدترین حالت از مرتبه $O(c.M^2.P\log P)$ خواهد بود. اگر فرض کنیم که در بدترین حالت $P \approx N$، آن گاه پیچیدگی رویکرد پیشنهادی برابر با $O(c.M^2.N\log N)$ است.

۴- آزمایش‌ها و نتایج

در این بخش بررسی‌های جامعی بر روی تأثیر مقادیر مختلف الگوریتم ارائه‌شده آمده و همچنین دقت رویکرد پیشنهادی با چهار روش دیگر مقایسه می‌شود. از پنج مجموعه داده موجود در پایگاه داده UCI [۲۶] برای آزمایش‌ها استفاده می‌شود. جدول ۲ مشخصات این مجموعه داده‌ها را نمایش می‌دهد که این مجموعه داده‌ها بدون مقادیر جاافتاده می‌باشند و پیاده‌سازی‌های لازم با استفاده از زبان برنامه‌نویسی پایتون انجام شده است.

۴-۱ مشخصات آزمایش‌ها

در این زیربخش جزئیات آزمایش‌های انجام‌شده بیان می‌شود. برای بررسی تأثیر متغیرهای رویکرد پیشنهادی، ابتدا به تعداد ۱۰ درصد کل نمونه‌های هر مجموعه داده، مقادیر جاافتاده به طور تصادفی در آن مجموعه داده ایجاد و سپس روش ارائه‌شده بر روی آن مجموعه داده اعمال می‌شود.

از معیار فاصله اقلیدسی برای سنجش فاصله دو نمونه استفاده شده است. مقدار این فاصله بین دو نمونه x_i و x_j از (۱۰) محاسبه می‌شود

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{M} (x_i^k - x_j^k)^2} \quad (۱۰)$$

که در آن M تعداد صفات هر نمونه می‌باشد.

پیش‌پردازش نرمال‌سازی داده‌ها به بازه ۰ تا ۱ به روش Min-Max با استفاده از (۱۱) برای هر مقدار در مجموعه داده صورت گرفته است

$$x_i^j = \frac{x_i^j - x^{j,\min}}{x^{j,\max} - x^{j,\min}} \quad (۱۱)$$

که در آن $x^{j,\min}$ و $x^{j,\max}$ به ترتیب مقادیر کمینه و بیشینه برای صفت j ام هستند. نرمال‌سازی از غلبه صفاتی با بازه مقادیر کوچک بر صفاتی با بازه مقادیر بزرگ در محاسبه فاصله جلوگیری می‌کند.

هر یک از آزمایش‌های انجام‌شده در ادامه، ۵ بار به صورت تصادفی انجام گردیده و میانگین خطای این ۵ بار به عنوان نتیجه نهایی، گزارش شده است. برای محاسبه خطا از معیار میانگین مربعات خطا[2] مطابق (۱۲) استفاده می‌شود

