

**باسمه تعالی**



**تعهدنامه اصالت اثر**

اینجانب امیرمسعود سفیدیان متعهد می­شوم که مطالب مندرج در این پایان­نامه/رساله حاصل کار پژوهشی اینجانب است و دستاوردهای پژوهشی دیگران که در این پژوهش استفاده­شده است، مطابق مقررات، ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان­نامه/رساله قبلاً برای احراز هیچ مدرک هم­سطح یا بالاتر ارائه نشده است. در صورت اثبات تخلف (در هر زمان) مدرک تحصیلی صادر شده توسط دانشگاه از اعتبار ساقط خواهدشد.

کلیه حقوق مادی و معنوی این اثر متعلق به دانشگاه تربیت دبیر شهید رجایی است.

امیرمسعود سفیدیان

امضاء

تهران- لویزان –کد پستی16788 – صندوق پستی163- 16785 تلفن9- 22970060 پست الکترونیکی [sru@sru.ac.ir](mailto:sru@sru.ac.ir)



دانشكده مهندسی کامپیوتر

بهبود تخمین مقادیر جاافتاده و تشخیص ناسازگاری در داده‌ها با استفاده از افراز داده‌ها

**نگارش**

**امیرمسعود سفیدیان**

**استاد راهنما: دکتر نگین دانشپور**

**پایان­نامه جهت دریافت درجه کارشناسی ارشد**

**در رشته مهندسی کامپیوتر-گرایش نرم افزار**

**شهریور ماه 1396**

**تقدیم به مهربان فرشتگانم**

این پایان‌نامه را ضمن تشکر و سپاس بی‌کران و در کمال افتخار تقدیم می‌نمایم به:

محضر ارزشمند پدر و مادر عزيزمان به‌خاطر همه‌ی تلاشهای محبت آمیزی که در دوران مختلف زندگی‌مان انجام داده‌اند و با مهربانی چگونه زیستن را به ما آموخته‌اند.

به آنان که نفس خیرشان و دعای روح پرورشان بدرقه‌ی راهمان بود.

الها به ما کمک کن تا بتوانیم ادای دين کنیم و به خواسته‌ی آنان جامه‌ی عمل بپوشانیم.

پروردگارا حسن عاقبت، سلامت و سعادت را برای آنان مقدر نما.

**قدردانی و تشکر**

اکنون با به پایان رساندن دوره­ی کارشناسی ارشد، وظیفه­ی خود می­دانم از تمام کسانی که مرا در طی این دوران یاری نمودند تشکر نمایم. در آغاز از استاد ارجمندم، سرکار خانم **دکتر نگین دانشپور** که زحمت راهنمایی من در طول این دوره بر عهده­ی ایشان بود بی­نهایت سپاسگزارم.

چكيده

یکی از مهم­ترین گام­ها در فرآیند کشف دانش، پیش­پردازش داده‌ها است. عملیات پاک­سازی داده‌ها در پیش­پردازش داده‌ها نقشی حیاتی در تضمین داده‌هایی باکیفیت بالا دارد. از جنبه­های ضروری داده‌هایی باکیفیت بالا، جنبه­ی کامل بودن داده­ها ­است. داده­های جاافتاده باید پیش از استفاده در الگوریتم­های داده‌کاوی و یادگیری ماشین جایگذاری شوند. در این پژوهش رویکردهای مختلف جدیدی برای جایگذاری مقادیر جاافتاده مطرح می­شود. مهم­ترین نکته در رویکردهای ارائه‌شده این است که توجه به افراز­هایی از داده‌های ورودی، به‌جای توجه به‌کل داده‌ها، احتمالا می‌تواند کیفیت تخمین داده‌ها را بهبود بخشد. سه دسته رویکرد مختلف در این پژوهش برای هدف جایگذاری مقادیر جاافتاده مطرح می‌شود. در دسته‌ی اول، روشی دومرحله‌ای مبتنی بر خوشه‌بندی داده‌ها و استفاده از ‌ترکیب روش k نزدیک‌ترین همسایگان وزن‌دار و رگرسیون خطی برای پرکردن مقادیر جاافتاده ارائه می‌شود. دسته‌ی دوم شامل ده رویکرد مبتنی بر همبستگی برای هدف جایگذاری مقادیر جاافتاده است. در دسته‌ی سوم، مقادیر جاافتاده با استفاده از خوشه‌بندی فازی c-means مبتنی بر گِرِی و انتخاب ویژگی مبتنی بر معیار اطلاعات متقابل، جایگذاری می‌شوند.

از سه معیار شناخته­شده‌ی ریشه میانگین مربعات خطا، ضریب تعیین و میانگین مطلق خطا برای ارزیابی عملکرد رویکردهای جایگذاری پیشنهادی استفاده‌شده است. نتایج رویکردهای پیشنهادی با دیگر رویکردهای موجود بر روی مجموعه­داده‌های شناخته­شده در زمینه­ی داده­کاوی و یادگیری ماشین مقایسه شده است. نتایج آزمایش­های تجربی نشان می­دهد که رویکردهای پیشنهادی در بیش‌تر موارد توانسته­اند بهتر از دیگر رویکردهای مقایسه شده عمل­ کنند.

**واژگان کلیدی**: پاک­سازی داده­ها، کیفیت داده­ها، تخمین مقادیر جاافتاده، افرازبندی

فهرست مطالب

[1 فصل اول: تعاریف پایه 1](#_Toc490694965)

[1-1 مقدمه 2](#_Toc490694966)

[1-2 داده‌کاوی 3](#_Toc490694967)

[1-3 کیفیت داده و اهمیت آن 5](#_Toc490694968)

[1-4 مشکلات کیفیت داده‌ها 7](#_Toc490694969)

[1-4-1 مشکلات تک منبع 8](#_Toc490694970)

[1-4-2 مشکلات چند منبع 9](#_Toc490694971)

[1-5 پاک‌سازی داده 10](#_Toc490694972)

[1-6 مقادیر جاافتاده 10](#_Toc490694973)

[1-6-1 علل ایجاد مقادیر جاافتاده 11](#_Toc490694974)

[1-6-2 مشکلات وجود مقادیر جاافتاده 12](#_Toc490694975)

[1-7 خلاصه و نتیجه‌گیری 13](#_Toc490694976)

[2 فصل دوم: پیشینه تحقیق 14](#_Toc490694977)

[2-1 مقدمه 15](#_Toc490694978)

[2-2 مکانیزم‌های ایجاد مقادیر جاافتاده 15](#_Toc490694979)

[2-3 رویکردهای جایگذاری مقادیر جاافتاده 16](#_Toc490694980)

[2-3-1 رویکردهای مبتنی بر شبکه‌های عصبی 18](#_Toc490694981)

[2-3-2 رویکردهای مبتنی بر روش k نزدیک‌ترین همسایگان 18](#_Toc490694982)

[2-3-3 رویکردهای مبتنی بر الگوریتم حداکثرسازی امید 21](#_Toc490694983)

[2-3-4 رویکردهای مبتنی رگرسیون 22](#_Toc490694984)

[2-3-5 رویکردهای مبتنی قوانین انجمنی 23](#_Toc490694985)

[2-3-6 رویکردهای مبتنی بر خوشه‌بندی 25](#_Toc490694986)

[2-3-7 رویکردهای ترکیبی 27](#_Toc490694987)

[2-3-8 دیگر روش‌ها 28](#_Toc490694988)

[2-4 ایده‌‌ی کلی پیشنهادی 29](#_Toc490694989)

[2-5 معیارهای ارزیابی برای سنجش کیفیت جایگذاری مقادیر جاافتاده 31](#_Toc490694990)

[2-6 خلاصه و نتیجه‌گیری 32](#_Toc490694991)

[3 فصل سوم: استفاده از خوشه‌بندی و رویکردی ترکیبی برای جایگذاری مقادیر جاافتاده عددی 35](#_Toc490694992)

[3-1 مقدمه 36](#_Toc490694993)

[3-2 رویکرد پیشنهادی 37](#_Toc490694994)

[3-3 تحلیل پیچیدگی زمانی رویکرد پیشنهادی 43](#_Toc490694995)

[3-4 آزمایش‌ها و نتایج 44](#_Toc490694996)

[3-4-1 مشخصات آزمایش‌ها 45](#_Toc490694997)

[3-4-2 بررسی تأثیر تعداد خوشه‌ها (c) بر کیفیت تخمین داده‌های جاافتاده 46](#_Toc490694998)

[3-4-3 بررسی تأثیر درصد مقایسات (p) برای پرکردن مقدار جاافتاده در هر خوشه 46](#_Toc490694999)

[3-4-4 بررسی تأثیر حد آستانه (t) برای تعیین روش تخمین 47](#_Toc490695000)

[3-4-5 مقایسه با روش پرکردن با مقدار میانگین و CKNNI 49](#_Toc490695001)

[3-5 خلاصه و نتیجه‌گیری 50](#_Toc490695002)

[4 فصل چهارم: جایگذاری مقادیر جاافتاده با استفاده از روشهای مبتنی بر همبستگی 54](#_Toc490695003)

[4-1 مقدمه 55](#_Toc490695004)

[4-2 جزئیات مراحل پیشنهادی 57](#_Toc490695005)

[4-2-1 محاسبه اولویت هر صفت جاافتاده برای هدف تخمین داده‌های جاافتاده 59](#_Toc490695006)

[4-2-2 انتخاب مجموعه پایه 60](#_Toc490695007)

[4-2-3 حداکثر کردن همبستگی بین صفات با بسط دادن مجموعه پایه 61](#_Toc490695008)

[4-2-3-1 حداکثر کردن همبستگی هر یک از صفات معلوم نسبت به صفت جاافتاده 63](#_Toc490695009)

[4-2-3-2 حداکثر کردن میانگین مقادیر همبستگی بین صفات معلوم و صفت جاافتاده 64](#_Toc490695010)