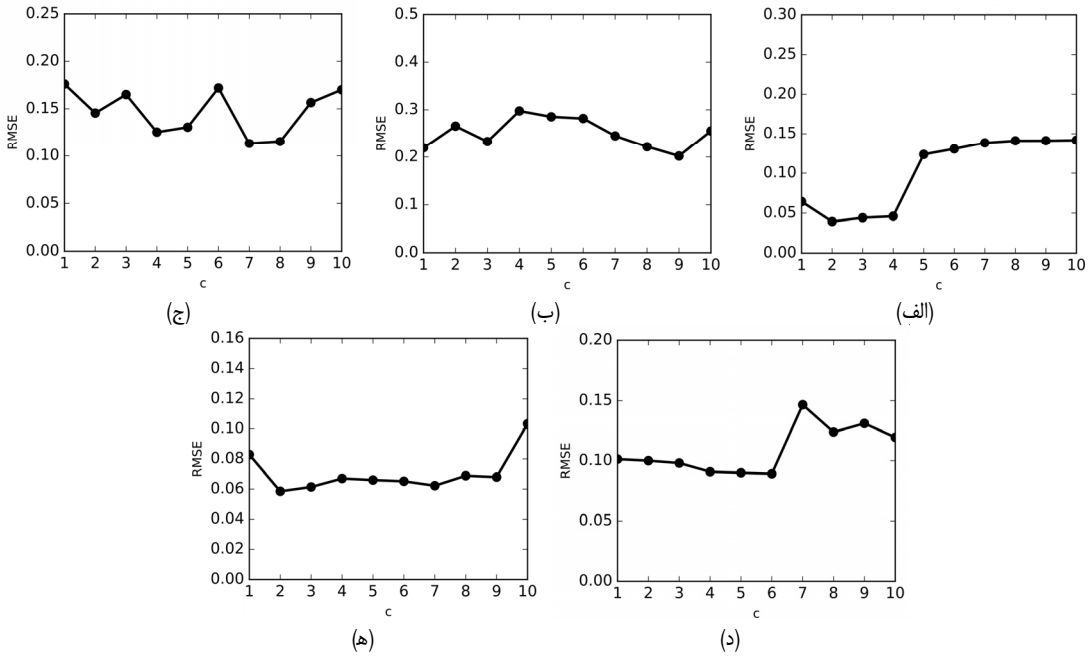
$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n} (y_k - y'_k)^2} \quad (۱۲)$$

که در آن y_k مقدار واقعی یک مقدار جاافتاده، y'_k مقدار تخمین زده شده برای آن مقدار جاافتاده و n تعداد مقادیر جاافتاده است.

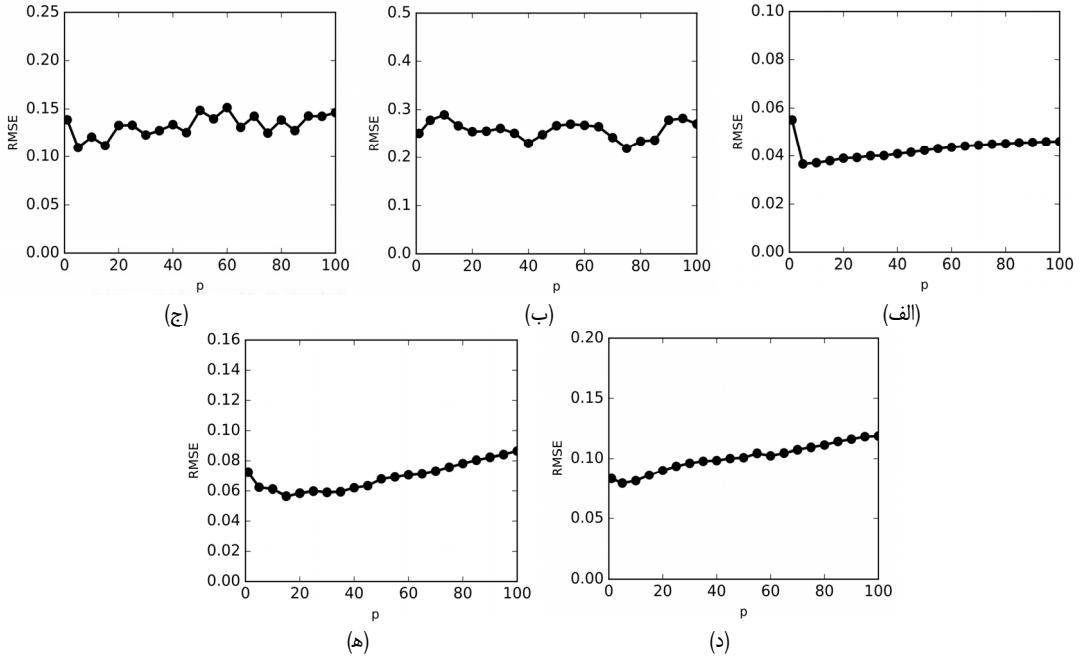
از خوشه‌بندی k-means برای خوشه‌بندی داده‌ها استفاده شده است. اگرچه این روش یکی از قدیمی‌ترین روش‌های خوشه‌بندی است با این حال هنوز هم یکی از پرکاربردترین روش‌های خوشه‌بندی است [۲۷]. حداکثر تعداد ۱۰۰ تکرار برای روش خوشه‌بندی k-means تنظیم شده است. در ادامه تأثیر متغیرهای مختلف رویکرد ارائه‌شده یعنی تعداد