[4-2-4 تخمین مقادیر جاافتاده با استفاده از زیرمجموعه‌های یافته شده 65](#_Toc490695011)

[4-2-5 ترکیب گام‌های حداکثرسازی و تخمین 67](#_Toc490695012)

[4-3 آزمایش‌های تجربی 69](#_Toc490695013)

[4-3-1 مجموعه‌داده‌های مورد آزمایش 69](#_Toc490695014)

[4-3-2 پیش‌پردازش داده‌ها و نحوه ایجاد مقادیر جاافتاده 70](#_Toc490695015)

[4-3-3 معیارهای ارزیابی 71](#_Toc490695016)

[4-3-4 مقایسه عملکرد رویکردهای ارائه‌شده با دیگر روش‌ها 71](#_Toc490695017)

[4-3-5 تأثیر پارامترهای و 78](#_Toc490695018)

[4-4 خلاصه و نتیجه‌گیری 83](#_Toc490695019)

[5 فصل پنجم: جایگذاری مقادیر جاافتاده با استفاده از خوشه‌بندی فازی c-means مبتنی بر گِرِی و انتخاب ویژگی مبتنی بر معیار اطلاعات متقابل 89](#_Toc490695024)

[5-1 مقدمه 90](#_Toc490695025)

[5-2 مفاهیم پیش‌نیاز 91](#_Toc490695026)

[5-2-1 خوشه‌بندی c-means فازی 91](#_Toc490695027)

[5-2-2 تحلیل رابطه‌ای گری 94](#_Toc490695028)

[5-2-3 معیار اطلاعات متقابل 95](#_Toc490695029)

[5-3 جزئیات مراحل پیشنهادی 97](#_Toc490695030)

[5-3-1 محاسبه اولویت هر صفت جاافتاده برای فرآیند تخمین داده‌های جاافتاده 97](#_Toc490695031)

[5-3-2 جایگذاری اولیه، خوشه‌بندی داده‌ها با استفاده از فازی c-means مبتنی بر گِرِی و عملیات انتخاب خوشه 98](#_Toc490695032)

[5-3-3 انتخاب ویژگی در هر خوشه‌ی انتخاب‌شده 102](#_Toc490695033)

[5-3-4 اعمال مدل رگرسیون بر هریک از خوشه‌ها و صفات انتخاب‌شده 103](#_Toc490695034)

[5-3-5 انجام نظرخواهی وزن‌دار برای تخمین نهایی 104](#_Toc490695035)

[5-4 آزمایش‌های تجربی 104](#_Toc490695036)

[5-4-1 طراحی آزمایش‌ها 105](#_Toc490695037)

[5-4-1-1 مجموعه‌داده‌های مورد آزمایش 105](#_Toc490695038)

[5-4-1-2 پیش‌پردازش داده‌ها و نحوه ایجاد مقادیر جاافتاده 106](#_Toc490695039)

[5-4-1-3 معیارهای ارزیابی 106](#_Toc490695040)

[5-4-2 آزمایش‌های انجام‌شده 106](#_Toc490695041)

[5-4-2-1 بررسی عملکرد الگوریتم خوشه‌بندی فازی c-means مبتنی بر گِری 106](#_Toc490695042)

[5-4-2-2 مقایسه عملکرد رویکردهای ارائه‌شده با دیگر روش‌ها توسط معیار میانگین مربعات خطا 110](#_Toc490695043)

[5-4-2-3 مقایسه عملکرد رویکردهای ارائه شده با دیگر روش‌ها توسط معیار ضریب تعیین 112](#_Toc490695044)

[5-4-2-4 مقایسه عملکرد رویکردهای ارائه شده با دیگر روش‌ها توسط معیار میانگین مطلق خطا 114](#_Toc490695045)

[5-4-2-5 تأثیر پارامترهای c و 116](#_Toc490695046)

[5-5 خلاصه و نتیجه‌گیری 118](#_Toc490695047)

[6 نتیجه‌گیری، بحث و پیشنهاد‌ها 119](#_Toc490695048)

[6-1 مقدمه 120](#_Toc490695049)

[6-2 نتایج تحقیق 120](#_Toc490695050)

[6-3 پیشنهادهایی برای پژوهش‌های آینده 121](#_Toc490695051)

[7 منابع و مراجع 123](#_Toc490695052)

فهرست جداول

[جدول ‏3‑1 مشخصات مجموعه‌داده‌های استفاده‌شده 44](#_Toc490695456)

[جدول ‏4‑1 ترکیب‌های مختلف برای گام‌های حداکثرسازی و تخمین 68](#_Toc490695457)

[جدول ‏4‑2 مقایسه عملکرد رویکردهای ارائه‌شده با سه روش دیگر برای مجموعه‌داده‌ی Iris 73](#_Toc490695458)

[جدول ‏4‑3 مقایسه عملکرد رویکردهای ارائه‌شده با سه روش دیگر برای مجموعه‌داده‌ی Wine 74](#_Toc490695459)

[جدول ‏4‑4 مقایسه عملکرد رویکردهای ارائه‌شده با سه روش دیگر برای مجموعه‌داده‌ی Glass 75](#_Toc490695460)

[جدول ‏4‑5 مقایسه عملکرد رویکردهای ارائه‌شده با سه روش دیگر برای مجموعه‌داده‌ی Haberman 76](#_Toc490695461)

[جدول ‏4‑6 مقایسه عملکرد رویکردهای ارائه‌شده با سه روش دیگر برای مجموعه‌داده‌ی Wholesale Customers 77](#_Toc490695462)

[جدول ‏5‑1 مقایسه عملکرد رویکردهای جایگذاری توسط معیار RMSE در مجموعه‌داده Wine 111](#_Toc490695463)

[جدول ‏5‑2 مقایسه عملکرد رویکردهای جایگذاری توسط معیار RMSE در مجموعه‌داده Wholesale Customers 111](#_Toc490695464)

[جدول ‏5‑3 مقایسه عملکرد رویکردهای جایگذاری توسط معیار RMSE در مجموعه‌داده Haberman 111](#_Toc490695465)

[جدول ‏5‑4 مقایسه عملکرد رویکردهای جایگذاری توسط معیار RMSE در مجموعه‌داده Iris 111](#_Toc490695466)

[جدول ‏5‑5 مقایسه عملکرد رویکردهای جایگذاری توسط معیار RMSE در مجموعه‌داده Glass 112](#_Toc490695467)

[جدول ‏5‑6 مقایسه عملکرد رویکردهای جایگذاری توسط معیار CoD در مجموعه‌داده Wine 113](#_Toc490695468)

[جدول ‏5‑7 مقایسه عملکرد رویکردهای جایگذاری توسط معیار CoD در مجموعه‌داده Wholesale Customers 113](#_Toc490695469)

[جدول ‏5‑8 مقایسه عملکرد رویکردهای جایگذاری توسط معیار CoD در مجموعه‌داده Haberman 113](#_Toc490695470)

[جدول ‏5‑9 مقایسه عملکرد رویکردهای جایگذاری توسط معیار CoD در مجموعه‌داده Iris 113](#_Toc490695471)

[جدول ‏5‑10 مقایسه عملکرد رویکردهای جایگذاری توسط معیار CoD در مجموعه‌داده Glass 114](#_Toc490695472)

[جدول ‏5‑11 مقایسه عملکرد رویکردهای جایگذاری توسط معیار MAE در مجموعه‌داده Wine 115](#_Toc490695473)

[جدول ‏5‑12 مقایسه عملکرد رویکردهای جایگذاری توسط معیار MAE در مجموعه‌داده Wholesale Customers 115](#_Toc490695474)

[جدول ‏5‑13 مقایسه عملکرد رویکردهای جایگذاری توسط معیار MAE در مجموعه‌داده Haberman 115](#_Toc490695475)

[جدول ‏5‑14 مقایسه عملکرد رویکردهای جایگذاری توسط معیار MAE در مجموعه‌داده Iris 115](#_Toc490695476)

[جدول ‏5‑15 مقایسه عملکرد رویکردهای جایگذاری توسط معیار MAE در مجموعه‌داده Glass 116](#_Toc490695477)

فهرست شکل‌ها

[شکل ‏1‑1 مراحل مختلف فرآیند کشف دانش [1] 5](#_Toc490695709)

[شکل ‏1‑2 دسته‌بندی مشکلات کیفیت داده 7](#_Toc490695710)

[شکل ‏2‑1 یک مثال از این‌که وابستگی قوی بین دو صفت روی کل مجموعه‌داده برقرار نیست (جدول سمت چپ) ولی روی افراز‌هایی از آن برقرار است (جدول سمت راست) [18]. 30](#_Toc490695711)

[شکل ‏2‑2 دسته‌بندی کلی رویکردهای تخمین مقادیر جاافتاده 33](#_Toc490695712)

[شکل ‏3‑1 روند کلی رویکرد پیشنهادی برای پرکردن مقادیر جاافتاده با استفاده از خوشه‌بندی و رویکردی ترکیبی 40](#_Toc490695713)

[شکل ‏3‑2 شبه کدِ رویکرد پیشنهادی برای پرکردن مقادیر جاافتاده با استفاده از خوشه‌بندی و رویکردی ترکیبی 41](#_Toc490695714)

[شکل ‏3‑3 تاثیر تعداد خوشه‌ها (c) بر روی مقدار RMSE در مجموعه‌داده‌های (آ) Wholesale Customers (ب) Haberman (ج) Wine (د) Glass (ه) Iris 48](#_Toc490695715)

[شکل ‏3‑4 تاثیر درصد مقایسات (p) بر روی مقدار RMSE در مجموعه‌داده‌های (آ) Wholesale Customers (ب) Haberman (ج) Wine (د) Glass (ه) Iris 48](#_Toc490695716)

[شکل ‏3‑5 تاثیر حدآستانه (t) بر روی مقدار RMSE در مجموعه‌داده‌های (آ) Wholesale Customers (ب) Haberman (ج) Wine (د) Glass (ه) Iris 49](#_Toc490695717)

[شکل ‏3‑6 مقایسه مقدار RMSE برای رویکرد ارائه‌شده و دیگر روش‌ها در مجموعه‌داده Iris 51](#_Toc490695718)

[شکل ‏3‑7 مقایسه مقدار RMSE برای رویکرد ارائه‌شده و دیگر روش‌ها در مجموعه‌داده Glass 51](#_Toc490695719)

[شکل ‏3‑8 مقایسه مقدار RMSE برای رویکرد ارائه‌شده و دیگر روش‌ها در مجموعه‌داده Wine 52](#_Toc490695720)

[شکل ‏3‑9 مقایسه مقدار RMSE برای رویکرد ارائه‌شده و دیگر روش‌ها در مجموعه‌داده Haberman 52](#_Toc490695721)

[شکل ‏3‑10 مقایسه مقدار RMSE برای رویکرد ارائه‌شده و دیگر روش‌ها در مجموعه‌داده Wholesale Customers 53](#_Toc490695722)

[شکل ‏4‑1 رویه کلی رویکردهای پیشنهادی برای تخمین مقادیر جاافتاده 58](#_Toc490695723)

[شکل ‏4‑2 الگوریتم حداکثرسازی همبستگی هریک از صفات نسبت صفات جاافتاده 64](#_Toc490695724)

[شکل ‏4‑3 الگوریتم حداکثرسازی میانگین مقادیر همبستگی بین صفات غیر جاافتاده نسبت به صفت جاافتاده 65](#_Toc490695725)

[شکل ‏4‑4 الگوریتم اول برای گام تخمین 66](#_Toc490695726)

[شکل ‏4‑5 الگوریتم دوم برای گام تخمین 67](#_Toc490695727)

[شکل ‏4‑6 تأثیر پارامتر P بر مقدار RMSE برای روش 1 79](#_Toc490695728)

[شکل ‏4‑7 تأثیر پارامتر P بر مقدار RMSE برای روش 2 79](#_Toc490695729)

[شکل ‏4‑8 تأثیر پارامتر P بر مقدار RMSE برای روش 3 80](#_Toc490695730)

[شکل ‏4‑9 تأثیر پارامتر P بر مقدار RMSE برای روش 4 80](#_Toc490695731)

[شکل ‏4‑10 تأثیر پارامتر P بر مقدار RMSE برای روش 5 80](#_Toc490695732)

[شکل ‏4‑11 تأثیر پارامتر P بر مقدار RMSE برای روش 6 81](#_Toc490695733)

[شکل ‏4‑12 تأثیر پارامتر P بر مقدار RMSE برای روش 7 81](#_Toc490695734)

[شکل ‏4‑13 تأثیر پارامتر P بر مقدار RMSE برای روش 8 81](#_Toc490695735)

[شکل ‏4‑14 تأثیر پارامتر P بر مقدار RMSE برای روش 9 82](#_Toc490695736)

[شکل ‏4‑15 تأثیر پارامتر P بر مقدار RMSE برای روش 10 82](#_Toc490695737)

[شکل ‏4‑16 تأثیر پارامتر بر مقدار RMSE برای روش 1 83](#_Toc490695738)

[شکل ‏4‑17 تأثیر پارامتر بر مقدار RMSE برای روش 2 84](#_Toc490695739)

[شکل ‏4‑18 تأثیر پارامتر بر مقدار RMSE برای روش 3 84](#_Toc490695740)

[شکل ‏4‑19 تأثیر پارامتر بر مقدار RMSE برای روش 4 85](#_Toc490695741)

[شکل ‏4‑20 تأثیر پارامتر بر مقدار RMSE برای روش 5 85](#_Toc490695742)

[شکل ‏4‑21 تأثیر پارامتر بر مقدار RMSE برای روش6 86](#_Toc490695743)

[شکل ‏4‑22 تأثیر پارامتر θ بر مقدار RMSE برای روش 7 86](#_Toc490695744)

[شکل ‏4‑23 تأثیر پارامتر بر مقدار RMSE برای روش 8 87](#_Toc490695745)

[شکل ‏4‑24 تأثیر پارامتر بر مقدار RMSE برای روش 9 87](#_Toc490695746)

[شکل ‏4‑25 تأثیر پارامتر بر مقدار RMSE برای روش 10 88](#_Toc490695747)

[شکل ‏5‑1 رویه کلی رویکرد پیشنهادی برای تخمین مقادیر جاافتاده 101](#_Toc490695748)

[شکل ‏5‑2 مراحل الگوریتم فازی c-means مبتنی بر گری (GFCM) 102](#_Toc490695749)

[شکل ‏5‑3 مقادیر تابع هدف برحسب تعداد تکرار الگوریتم GFCM در مجموعه‌داده Iris 107](#_Toc490695750)

[شکل ‏5‑4 مقادیر تابع هدف برحسب تعداد تکرار الگوریتم GFCM در مجموعه‌داده Glass 108](#_Toc490695751)

[شکل ‏5‑5 مقادیر تابع هدف برحسب تعداد تکرار الگوریتم GFCM در مجموعه‌داده Haberman 108](#_Toc490695752)

[شکل ‏5‑6 مقادیر تابع هدف برحسب تعداد تکرار الگوریتم GFCM در مجموعه‌داده Wholesale Customers 109](#_Toc490695753)

[شکل ‏5‑7 مقادیر تابع هدف برحسب تعداد تکرار الگوریتم GFCM در مجموعه‌داده Wine 109](#_Toc490695754)

[شکل ‏5‑8 تأثیر پارامتر θ بر مقدار RMSE در مجموعه‌داده Wine 117](#_Toc490695755)

[شکل ‏5‑9 تأثیر پارامتر c بر مقدار RMSE در مجموعه‌داده Wine 118](#_Toc490695756)

# فصل اول: تعاریف پایه

## مقدمه

کامپیوتری شدن جامعه­ی جهانی، توانایی ما را در تولید و جمع­آوری داده از منابع مختلف به شکل چشم­گیری افزایش داده است. حجم عظیم داده‌ها ابعاد مختلف زندگی ما را به‌شدت تحت تأثیر قرار داده است. رشد انفجاری در داده‌های ذخیره‌شده یا گذرا[[1]](#footnote-1)، نیاز به تکنیک­هایی جدید و ابزارهای خودکار را دوچندان کرده­ است. وظیفه­ی اصلی این ابزارها تبدیل حجم عظیم داده‌های خام[[2]](#footnote-2) به دانش و اطلاعات مفید است. این امر منجر به شکوفایی و رشد سریع دانشی به نام داده­کاوی[[3]](#footnote-3) و کاربردهای مختلف آن در علوم کامپیوتر در سال­های اخیر شده است. داده­کاوی که از آن به نام کشف دانش از داده‌ها[[4]](#footnote-4) (KDD) نیز یاد می­شود، دانش و ابزار استخراج خودکار و الگوهایی مفید است که دانشِ نهان در پایگاه­ داده‌های بزرگ، وب[[5]](#footnote-5) و یا دیگر مخازن عظیم اطلاعات را به کاربران ارائه می­کند [1]. فرآیند کشف دانش مراحل مختلفی را شامل می­شود. یکی از مراحل ابتدایی و مهم­ترین مراحل، پیش­پردازش[[6]](#footnote-6) داده‌ها است. در این گام داده‌های خامِ جمع‌آوری‌شده برای استفاده در مراحل بعدی آماده می­شوند. ازجمله مهم­ترین وظایفی که در پیش­پردازش داده‌ها انجام می­شود، فرآیند پاک­سازی داده‌ها[[7]](#footnote-7) برای افزایش کیفیت داده‌های خام است. یکی از جنبه­های مهم کیفیت داده‌ها که در پاک­سازی داده‌ها مطرح می­شود، مبحث کامل بودن[[8]](#footnote-8) داده‌ها است. تأکید این جنبه از کیفیت داده‌ها بر عدم حضور مقادیر نامعلوم در داده‌ها است؛ اگرچه حضور مقادیر جاافتاده در داده‌های دنیای واقعی به دلایل مختلفی غیرقابل‌اجتناب است. بنابراین روش­ها و رویکردهای متعددی برای پُرکردن مقادیر نامعلوم در گام پیش­پردازش داده‌ها مطرح‌شده ­است.

در این فصل مفاهیم کلی داده­کاوی، اهمیت پیش­پردازش و پاک­سازی داده­ها در داده­کاوی، جایگاه فرآیند پر کردن مقادیر جاافتاده در داده­کاوی، دلایل رخ دادن این پدیده و اهمیت آن مطرح خواهد شد.