1. Worst Case

2. Root Mean Squared Error



شکل ۳: تأثیر تعداد خوشه‌ها (c) بر روی مقدار RMSE در مجموعه داده‌های (الف) Wholesale Customers، (ب) Haberman، (ج) Wine، (د) Glass و (ه) Iris.



شکل ۴: تأثیر درصد مقایسات (p) بر روی مقدار RMSE در مجموعه داده‌های (الف) Wholesale Customers، (ب) Haberman، (ج) Wine، (د) Glass و (ه) Iris.

خوشه‌ها (c)، درصد انتخاب نزدیک‌ترین همسایه‌ها (p) و حد آستانه برای تعیین روش تخمین (t) بررسی می‌گردند.

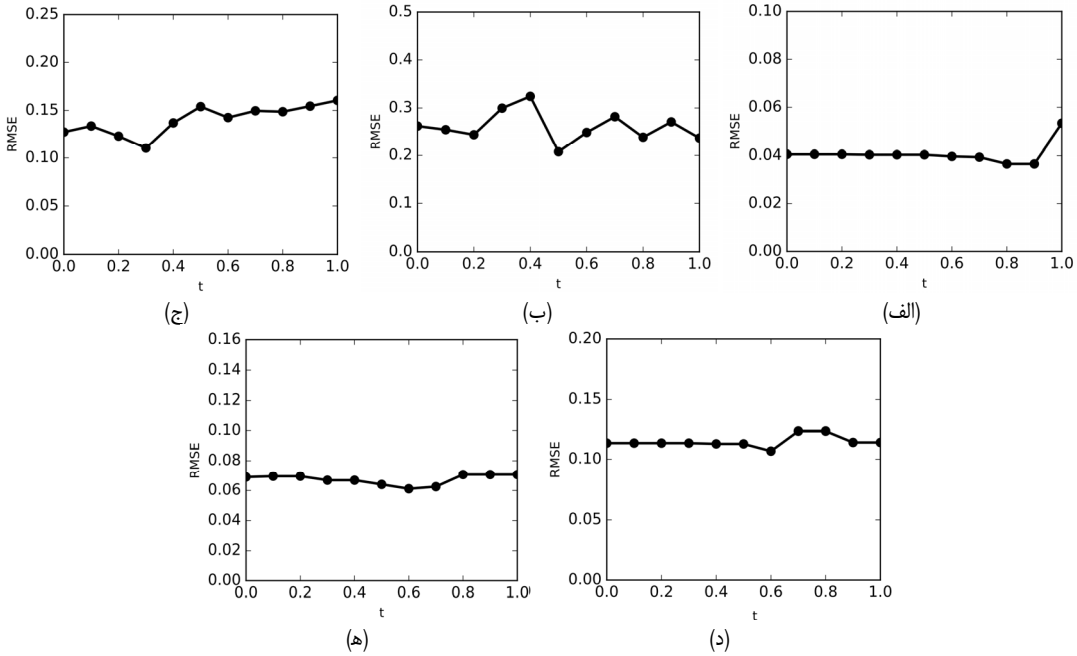
٤-٢ بررسی تأثیر تعداد خوشه‌ها (c) بر کیفیت تخمین داده‌های جاافتاده

در این قسمت نتایج آزمایش‌هایی برای بررسی تأثیر مقدار c در میزان خطا بررسی می‌شود. ابتدا به تعداد ۱۰٪ کل نمونه‌ها، مقدار جاافتاده به صورت تصادفی در هر مجموعه داده ایجاد شده است. سپس مقدار c بین ۱ تا ۱۰ برای مقادیر ثابت p و t، تغییر می‌یابند و نتایج ثبت می‌شوند. شکل ۳ مقدار خطای RMSE را برای مقادیر مختلف c برای مجموعه داده‌های مختلف نمایش می‌دهد. همان طور که از نمودارها پیداست برای مجموعه داده‌های مختلف خوشه‌بندی داده‌ها مقدار خطای تخمین را کاهش داده و در واقع مقدار خطا در یک مقدار $c \neq ۱$ بهینه شده است.