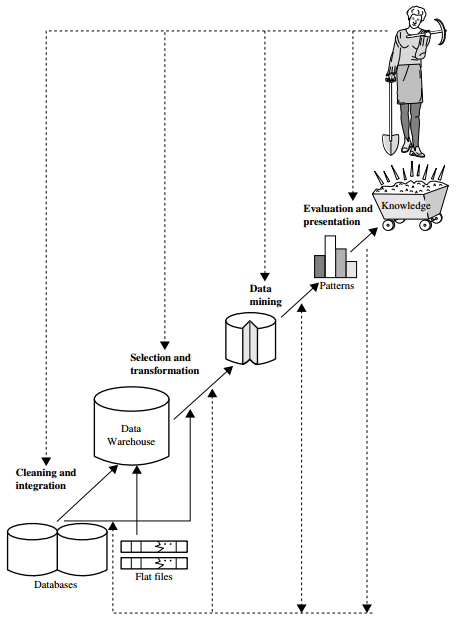
## داده‌کاوی

امروزه با توجه به حجم بالای داده‌های جمع‌آوری‌شده در زمینه­های مختلف، لازم است دانش مفید از داخل مجموعه‌داده‌ها استخراج شود [2]. زمانی كه حجم داده‌ها به‌طور چشم­گیری افزایش می­یابد، كاربران خبره هم نمي­توانند الگوهاي مفيد را در ميان حجم انبوه داده‌ها تشخيص دهند [3]. یکی از علوم کامپیوتری که طی سالیان اخیر رشدِ بدون توقفی جهت انجام این امر داشته است، علم داده­کاوی است. داده‌کاوی به‌عنوان يك ابزار قوي براي استخراج اطلاعات و دانش مفید از داده‌های خام، شناخته‌شده است. براي داده­كاوي تعاريف مختلفي وجود دارد كه معمول­ترين آن عبارت است از: استخراج دانش و اطلاعات از يك پايگاه داده بسيار بزرگ و پيچيده و همچنين كشف الگوهاي پنهان درون داده‌ها [4,5]. ابزارهاي داده­كاوي الگوهاي پنهاني را كشف و پيش­بيني مي­كنند كه متخصصان ممكن است به دليل اينكه اين اطلاعات و الگوها خارج از انتظار آن­ها باشد، آن­ها را مدنظر قرار ندهند و به آن­ها دست نيابند [6].

کشف دانش شامل مراحل مختلفی است. همان­طور که در شکل ‏1‑1 مشاهده می‌شود، مراحل کشف دانش به شرح زیر است [1]:

* *پاک­سازی داده‌ها: در این مرحله داده‌های مغشوش*[[9]](#footnote-9) *و ناسازگار حذف می‌شوند.*
* *یکپارچه‌سازی داده‌ها*[[10]](#footnote-10)*: در این مرحله داده‌هایی که در چند منبع مختلف قرار دارند تجمیع و یکپارچه می‌شوند. گاهی مراحل 1 و 2 را باهم مرحله­ی پیش‌پردازش می‌نامند. در این مرحله بر روی داده‌های خام پردازش انجام می‌شود و نتایج در مخزن داده‌ها ذخیره‌سازی می‌شود.*
* *انتخاب داده‌ها*[[11]](#footnote-11)*: در این مرحله داده‌هایی که مرتبط به وظیفه­ی تحلیل هستند، از پایگاه­ داده بازیابی*[[12]](#footnote-12) *می‌شوند.*
* *تبدیل داده‌ها*[[13]](#footnote-13)*: در این مرحله داده‌های بازیابی شده به قالبی مناسب برای داده­کاوی تبدیل می‌شوند.*
* *داده‌کاوی: مرحله­ای است که در آن الگو‌ها با استفاده از روش‌های هوشمند از داده‌ها استخراج می‌شوند.*
* *ارزیابی الگوها*[[14]](#footnote-14)*: در این مرحله الگوهایی که واقعاً مفید هستند توسط معیارهای مربوطه شناسایی می­شوند.*
* *ارائه دانش*[[15]](#footnote-15)*: تکنیک‌های مختلف نمایش و بصری سازیِ*[[16]](#footnote-16) *دانشِ کاوش شده برای کاربران در این مرحله به کار گرفته می‌شود.*

برخی کل این فرایند را کشف دانش در پایگاه داده می‌نامند و برخی نیز کل این فرآیند داده‌کاوی را اطلاق می‌کنند. الگوریتم­های داده­کاوی مانند دسته­بندی[[17]](#footnote-17)، خوشه­بندی[[18]](#footnote-18) و رگرسیون[[19]](#footnote-19) به‌طور فراگیری در فرآیند کشف دانش از مجموعه‌داده‌ها کاربرد دارند. این­ها فرآیند کشف حقیقت­ها[[20]](#footnote-20)ی جدید و شناسایی روابط یا الگوهای مفید در داده‌ها هستند [7].



شکل ‏1‑1 مراحل مختلف فرآیند کشف دانش [1]

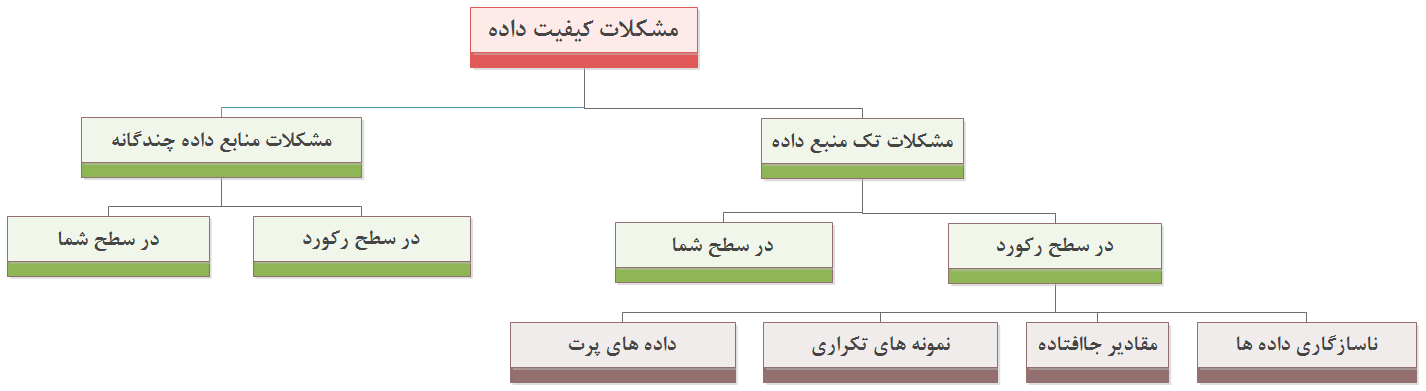
## کیفیت داده و اهمیت آن

اولین گام پیش از شروع فرآیند کشف دانش، جمع­آوری داده‌ها است که در آن نمونه­ها از تک منبع[[21]](#footnote-21) یا منابع چندگانه[[22]](#footnote-22) جمع­­آوری می­شوند [8]. داده­کاویِ باکیفیت بالا فقط هنگامی حاصل می­شود که داده‌های جمع‌آوری‌شده کیفیت بالایی داشته باشند [9]. متأسفانه موانع زیادی برای استفاده­ی مستقیم از داده‌های جمع‌آوری‌شده در گام­های بعدی کشف دانش­ وجود دارد. *داده‌های جهان واقعی به‌شدت مستعدِ پذیرش داده‌های مغشوش، جاافتاده و ناسازگار هستند. متأسفانه* داده‌های جمع‌آوری‌شده پر از نواقص[[23]](#footnote-23) مختلفی هستند. ازجمله مهم­ترین نواقص می‌توان به ناسازگاری­ها[[24]](#footnote-24)، خطاها[[25]](#footnote-25)، داده‌های پرت[[26]](#footnote-26) و مقادیر جاافتاده[[27]](#footnote-27) اشاره کرد *[10]*. *علت این امر معمولا حجم بالا داده‌ها و جمع‌آوری شدن داده‌ها از چندین منبع غیر همگن*[[28]](#footnote-28) *است* [1]*.* جمع­آوری این حجم از داده‌ها موجب ایجاد مشکل جدیدی به نام عدم کیفیت داده‌ها می‌شود. موضوع کیفیت داده‌ها در حوزه­های مختلفی از قبیل داده­کاوی، یادگیری ماشین و بازشناسی الگو[[29]](#footnote-29) مطرح ­است. *کیفیت داده‌ها شامل ابعاد متنوعی از قبیل صحت*[[30]](#footnote-30)*، دقت*[[31]](#footnote-31)*، یکپارچگی*[[32]](#footnote-32)*، سازگاری*[[33]](#footnote-33) *و کامل­ بودن،* به ­هنگام­ بودن[[34]](#footnote-34)، قابل‌اعتماد ­بودن و تفسیرپذیری[[35]](#footnote-35) *است* [1]*.*

تصمیم­گیری بر مبنای داده‌های فاقد کیفیت هزینه­های جبران‌ناپذیری را وارد­ می­کند. مطالعات انجام‌شده نشان می­دهد که بیش از 30% داده‌ها فاقد کیفیت هستند [11]. فراهم­ نمودن داده‌هایی باکیفیت بالا کاری زمان­بر و هزینه­بر است، اما در قبال هزینه­های شکست براثر داده‌های فاقد کیفیت بسیار ارزشمند است. بسیاری از پروژه­های نرم­افزاری به دلیل داده‌های فاقد کیفیت دچار شکست ­شده­اند [12]. بنابراین برای حاصل شدن نتایج مطلوب از هدف کشف دانش لازم است که این داده‌های معیوب پاک­سازی شوند. این پاک­سازی­ معمولاً در گام پیش­پردازش انجام می­شود. پیش­پردازش داده‌ها اهمیت فراوانی دارد [13]، به‌نحوی‌که حدود 80% از کل تلاش برای مهندسی داده[[36]](#footnote-36) صرف آماده­سازی داده‌ها[[37]](#footnote-37) برای مراحل بعد می­شود [14]. پیش­پردازش داده‌ها شامل گام­هایی مثل تبدیل نوع داده[[38]](#footnote-38)، مدیریت داده‌های پرت، نرمال‌سازی مقادیر[[39]](#footnote-39) و مدیریت مقادیر جاافتاده است *[10]*. *در ادامه پس از تعریف و دسته‌بندی مشکلات کیفیت داده، به‌طور خاص بر مشکل مقادیر جاافتاده در داده‌ها تمرکز می­شود.*

## مشکلات کیفیت داده‌ها

در این بخش مسائل اصلی کیفیت داده که باید توسط پاک‌سازی و تبدیل داده­ها حل شوند، دسته‌بندی می­شوند. این مسائل اکثراً بسیار به یکدیگر مرتبط هستند و با یکدیگر هم‌پوشانی دارند. [15] مشکلات مربوط را بر اساس تکی یا چندگانه بودن منابع و در سطح شِما یا رکورد بودنِ مشکل دسته‌بندی کرده است. شکل ‏1‑2 یک دسته­بندی پیشنهادی از مشکلات کیفیت داده­ها را نمایش می­دهد. (این شکل مال خودتون بود؟ ارجاع نداشت؟)



شکل ‏1‑2 دسته‌بندی مشکلات کیفیت داده

مشکلات در سطح شِما در سطح نمونه نیز منعکس خواهند شد. این مشکلات می‌توانند با بهبود طراحی شِما (تکامل شِما[[40]](#footnote-40))، تبدیل شِما و یکپارچه‌سازی شِما برطرف شوند. مشکلات در سطح نمونه به مشکلاتی گفته می‌شود در داخل خود داده وجود دارند و از روی شِما قابل‌تشخیص نیستند.