بنابراین فرضیه کاهش خطای تخمین با خوشه‌بندی داده‌ها با توجه به نمودارها قابل درک است. مقدار c بهینه برای مجموعه داده‌های مختلف، متفاوت است. آزمایش‌ها نشان می‌دهند افزایش مقدار c از یک حد به بعد، به تدریج باعث افزایش مقدار خطای تخمین می‌شود.

٤-٣ بررسی تأثیر درصد مقایسات (p) برای پرکردن مقدار جاافتاده در هر خوشه

در این قسمت نتایج آزمایش‌هایی برای بررسی تأثیر مقدار p در میزان خطا بررسی می‌شود. ابتدا به تعداد ۱۰٪ کل نمونه‌ها، مقدار جاافتاده به صورت تصادفی در هر مجموعه داده ایجاد شده است. سپس مقدار p بین ۱ تا ۱۰۰ درصد با گام ۵ درصد، برای مقادیر ثابت c و t، تغییر می‌یابند و نتایج ثبت می‌شوند. شکل ۴ مقدار خطای RMSE را برای مقادیر مختلف p برای مجموعه داده‌های مختلف نمایش می‌دهد. همان



شکل ۵: تأثیر حد آستانه (t) بر روی مقدار RMSE در مجموعه داده‌های، (الف) Wholesale Customers، (ب) Haberman، (ج) Wine، (د) Glass و (ه) Iris.

طور که از نمودارها پیداست برای مجموعه داده‌های Iris، Glass و Wholesale Customers پس از کاهش مقدار خطا در مقادیر کوچک p، با افزایش مقدار p، مقدار خطا نیز یک سیر تقریباً صعودی یکنوا داشته است. برای مجموعه داده Wine نیز اگرچه نوسان برای مقدار خطا وجود داشته است با این حال پس از یک مقدار بهینه ($p = 5\%$)، مقدار خطا یک سیر به طور کلی صعودی را داشته است. برای مجموعه داده Haberman مقدار بهینه p در مقادیر بالاتر حاصل شده است. در واقع نمودارها این مفهوم را بیان می‌کنند که مقدار بهینه k در روش k نزدیک‌ترین همسایه نه بسیار کوچک است و نه بسیار نزدیک به تعداد کل نمونه‌های موجود در هر خوشه. مقدار بهینه p برای هر مجموعه داده را می‌توان طی آزمایش‌ها تعیین نمود.

٤-٤ بررسی تأثیر حد آستانه (t) برای تعیین روش تخمین

در این قسمت نتایج آزمایش‌هایی برای بررسی تأثیر مقدار t در میزان خطا بررسی می‌شود. ابتدا به تعداد ۱۰٪ کل نمونه‌ها، مقدار جاافتاده به صورت تصادفی در هر مجموعه داده ایجاد شده است. سپس مقدار t بین ۰ تا ۱ با گام ۰/۱ برای مقادیر ثابت p و c، تغییر می‌یابند و نتایج ثبت می‌شوند. شکل ۵ مقدار خطای RMSE را برای مقادیر مختلف t برای مجموعه داده‌های مختلف نمایش می‌دهد. همان طور که از نمودارها پیداست برای مجموعه داده‌های مختلف یک مقدار بهینه $0 < t < 1$ توانسته است مقدار خطا را کمینه کند. این نمودارها بیانگر این هستند که ترکیب دو روش k نزدیک‌ترین همسایگان وزن‌دار و رگرسیون خطی توانسته است بهتر از این دو روش به تنهایی عمل کند. همان طور که پیداست در $t = 0$ که بیانگر روش k نزدیک‌ترین همسایگان وزن‌دار به تنهایی و در $t = 1$ که بیانگر روش رگرسیون خطی به تنهایی است، مقادیر خطا بیش از مقدار کمینه برای خطاست. مختلف‌بودن مقادیر بهینه برای t به این معناست که برای مجوعه داده‌های مختلف یکی از روش‌ها بر دیگری غلبه دارد. در واقع میزان همبستگی درون نمونه‌های افرازهای تشکیل‌شده از مجموعه داده‌های مختلف، متفاوت است.