### مشکلات تک منبع

کیفیت داده در یک منبع به‌شدت به میزان محدودیت‌های شِما[[41]](#footnote-41) و یکپارچگی که کنترل­کننده‌ی مقادیر داده هستند، وابسته است. برای منابع بدون شِما مانند فایل‌ها، محدودیت کمی بر روی ذخیره‌ و ورود داده‌ها وجود دارد؛ بنابراین احتمال بروز خطا و ناسازگاری در این منابع بالاست. در سوی مقابل در پایگاه داده‌ها، محدودیت‌های زیادی بر روی مدل داده مانند محدودیت‌های ارجاع[[42]](#footnote-42) اعمال می‌شود. اکثر خطا‌ها در سطح شِما به ‌دلیل طراحی ضعیف شِما رخ می‌دهند. خطاهای در سطح نمونه، خطاها و ناسازگاری‌هایی هستند که نمی‌توان از رخ داد آن‌ها در سطح شِما جلوگیری نمود؛ مانند غلط‌های املایی هنگام ورود داده. برای هر دو نوع خطا می‌توان چهار محدوده برای رخ داد خطا تعیین نمود: صفت[[43]](#footnote-43)، رکورد، نوع رکورد[[44]](#footnote-44) و منبع.

در این گزارش به مشکلات تک منبع داده و در سطح رکورد و به‌طور خاص به مشکل مقادیر جاافتاده پرداخته می‌شود. مسائل سطح رکورد در تک منبع داده را می‌توان به چهار دسته‌ی کلی زیر تقسیم کرد:

* **نمونه‌های تکراری**[[45]](#footnote-45)**:**

یکی از مشکلات موجود در مجموعه‌داده­ها حضور نمونه­هایی است که همگی به یک موجودیت در جهان واقعیت اشاره دارند و همگی بیانگر اطلاعاتی یکسان هستند. بنابراین با پدیده افزونگی[[46]](#footnote-46) داده روبه‌رو هستیم. فرآیند مشخص کردن این‌که آیا دو رکورد به یک موجودیت یکسان در جهان واقع اشاره دارند یا خیر را عمل ارتباط رکورد[[47]](#footnote-47) می‌نامند. این فرآیند به‌ویژه هنگام یکپارچه‌سازی دو یا چند منبع داده با شمای یکسان ضروری است [16]. عمل شناسایی و تطابق نمونه‌های تکراری و حذف یکی از نمونه‌های تکراری را عمل حذف نمونه‌های تکراری می‌نامند.

* **ناسازگاری داده‌ها:**

یکی از مسائل اصلی داده‌های دنیای واقعی سازگاری داده‌ها است. درواقع اغلب داده‌های دنیای واقعی مغشوش و همراه با خطا و تناقضات هستند. ناسازگاری‌ها می‌توانند بسیار پرهزینه باشند. برای مثال سالانه 600 بیلیون دلار در تجارت‌های آمریکا صرف تصحیح ناسازگاری‌ها می‌شود [17].

* **داده‌های پرت:**

فرآیند پیدا کردن داده‌هایی با رفتاری بسیار متفاوت‌تر از رفتار مورد انتظار با توجه به کل نمونه‌ها را شناسایی داده‌های پرت می‌نامند. این فرآیند در کاربردهایی مانند تشخیص تقلب، پزشکی، پردازش تصویر و تشخیص نفوذ کاربرد دارد [1].

* **داده­های جاافتاده:**

یکی از رایج­ترین مشکلات دنیای واقعی عدم حضور برخی مقادیر در مجموعه‌داده‌ها به دلایل مختلف است. بنابراین با روش­های مختلف پاک­سازی داده­ها تلاش می­شود تا مقادیر جاافتاده در یک مجموعه‌داده تا حد ممکن به‌طور دقیق تخمین زده شوند.

### مشکلات چند منبع

مشکلات داده‌ی تک منبع هنگامی‌که نیاز به چندین منبع است، حادتر خواهد شد. هریک از منابع ممکن است شامل داده‌های معیوب بوده و ممکن است هرکدام به‌صورت متفاوت نمایش‌ داده شوند و با یکدیگر تناقض یا هم‌پوشانی داشته باشند. دلیل امر این است که اغلب منابع داده به‌صورت مستقل از یکدیگر توسعه می‌یابند و نگه‌داری می‌شوند. در سطح شِما، مشکل اصلی برای طراحی شِما تضادهای نام‌گذاری[[48]](#footnote-48) و تضادهای ساختاری[[49]](#footnote-49) بین منابع داده است [15].

## پاک‌سازی داده

پاک‌سازی داده (سایش داده[[50]](#footnote-50)) عملیات شناسایی و حذف خطا‌ها و ناسازگاری‌ها از داده به‌منظور بهبود کیفیت داده است. مشکلات کیفیت داده در منابع داده‌ی تکی مانند فایل‌ها یا پایگاه داده‌ها به دلایل اشتباه املایی در هنگام ورود داده، اطلاعات اشتباه یا دیگر داده‌های نامعتبر رخ دهد. نیاز به پاک‌سازی داده‌ها برای رفع مشکلات کیفیت داده در منابع داده‌ی چندگانه که نیاز به یکپارچه­سازی منابع دارند، مثل پایگاه داده‌ی تحلیلی[[51]](#footnote-51) و پایگاه داده فدرال[[52]](#footnote-52) و یا سیستم‌های اطلاعاتی مبتنی بر وب، به‌طور چشم‌گیری افزایش می‌یابد. فرآیند پاک‌سازی داده هزینه­بر است و جلوگیری از ورود داده‌ی خطا یکی از مهم‌ترین گام‌ها در ساده‌تر کردن مسئله‌ی پاک‌سازی است. برای این منظور طراحی شِمای مناسب، اعمال محدودیت‌های لازم و نحوه‌ی ورودی برنامه کاربردی مناسب لازم است. درواقع هدف از پاک­سازی داده‌ها دستیابی به داده‌هایی بدون تکرار، خطای املایی و ناسازگاری و مطابق با ساختار تعریف­شده توسط کاربران است. با استفاده از پاک­سازی داده پیش از داده‌کاوی می‌توان کیفیت داده را بهبود بخشید. کیفیت بالای داده‌ برای هر کاربردی مهم است. تشخیص و تصحیح داده‌ فرآیندی زمان‌بر است؛ اگرچه امری غیرقابل‌اجتناب است. ناهنجاری و ناخالصی داده‌ ممکن است منجر به تحلیل ناکارای داده‌، تصمیم‌گیری غلط و عدم رضایت کاربران شود [18]. عملیات پاک‌سازی داده در­ حوزه‌ی داده‌های ایستا‌ معمولاً تنها یک‌بار پیش از پردازش اصلی انجام می‌شود.

## مقادیر جاافتاده

در دنیای واقعی، داده‌ها اغلب دارای مقادیر جاافتاده هستند. یک وظیفه­ی مهم در پاک­سازی داده‌ها و پیش­پردازش، تخمین تا حد ممکن دقیقِ مقادیر جاافتاده برای افزایش کیفیت داده‌ها است. به بیان ساده، یک مقدار جاافتاده مقداری است که به هر دلیلی معرفی نشده ­است و یا در گام جمع­آوری داده‌ها گم­شده است. بنابراین مقداری نامعلوم[[53]](#footnote-53) یا تهی[[54]](#footnote-54) دارد [19]. وجود مقادیر جاافتاده در داده‌ها نه‌تنها مشکلی رایج در بسیاری از تحقیقات تجربی مثل تحقیقات صنعتی، تجاری و علمی است، بلکه مشکلی غیرقابل‌اجتناب است [20,21]. برای مثال، حدود 45% از مجموعه­های موجود در مخزن داده­ی UCI [22] که یکی از مشهورترین مخازن داده برای اهداف یادگیری ماشین است، شامل مقادیر جاافتاده هستند [23]. اینجا جایی است که نقشِ عملیات جایگذاری مقادیر جاافتاده[[55]](#footnote-55) ظاهر می­شود. هدف از جایگذاری مقادیر جاافتاده که نام­های دیگر آن تخمین مقادیر جاافتاده[[56]](#footnote-56) یا پر کردن مقادیر جاافتاده[[57]](#footnote-57) است، تخمین مقادیری برای پر کردن مقادیر جاافتاده با استدلال از داده‌های کاملِ مشاهده‌شده است [8]. بنابراین پس از انجام عملیات جایگذاری مقادیر جاافتاده، مجموعه­داده­ی کامل به دست خواهد آمد. مستقل از این‌که یک پرس‌وجو[[58]](#footnote-58) و پایگاه داده چقدر کارا باشد، باوجود مقادیر جاافتاده نمی‌توان تضمین کرد که داده‌ها چه اندازه صحیح و دقیق خواهند بود. مقادیر جاافتاده به دلایل متفاوتی ازجمله غفلت انسان، نقض قوانین و یا محدودیت‌های خارجی رخ دهند. گستره‌ی این مشکل بسته به کاربرد می‌تواند متفاوت باشد. داده‌های ناکامل استفاده­پذیری[[59]](#footnote-59) از داده را تحت تأثیر قرار می‌دهد، بنابراین لازم است که مقادیر جاافتاده تخمین زده شوند [24]. جایگذاری مقادیر جاافتاده همانند دیگر وظایف پیش­پردازش امری زمان­بر است [25].

### علل ایجاد مقادیر جاافتاده

مقادیر جاافتاده ممکن است به دلایل مختلفی در داده‌ها وجود داشته باشند. موارد زیر علل اصلی ایجاد مقادیر جاافتاده در فرآیند جمع­آوری و ذخیره داده‌ها هستند [26,27,28,18,20,29]:

* خرابی سیستم قدرت[[60]](#footnote-60)
* عوامل محیطی مثل رطوبت و دما
* عدم پاسخ در آزمایش­های علمی
* اندازه­گیری­های غلط
* خطای انتقال داده در سیستم­های دیجیتال
* رویه­های دستی برای ثبت داده­[[61]](#footnote-61)
* اختلال در عملکرد تجهیزات ذخیره­سازی داده

### مشکلات وجود مقادیر جاافتاده

مقادیر جاافتاده به‌ویژه هنگامی‌که نرخ جاافتادگی[[62]](#footnote-62) بسیار بالا باشد، موجب ایجاد مشکلاتِ جدی و متعددی می­شود. حضور مقادیر جاافتاده به‌طور مستقیم کیفیت نتایج نهایی به‌دست‌آمده از فرآیند یادگیری یا کاوش از داده‌ها را تحت تأثیر قرار می­دهد و ممکن است منجر به تنزل شدید در عملکرد این الگوریتم­ها شود [8]. وجود مقادیر جاافتاده ممکن است منجر به از دست رفتن اطلاعات مهم شود [30]. غیرقابل اعمال بودن[[63]](#footnote-63) روش­های تحلیل داده­ به مجموعه­داده‌های دارای مقادیر جاافتاده یکی از جدی­ترین مشکلات وجود مقادیر جاافتاده در داده‌ها است. بسیاری از الگوریتم­های تحلیل داده‌ها قابلیت اعمال به مجموعه‌داده‌هایی با مقادیر جاافتاده را ندارند، چراکه اصولاً این الگوریتم­ها به‌گونه‌ای طراحی‌شده‌اند که فقط با داده‌های کامل می‌توانند کار کنند. بنابراین این روش­های تحلیل داده نمی‌توانند به داده‌های اصلی که شامل مقادیر نامعلوم هستند، اعمال شوند [31].

وجود مقادیر جاافتاده می‌تواند منجر به نتیجه­گیری­های گمراه‌کننده، غیر صحیح و غیرمعقول از یک تحقیق علمی شود و ممکن است کل فرآیند جمع­آوری و تحلیل داده‌ها را برای کاربران بی­فایده کند [32]. علاوه بر این، این نقص در داده‌ها می‌تواند قابلیت تعمیم[[64]](#footnote-64) را محدود ­کند [29] و اعتبار یافته­های یک تحقیق را زیر سؤال ببرد [9]. مدیریت نامناسب مقادیر جاافتاده ممکن است به دلیل اختلاف­های موجود بین داده‌های جاافتاده و داده‌های کامل نتایجی جانب‌دارانه[[65]](#footnote-65) را حاصل کند [33,34,31].

## خلاصه و نتیجه‌گیری

در این فصل ابتدا به بررسی مفاهیم داده‌کاوی، کیفیت داده‌ها و پاک­سازی داده‌ها پرداخته شد. سپس به‌طور دقیق مشکل مقادیر جاافتاده در داده‌ها، علل ایجاد این پدیده و عواقب وجود آن بررسی شد.

در این پایان­نامه ابتدا در فصل ‏2 به بررسی کارهای پیشینِ انجام‌شده در زمینه­ی جایگذاری مقادیر جاافتاده پرداخته‌شده­ است. در فصل­های بعدی رویکردهای جدیدی برای این هدف ارائه‌شده است. در فصل ‏3 با استفاده از روشی جدید مبتنی بر خوشه‌بندی و رویکردی ترکیبی مقادیر جاافتاده عددی پر می­شوند. همچنین جزئیات روش ارائه‌شده و نتایج آزمایش­های انجام‌شده به‌تفصیل ارائه‌شده‌اند. در فصل ‏4 چندین رویکرد جدید به نام روش­های جایگذاری مقادیر جاافتاده مبتنی بر همبستگی[[66]](#footnote-66) (CMIM) برای تخمین مقادیر جاافتاده عددی مطرح می­شود. در فصل ‏5 رویکردی جدید مبتنی بر خوشه‌بندی فازی c-means مبتنی بر گِرِی[[67]](#footnote-67) و انتخاب ویژگی مبتنی بر معیار اطلاعات متقابل[[68]](#footnote-68) برای تخمین مقادیر جاافتاده عددی مطرح می‌شود. درنهایت، در فصل ‏6 نتیجه‌گیری، بحث و پیشنهاد‌ها بیان می‌شوند.

# فصل دوم: پیشینه تحقیق

## مقدمه

همان­طور که در فصل پیش بیان شد، حضور مقادیر جاافتاده در مجموعه­داده­های جهان واقعی یک مشکل بسیار رایج است. راه‌حل رفع این مشکل، جایگذاری مقادیر جاافتاده است. در این فصل، مروری بر روش­های مختلف مطرح‌شده برای هدف جایگذاری مقادیر جاافتاده انجام می­شود. این روش­ها طیف گسترده­ای از علوم مختلف مانند داده­کاوی و یادگیری ماشین را شامل می­شود. قبل از بحث درباره­­ی روش­های مختلف جایگذاری مقادیر جاافتاده، مروری بر انواع مکانیزم (تولید) داده‌های جاافتاده[[69]](#footnote-69) می­شود. سپس، سه معیار شناخته‌شده برای ارزیابی دقت روش­های جایگذاری مقادیر جاافتاده معرفی می­شود.

## مکانیزم‌های ایجاد مقادیر جاافتاده

مقادیر جاافتاده ممکن است توسط مکانیزم­های مختلفی ایجاد شوند. اگر نشان‌دهنده‌ی یک متغیر تصادفی باشد و یک مجموعه‌داده صفت داشته باشد، آنگاه سه نوع مکانیزم متفاوت برای تولید مقادیر جاافتاده وجود خواهد داشت [35,25,36]:

1. جاافتاده به‌طور کاملاً تصادفی[[70]](#footnote-70) (MCAR):

در این حالت احتمال این­که برای هر نمونه، مقدار یک صفت مثل معلوم یا جاافتاده باشد، وابسته به هیچ‌یک از دیگر صفات نیست و کاملاً مستقل از دیگر صفات است. یعنی:

|  |  |
| --- | --- |
| ‏2‑1 |  |

2. جاافتاده به‌طور تصادفی[[71]](#footnote-71) (MAR):

در این حالت احتمال این­که برای هر نمونه، مقدار یک صفت مثل معلوم یا جاافتاده باشد، ممکن است به دیگر صفات وابسته باشد ولی به مقدار خود آن صفت وابسته نیست. یعنی:

|  |  |
| --- | --- |
| ‏2‑2 |  |

3. جاافتاده به‌طور غیر تصادفی (MNAR[[72]](#footnote-72)):

در این حالت احتمال این­که برای هر نمونه مقدار یک صفت مثل معلوم یا جاافتاده باشد، ممکن است به مقدار خود آن صفت نیز وابسته باشد.

به بیان ساده، این اصطلاحات روابط بین صفات جاافتاده و احتمال وجود داده‌های جاافتاده را بیان می­کنند [36]. اگرچه در دنیای واقعی، اینکه داده‌ی جاافتاده در کدام دسته­بندی ذکرشده قرار می‌‌گیرد معمولاً غیرممکن است، زیرا نحوه­ی ایجاد شدن مقدار جاافتاده نامعلوم است [37].

## رویکردهای جایگذاری مقادیر جاافتاده

محققان زیادی به مسئله­ی جایگذاری مقادیر جاافتاده توجه کرد­ه­اند تا جایگذاری را دقیق و کارا سازند. این تحقیقات شامل طیف وسیعی از روش­ها می­شود؛ از روش­های ساده­ی آماری تا روش­های پیچیده­ی مبتنی بر داده­کاوی و یادگیری ماشین. در بین این روش­ها، تعدادی روش شناخته‌شده و ساده ازلحاظ محاسباتی وجود دارد:

• نادیده گرفتن/حذف نمونه­هایی با مقادیر جاافتاده

• پر کردن دستی مقادیر جاافتاده

• پر کردن با مقدار ثابت سراسری یا مقادیر محتمل[[73]](#footnote-73) (مانند مقدار میانگین یا مقدار میانه)

ساده­ترین راه مدیریت مقادیر جاافتاده نادیده گرفتن نمونه­هایی با مقادیر جاافتاده است (رویکرد اول). بااین‌حال این رویکرد ممکن است منجر به از دست رفتن دیگر اطلاعات مفیدِ موجود در نمونه­های جاافتاده شود؛ به‌ویژه زمانی که نرخ جاافتادگی بالاست. به‌علاوه، این رویکرد ممکن است منجر به سوی­گیری[[74]](#footnote-74) احتمالی نتایج تخمین­ها شود [38]. [35] خطرات حذف نمونه­های جاافتاده را بررسی کرده است.

رویکرد دوم فقط وقتی عملی است که در داده‌ها تعداد نسبت کمی مقدار جاافتاده وجود دارد. در غیر این صورت این رویکرد بسیار زمان­بر و هزینه­بر خواهد بود. به‌علاوه خطای انسانی نیز می‌تواند نتیجه نهایی را تحت تأثیر قرار دهد.