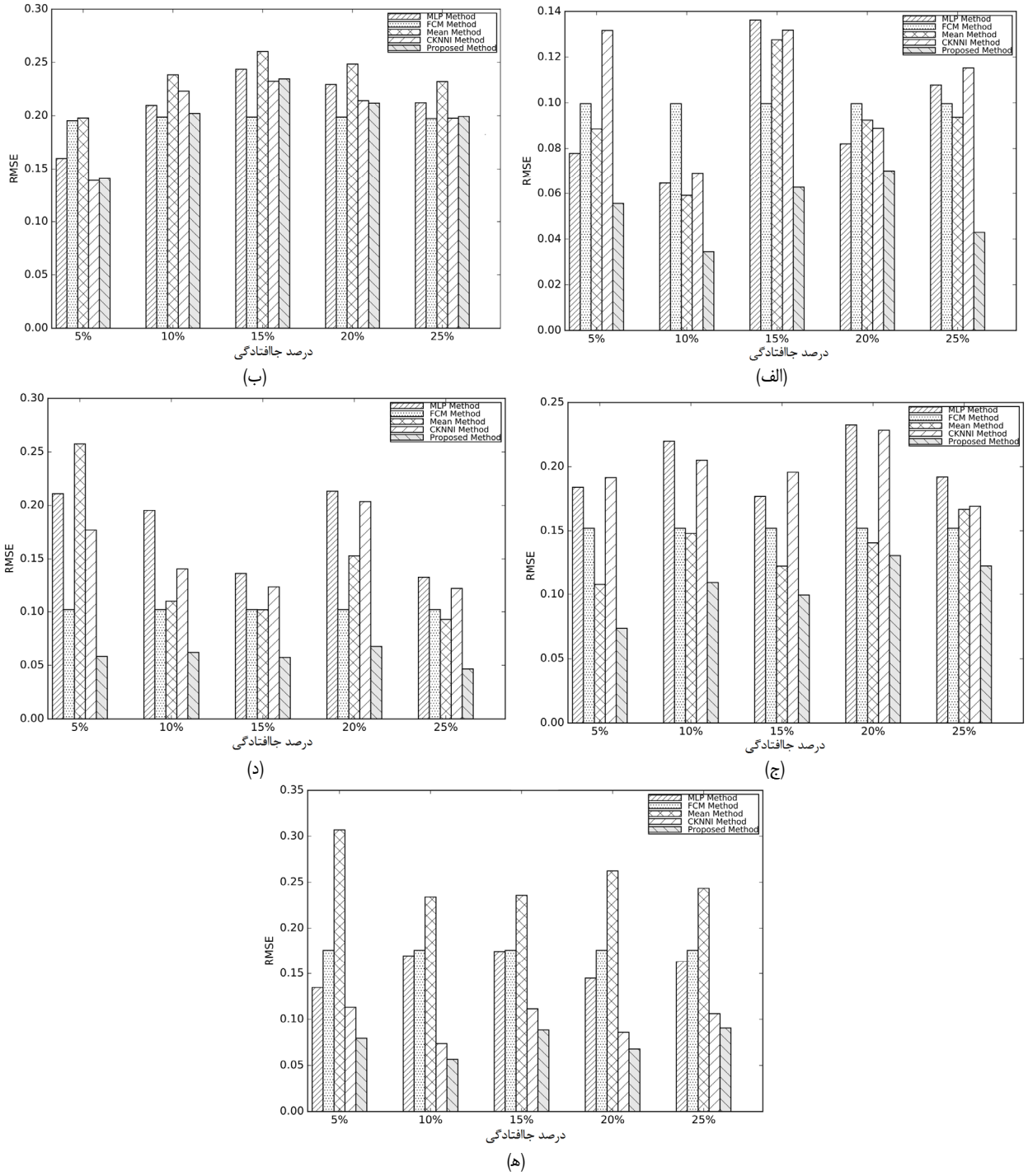
٤-٥ مقایسه با روش پرکردن با دیگر روش‌ها

در این قسمت میزان خطای تخمین داده‌ها در مقایسه با روش پرکردن با مقدار میانگین (Mean)، روش تخمین با شبکه عصبی پرسپترون چندلایه (MLP) [۵]، روش تخمین با خوشه‌بندی c-means فازی (FCM)[1] [۶] و همچنین روش CKNNI [۷] مقایسه می‌شود. ابتدا به تعداد ۵٪، ۱۰٪، ۱۵٪، ۲۰٪ و ۲۵٪ کل نمونه‌ها، مقدار جاافتاده به صورت تصادفی در هر مجموعه داده ایجاد شده است. سپس مقادیر جاافتاده با چهار روش ذکرشده و روش پیشنهادی در این مقاله، پر شده‌اند. مقدار خطای RMSE برای این رویکردها محاسبه شده و نتایج برای مجموعه داده‌های مختلف ثبت شدند. شکل ۶ مقدار خطای RMSE محاسبه‌شده برای مجموعه داده‌های مختلف برای درصدهای مختلف مقادیر جاافتاده تولیدشده را برای روش‌های ذکرشده نشان می‌دهد. همان طور که از شکل‌ها پیداست برای همه مجموعه داده‌ها به غیر از Haberman رویکرد پیشنهادی خطای تخمین کمتری نسبت به چهار روش دیگر داشته است. اگرچه در مجموعه داده Haberman دقت تخمین روش پیشنهادی نسبت به دیگرها کمتر است با این حال روش پیشنهادی عملکرد بسیار نزدیکی نسبت به بهترین روش در ۵٪، ۲۰٪ و ۲۵٪ جاافتادگی داشته است. علت ضعف نسبی روش پیشنهادی در این مجموعه داده احتمالاً تعداد کم صفات پیش‌بینی کننده در این مجموعه داده است. به طور کلی با توجه به نتایج موجود می‌توان مشاهده کرد که خطای روش پیشنهادی در اکثر موارد (به غیر از مجموعه داده Haberman) از دیگر روش‌های مقایسه‌شده کمتر است.

۵- نتیجه‌گیری

در این مقاله روشی جدید و دومرحله‌ای برای تخمین داده‌های جاافتاده ارائه گردید. روش پیشنهادی شامل یک مرحله خوشه‌بندی و ترکیب دو روش پایه برای تخمین مقادیر جاافتاده است. تأکید روش پیشنهادی بر استفاده از همبستگی‌های موجود بر مجموعه داده‌ها تا حد ممکن برای

1. Fuzzy C-Means



شکل ۶: مقایسه مقدار RMSE برای رویکرد ارائه‌شده، روش CKNNI و روش میانگین، برای درصدهای مختلف جاافتادگی در مجموعه داده‌های (الف)Wholesale Customers، (ب) Haberman، (ج) Wine، (د) Glass و (ه) Iris.

تخمین مقادیر جاافتاده است. نتایج بررسی‌شده بر روی پنج مجموعه داده نشان داد که رویکرد پیشنهادی میانگین مربعات خطای کمتری در اکثر موارد در مقایسه با چهار روش مختلف دیگر دارد. تأثیر پارامترهای رویکرد پیشنهادی در میزان خطا نیز بررسی شدند.