جایگذاری مقادیر جاافتاده با مقادیر محتمل (رویکرد سوم)، مثل مقدار میانگین برای صفات کمّی[[75]](#footnote-75) و مقدار مد[[76]](#footnote-76) برای صفات کِیفی[[77]](#footnote-77)، انتخاب دیگری است. مشکل این رویکرد نادیده گرفتن روابط بین صفات است. به‌علاوه در این رویکرد بحث توزیع[[78]](#footnote-78) داده‌ها نادیده گرفته خواهد شد؛ چراکه تمام مقادیر جاافتاده با یک مقدار یکسان پر خواهند شد. [39] معتقد است که این رویکرد نیز ممکن است منجر به تولید تخمین­های جانبدارنه شود.

درنتیجه‌ی ضعف­های روش­های ذکرشده، روش­هایی پیچیده­تر پیشنهادشده‌اند. در ادامه دسته­بندی­های مختلف برای رویکردهای جایگذاری مقادیر جاافتاده ارائه می­شوند.

### رویکردهای مبتنی بر شبکه‌های عصبی

رویکردهای مبتنی بر شبکه­های عصبی مصنوعی[[79]](#footnote-79) به‌طور گسترده برای هدف جایگذاری مقادیر جاافتاده استفاده‌شده­اند. [40] و [41] با استفاده از آموزش شبکه­های عصبی نگاشت خودسازمان‌ده[[80]](#footnote-80) (SOM) برای پر کردن مقادیر جاافتاده استفاده کرده است. در این روش، پس از آموزش شبکه­ی SOM توسط نمونه­های کامل، نمونه­های جاافتاده به شبکه­ی آموزش­دیده داده می‌شوند تا پر شوند.

در مطالعات زیادی از شبکه­های عصبی پرسپترون چندلایه[[81]](#footnote-81) (MLP) برای هدف جایگذاری مقادیر جاافتاده استفاده‌شده است [42,43,44,45,46,47,48]. پر کردن با استفاده از آموزش شبکه­های MLP به‌عنوان یک مدل رگرسیون غیرخطی بر روی نمونه­های کامل و انتخاب یک صفت به‌عنوان صفت هدف[[82]](#footnote-82) انجام می­شود. در [25] با استفاده از شبکه‌ی عصبی پرسپترون چندلایه، داده‌های جاافتاده برای هجده مجموعه‌داده تخمین زده‌شده‌اند. همچنین تأثیر روش‌های مختلف یادگیری و پارامترهای مختلف شبکه عصبی بر روی دقت نهایی نیز بررسی‌شده‌اند. [49] و [50] از شبکه­های عصبی انجمنی خودکار[[83]](#footnote-83) برای هدف جایگذاری مقادیر جاافتاده استفاده کرده­است. مشکل اصلی شبکه‌های عصبی زمان‌بر بودن فرآیند یادگیری و یافتن مدل بهینه برای یک مجموعه‌داده نسبت به دیگر روش‌ها و همچنین قابلیت تفسیر پایین است [1]. به‌علاوه، اصولاً آموزش مناسب شبکه­های عصبی نیاز به تعداد نمونه­های ورودی بالایی دارد؛ بنابراین هنگامی‌که تعداد نمونه­های جاافتاده بالا است، دقت تخمین این روش­ها کاهش می­یابد.

### رویکردهای مبتنی بر روش k نزدیک‌ترین همسایگان

جایگذاری با استفاده از k نزدیک­ترین همسایگان[[84]](#footnote-84) (k-NNI) [51,52,53,54,55] یکی از معروف­ترین و پایه‌ترین روش­های تخمین مقادیر جاافتاده است. دلایل اصلی این امر سادگی مفهومی، راحتی پیاده­سازی و تولید نتایج نسبت قابل‌قبول نسبت به دیگر روش­ها است [56]. در این روش ابتدا تعداد k نزدیک­ترین همسایگان نسبت به نمونه­ی جاافتاده پیدا می­شوند. سپس مقدار جاافتاده با استفاده از اطلاعات موجود در نمونه­های انتخاب‌شده محاسبه می­شود. معمولاً از پُرتکرارترین مقدار در نمونه­های پیداشده برای جایگذاری صفت جاافتاده در صفات اسمی[[85]](#footnote-85) و از مقدار میانگین برای جایگذاری صفت جاافتاده در صفات عددی استفاده می­شود.

*در* [57] *از یک رویکرد مبتنی بر روش* k *نزدیک‌ترین همسایگان برای پر کردن مقادیر جاافتاده استفاده‌شده است. در این رویکرد ابتدا درجه‌ی جاافتادگی هر نمونه که برابر است با تعداد صفات جاافتاده در هر نمونه، محاسبه می‌شود. نمونه‌هایی که درجه‌ی جاافتادگی آن‌ها بزرگ‌تر از دو است از ادامه‌ی پردازش خارج می‌شوند. برای نمونه‌های باقی‌مانده، نمونه‌ها به دو مجموعه‌ نمونه‌های با تک مقدار جاافتاده () و نمونه‌های بدون مقدار جاافتاده () تقسیم می‌شوند. برای هر نمونه از مجموعه‌ی نزدیک­ترین رکورد به آن نمونه از بین رکوردهای پیداشده و برای مقدار جاافتاده، مقدار آن صفت با نزدیک‌ترین نمونه­ی یافت شده پر خواهد شد. از نقاط قوت این روش می‌توان به‌سادگی و قابل‌فهم بودن پیاده‌سازی رویکرد ارائه‌شده (نه لزوماً پیچیدگی محاسباتی) و قدرت تعمیم بالا برای کاربردهای مختلف با تغییر معیار شباهت اشاره کرد. عدم توانایی در پر کردن مقادیر جاافتاده‌ی چندگانه، حساسیت به داده‌های پرت، حساسیت بالا به تابع شباهت از نقاط ضعف این روش است. همچنین بررسی فقط نزدیک‌ترین نمونه لزوماً بهترین گزینه نیست.*

[58] و [59] از روش­های جایگذاری k-NN مبتنی بر گِرِی[[86]](#footnote-86) استفاده نموده­اند. در این روش­ها، از نمره‌ی رابطه­ای گِرِی[[87]](#footnote-87) (GRG) برای یافتن نزدیک­ترین همسایگانِ یک نمونه­ی جاافتاده استفاده‌شده است. [58] از یک رویکرد تکراری مبتنی بر گِرِی به نام [[88]](#footnote-88)GBKII استفاده کرده­ است. رویکرد مطرح‌شده در یک فرآیند تکراری بر روی مقادیر جاافتاده، تک‌به‌تک مقادیر جاافتاده را پر می‌نماید. *در این روش، شباهت دو نمونه (برای جستجوی نزدیک‌ترین همسایگان) فقط روی نمونه‌های هم برچسب*[[89]](#footnote-89) *محاسبه می‌شود.* امکان پر کردن مقادیر جاافتاده چندگانه و کاراتر بودن پیچیدگی زمانی به دلیل اینکه تنها بین نمونه‌های هم برچسب مقایسه شباهت انجام می‌شود، از مزیت‌های این روش است. نیاز به دانستن برچسب کلاسِ نمونه‌ها، ضعف اصلی این روش است. *معیار شباهت استفاده‌شده در این روش ضریب رابطه‌ای گِرِی*[[90]](#footnote-90) *(*GRC*) است که برگرفته از مفهوم تحلیل رابطه‌ای گِری در نظریه اطلاعات است.*

در [60] دقت رویکرد جایگذاری با استفاده از k نزدیک­ترین همسایگان بر روی چهار مجموعه‌داده با دو معیار سنجش فاصله‌ی مَنهَتن[[91]](#footnote-91) و اقلیدسی[[92]](#footnote-92) بررسی‌شده است و نتایج با روش پر کردن با میانگین مقایسه شده‌ است. نتایج نشان داده است که روش k-NNI بهتر از روش جایگذاری با مقدار میانگین عمل می­کند. [61] به‌طور خاص‌منظوره از این روش برای تخمین مقادیر جاافتاده در کاربرد برای تشخیص سرطان سینه استفاده کرده است. در [56] از یک نسخه بهبودیافته روش k نزدیک‌ترین همسایه و استفاده از خوشه‌بندی داده‌های قرارگرفته در یک دسته به نام "k خوشه‌ و نزدیک‌ترین همسایه‌ی مبتنی بر دسته[[93]](#footnote-93)" (CKNNI) برای تخمین مقادیر جاافتاده استفاده‌شده است. مشکل اصلی این روش نیاز به دانستن برچسب دسته­ی تمام نمونه‌ها است. درواقع اگر دسته­ی نمونه‌ها مشخص نباشند، نمی‌توان از روش ذکرشده استفاده نمود.

نوع دیگری از روش­های جایگذاری با k نزدیک­ترین همسایگان، نسخه­ی وزن­دار[[94]](#footnote-94) آن است. در این روش، فاصله­ هر نزدیک­ترین همسایگان تا نمونه­ی جاافتاده محاسبه می­شود. سپس از فواصل محاسبه‌شده به نحوی در تخمین نهایی مقادیر جاافتاده استفاده می­شود [62].

مزیت اصلی روش­های مبتنی بر k-NN، راحت بودن قابلیت استفاده هم برای صفات عددی و هم صفات اسمی با انتخاب مناسب تابع فاصله است. به‌علاوه در این روش­های نیازی به ساخت مدل پیش­بینی کننده[[95]](#footnote-95) نیست. بنابراین زمان محاسباتی صرف ساخت مدل نمی­شود [56,39]. مشکل اصلی این روش­ها این است که ممکن است برای مجموعه­داده‌های بزرگ زمان زیادی برای جایگذاری صرف شود، چراکه در این روش­ها برای هر نمونه­ی جاافتاده، روی کل مجموعه‌داده جستجوی کامل انجام می­شود. به‌علاوه، تعیین دقیق تابع فاصله و مقدار مناسب برای k ممکن است چالش‌برانگیز باشد [63].