از مشکلات اصلی رویکرد پیشنهادی می‌توان به چالش انتخاب بهینه برای پارامترهای رویکرد پیشنهادی اشاره کرد اگرچه این چالش نیز با عملیات زمان‌بر جستجوی کامل برای انتخاب بهترین ترکیب پارامترها قابل برطرف‌شدن است. تأثیر دیگر روش‌های خوشه‌بندی در مرحله اول و ترکیب دیگر روش‌ها در دست بررسی است. به علاوه، تأثیر دیگر معیارهای سنجش همبستگی مانند معیار اطلاعات متقابل[1] در کیفیت پرکردن مقادیر جاافتاده در پژوهش‌های آینده بررسی خواهد شد.

مراجع

[1] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Second Edition, John Wiley Sons, Inc., pp. 11-15, 2002.

[2] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," *Information Sciences*, vol. 233, no. 2, pp. 25-35, Jun. 2013.

1. Mutual Information

[3] B. van Stein and W. Kowalczyk, "An incremental algorithm for repairing training sets with missing values," in *Proc. Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems,* vol. 611, pp. 175-186, Jun. 2016.
[4] A. A. Chavan and V. K. Verma, "Treatment of missing values for association rules: a recent survey," *International J. of Computer Applications*, vol. 70, no. 26, pp. 1-4, May 2013.
[5] E. L. Silva-Ramrez, R. Pino-Mejas, and M. Lopez-Coello, "Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns," *Applied Soft Computing*, vol. 29, no. 1, pp. 65-74, Apr. 2015.
[6] P. Raja and K. Thangavel, "Soft clustering based missing value imputation," in *Proc. Annual Convention of the Computer Society of India: Digital Connectivity-Social Impact,* vol. 679, pp. 119-133, Dec. 2016.
[7] C. Jiang and Z. Yang, "CKNNI: an improved knn-based missing value handling technique," in *Proc. Int. Conf. on Intelligent Computing,* pp. 441-452, Aug. 2015.
[8] C. H. Wu, C. H. Wun, and H. J. Chou, "Using association rules for completing missing data," in *Proc. IEEE Fourth Int. Conf. on Hybrid Intelligent Systems, HIS'04,* pp. 236-241, 5-8 Dec. 2004.
[9] J. Wu, Q. Song, and J. Shen, "An novel association rule mining based missing nominal data imputation method," in *Proc. IEEE Eighth ACIS Int. Conf. on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, SNPD, vol. 3, pp. 244-249, Jul. 2007.
[10] N. Singh, A. Javeed, S. Chhabra, and P. Kumar, "Missing value imputation with unsupervised Kohonen self organizing map," in *Emerging Research in Computing, Information, Communication and Applications,* pp. 61-76, Jul. 2015.
[11] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 3rd Ed., pp. 398-408, 2011.
[12] R. Krishnamoorthy, S. Sreedhar Kumar, and B. Neelagund, "A new approach for data cleaning process," in *Proc. IEEE Recent Advances and Innovations in Engineering, ICRAIE'14,* 5 pp. Jul. 2014.
[13] C. Zhang, X. Zhu, J. Zhang, Y. Qin, and S. Zhang, "GBKII: an imputation method for missing values," in *Advances in Knowledge Discovery and Data Mining,* vol. 11, pp. 1080-1087, May 2007.
[14] E. R. Hruschka, E. R. Hruschka, and N. F. F. Ebecken, "Evaluating a nearest-neighbor method to substitute continuous missing values," in *Proc. Australasian Joint Conf. on Artificial Intelligence,* vol. 16, pp. 723-734, Dec. 2003.
[15] C. Ye and H. Wang, "Capture missing values based on crowdsourcing," in *Proc. of the 9th Int. Conf. on Wireless Algorithms, Systems, and Applications, WASA'14,* pp. 783-792, Jun. 2014.
[16] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Missing value imputation based on k-mean clustering with weighted distance," in *Proc. Int. Conf. on Contemporary Computing,* vol. 3, pp. 600-609, Aug. 2010.
[17] V. V. Ayuyev, J. Jupin, P. W. Harris, and Z. Obradovic, "Dynamic clustering-based estimation of missing values in mixed type data," in *Proc. Int. Conf. on Data Warehousing and Knowledge Discovery,* vol. 11, pp. 366-377, Aug. 2009.
[18] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation: a study of fuzzy k-means clustering method," in *Rough Sets and Current Trends in Computing,* vol. 3066, pp. 573-579, Jun. 2004.
[19] N. Ankaiah and V. Ravi, "A novel soft computing hybrid for data imputation," in *Proc. of the 7th Int. Conf. on Data Mining, DMIN'11*, pp. 65-71, Jul. 2011.
[20] S. Azim and S. Aggarwal, "Hybrid model for data imputation: using fuzzy c means and multi layer perceptron," in *Proc. IEEE Int. Advance Computing Conf., IACC'14*, vol. 4, pp. 1281-1285, Feb. 2014.
[21] S. Bashir, S. Razzaq, U. Maqbool, S. Tahir, and A. R. Baig, *Using Association Rules for Better Treatment of Missing Values*, arXiv preprint arXiv: 0904.3320, 2009.
[22] G. Rahman and Z. Islam, "A decision tree-based missing value imputation technique for data pre-processing," in *Proc. of the Ninth Australasian Data Mining Conf.*, *AusDM'11*, vol. 121, pp. 41-50, Dec. 2011.
[23] C. F. Tsai and F. Y. Chang, "Combining instance selection for better missing value imputation," *J. of Systems and Software*, vol. 122, no. 1, pp. 63-71, Dec. 2016.
[24] D. R. Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 257-286, Mar. 2000.
[25] M. Amiri and R. Jensen, "Missing data imputation using fuzzy-rough methods," *Neurocomputing*, vol. 205, no. 1, pp. 152-164, Sept. 2016.
[26] M. Lichman, *UCI Machine Learning Repository* School of Information and Computer Science, Irvine, CA: University of California, 2013.
[27] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recogn. Lett.*, vol. 31, no. 8, pp. 651-666, Jun. 2010.

امیرمسعود سفیدیان فارغ التحصیل مقطع کارشناسی ارشد در رشته‌ی مهندسی کامپیوتر- نرم افزار از دانشگاه تربیت دبیر شهید رجایی می‌باشد. نام‌برده تحصیلات خود را در مقطع کارشناسی مهندسی کامپیوتر- نرم‌افزار در سال ۱۳۹۴ با کسب رتبه اول در دانشگاه تربیت دبیر شهید رجایی به اتمام رسانده است. وی در سال ۱۳۹۴ بدون کنکور و با شرایط استعداد درخشان به‌طور مستقیم در مقطع کارشناسی ارشد دانشگاه تربیت دبیر شهید رجایی در رشته‌ی مهندسی کامپیوتر- نرم‌افزار پذیرفته شد. وی در تابستان ۱۳۹۶ تحصیلات خود را در مقطع کارشناسی ارشد با کسب رتبه‌ی اول به پایان رساند. زمینه‌های تحقیقاتی ایشان عبارتند از: جایگذاری مقادیر جاافتاده در داده‌ها، پیش‌پردازش داده‌ها و داده‌کاوی.

نگین دانشپور استادیار دانشکده مهندسی کامپیوتر دانشگاه تربیت دبیر شهید رجایی می‌باشد. نام‌برده تحصیلات خود را در مقطع کارشناسی مهندسی کامپیوتر- سخت‌افزار در سال ۱۳۷۸ با کسب رتبه اول در دانشگاه شهید بهشتی، و کارشناسی ارشد مهندسی کامپیوتر- نرم‌افزار در سال ۱۳۸۱ در دانشگاه صنعتی امیرکبیر به پایان رسانده است، و در سال ۱۳۸۹ دکتری خود در رشته مهندسی کامپیوتر- نرم‌افزار را از دانشگاه صنعتی امیرکبیر اخذ کرده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پایگاه داده تحلیلی، سیستم‌های تصمیم‌یار، پیش‌پردازش داده‌ها و داده‌کاوی.