### رویکردهای مبتنی بر الگوریتم حداکثرسازی امید

جایگذاری با رویکرد حداکثرسازی امید[[96]](#footnote-96) (EMI) [64] روش شناخته‌شده‌ی دیگری برای جایگذاری مقادیر جاافتاده­ی عددی است. در این روش ماتریس­های میانگین و کوواریانس[[97]](#footnote-97) با استفاده از داده‌ها موجود تخمین زده ­می­شوند [65]. در ابتدا الگوریتم EMI با تخمین­های اولیه از ماتریس­های میانگین و کوواریانس شروع می­کند. سپس، الگوریتم آن­قدر ادامه می­یابد تا مقادیر تخمین زده‌شده و تخمین­های ماتریس­های میانگین و کوواریانس به‌طور چشم­گیری از تکرار فعلی به تکرار بعدی تغییر نکنند [66]. مشکل اصلی این روش این است که در این روش از اطلاعات کل مجموعه‌داده‌ها استفاده می­شود و هنگامی خوب عمل می­کند که همبستگی بین صفات قوی باشد. بنابراین برای مجموعه‌داده‌هایی مناسب خواهد بود که در آن­ها همبستگی­های قوی­ بین صفات بر روی کل مجموعه‌داده برقرار باشد. در [18] سعی کردند تا این مشکل را با اعمال الگوریتم EMI بر روی بخش­هایی از داده‌ها تا حدودی برطرف کند. ایده‌ی اصلی آن­ها این است که همبستگی و روابط بین صفات درون یک افراز از داده‌ها می‌تواند قوی‌تر از همبستگی بر روی کل مجموعه‌داده باشد و بنابراین الگوریتم EM بر روی افرازهایی از مجموعه‌داده بهتر عمل خواهد کرد و درنهایت نتیجه به‌دست‌آمده بهتر خواهد شد. ازاین‌رو الگوریتم پیشنهادی آن­ها، که "جایگذاری مقادیر جاافتاده مبتنی بر درخت تصمیم[[98]](#footnote-98) (DMI)" نام دارد، بخش­هایی از داده‌های ورودی را توسط درخت تصمیم پیدا می­کند که همبستگی قوی بر آن­ها حاکم است. سپس الگوریتم EMI را برافرازهای یافته شده اعمال می­کند. *الگوریتم ارائه‌شده دارای پنج مرحله به شرح زیر است:*

1. *تقسیم مجموعه‌داده­ها به دودسته‌ی و ، به‌نحوی‌که تنها شامل داده­های صحیح و تنها شامل داده­هایی که مقادیر جاافتاده دارند، است.*
2. *ساخت درخت تصمیم از روی مجموعه برای ویژگی­هایی که مقادیر جاافتاده دارند به‌طوری‌که برچسب برای ساختن درخت را برابر با صفت جاافتاده در نظر گرفته شود.*
3. *نسبت­ دادن هر یک از رکورد­های مجموعه به یکی از برگ­های درخت. درصورتی‌که در یک رکورد بیش از یک مقدار جاافتاده وجود­ داشت، آن را به چند درخت نسبت­ می­دهد.*
4. *درصورتی‌که مقدار جاافتاده جزء مقادیر عددی باشد، مقدار جاافتاده را با اجرای الگوریتم* EM *بر روی رکورد­هایی که در آن برگ واقع‌شده‌اند، محاسبه ­می­کند؛ اما درصورتی‌که جزء مقادیر دسته­ای باشد، کلاس اکثریت آن برگ را به‌عنوان­ جایگزین مقدار جاافتاده در­ نظر ­گرفته­ می­شود.*
5. *ادغام داده­های پرشده* *با داده­های صحیح.*

اگرچه این روش بخش­هایی با همبستگی بالا را توسط درخت تصمیم پیدا می­کند، بااین‌حال از معیار دقیقی برای سنجش میزان همبستگی بین صفات استفاده نمی‌کند و همواره فرض می‌کند که استفاده از درخت تصمیم منجر به افزایش همبستگی در داده‌ها می‌شود. ازاین‌رو، ضروری به نظر می­رسد که مقدار همبستگی را توسط معیاری دقیق برای اهداف ذکرشده اندازه­گیری کرد.

### رویکردهای مبتنی رگرسیون

معروف­ترین روش­های جایگذاری مقادیر جاافتاده در آمار، روش­های مبتنی بر رگرسیون است [67,68]. در این روش­ها داده­ی جاافتاده با استفاده از روابط موجود بین صفت جاافتاده و دیگر صفات از طریق ساخت معادله­ی رگرسیون تخمین زده می­شود [68]. مدل رگرسیون بر روی داده‌های ورودی برازش[[99]](#footnote-99) می­شود تا بتواند مقدار یک متغیر را از روی دیگر متغیرها، پیش­بینی کند. متغیری که پیش­بینی می­شود متغیر پاسخ[[100]](#footnote-100)، متغیر وابسته[[101]](#footnote-101) یا متغیر هدف[[102]](#footnote-102) نامیده می­شود. متغیرهایی که متغیر پاسخ از روی آن پیش­بینی به می­شود، پیش­بینی کننده­ها[[103]](#footnote-103) یا متغیرهای مستقل[[104]](#footnote-104) نامیده می­شوند. مهم­ترین ویژگی این روش­ها، استفاده از روابط موجود بین صفات برای پیش­بینی مقادیر جاافتاده است. انواع مختلفی از رگرسیون­ها برای مدیریت داده‌های جاافتاده وجود دارد. رگرسیون خطی چندگانه (چند صفتی)[[105]](#footnote-105) برای صفات کمّی، رگرسیون منطقی[[106]](#footnote-106) برای هنگامی‌که متغیر وابسته دوبخشی (دو مقداره)[[107]](#footnote-107) است و رگرسیون منطقی چندجمله­ای[[108]](#footnote-108) برای رسیدگی به صفات جاافتاده­ی دسته­ای[[109]](#footnote-109) هنگامی‌که بیش از دودسته وجود دارد، از مهم­ترین انواع رگرسیون هستند [47]. [37] دنباله­ای از مدل­های رگرسیون را به‌طور تکراری[[110]](#footnote-110) به مقادیر جاافتاده برای هدف جایگذاری مقادیر جاافتاده اعمال می­کند. این روش از برچسب دسته­ی هر نمونه نیز علاوه بر دیگر صفات به‌عنوان متغیر پیش­بینی کننده استفاده می­کند.

### رویکردهای مبتنی قوانین انجمنی

دسته‌ای از روش‌ها با استفاده از قوانین انجمنی[[111]](#footnote-111) داده‌های جاافتاده را می­کنند. در [69] ابتدا مجموعه‌داده‌ی رابطه‌ای به مجموعه‌داده‌ی تراکنشی[[112]](#footnote-112) تبدیل می‌شود تا بتوان از آن قوانین انجمنی را استخراج نمود. *پس از یافتن قوانین یافت شده از مجموعه‌داده‌ی 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 طبق رابطه‌ی* ‏2‑3 *محاسبه می‌شود.*

|  |  |
| --- | --- |
| ‏2‑3 |  |

*که در آن r یک قانون، طول قانونِ r ، n تعداد صفات داده، برابر با مقدار بالابری*[[113]](#footnote-113)*قانون r و*

|  |  |
| --- | --- |
| ‏2‑4 |  |

*است که در آن یک تابع انتخاب است که اگر سمت چپ قانون r بتواند یک رکورد با مقدار جاافتاده را توصیف کند برابر با یک خواهد بود و درغیر این صورت صفر خواهد بود و*

|  |  |
| --- | --- |
| ‏2‑5 |  |

*است که بیانگر صفت j ام رکورد iام و مقدار صفت j ام رکورد iام است. قانونی با بیش‌ترین امتیاز به‌عنوان قانون برگزیده انتخاب‌شده و از مقدار سمت راست آن قانون به‌عنوان مقدار جایگزین صفت مقدار صفت جاافتاده استفاده می‌شود. از نقاط قوت این روش می‌توان به دخیل کردن معیار وابستگیِ دو طرف قوانین انجمنی و طول قوانین و حمایت از مقادیر جاافتاده‌ی چندگانه اشاره کرد.*

*در* [70] *قوانین یافت شده رتبه‌بندی می‌شوند و قانونی با بهترین رتبه به‌عنوان قانون انتخاب‌شده برای پر کردن مقدار جاافتاده استفاده خواهد شد. قوانین گزینش‌شده به ترتیب بر اساس طول قانون،* درصد اعتماد[[114]](#footnote-114) و درصد پشتیبانی[[115]](#footnote-115)*رتبه‌بندی می‌شوند. پس از یافتن قوانین یافت شده از مجموعه‌داده، برای هر مقدار جاافتاده مجموعه قوانینی که در سمت راست آن‌ها همان صفت جاافتاده حضور دارند و همچنین سمت چپ آن قوانین با باقی صفات رکورد مطابق است، انتخاب می‌شوند. سپس با استفاده از ترتیب ذکرشده قوانین انتخاب‌شده مرتب می‌شوند و بهترین قانون انتخاب می‌شود. سپس مقدار جاافتاده با استفاده از سمت راست بهترین قانون انتخاب‌شده جایگزین می‌شود.**از نقاط قوت این روش می‌توان به استفاده دخیل کردن طول قوانین و رتبه‌بندی قوانین اشاره کرد. در نظر گرفتن تنها بهترین قانون هنگام پر کردن مقادیر از نقاط ضعف این روش است***.**در [71] از روش توسعه‌یافته‌ی قوانین انجمنیِ مقاوم[[116]](#footnote-116) برای جایگذاری مقادیر جاافتاده در مجموعه‌داده‌هایی با مقادیر جاافتاده چندگانه استفاده‌شده است. محدودیت اصلی روش­های مبتنی بر قوانین انجمنی این است که تنها قابل‌استفاده برای داده‌های غیر عددی هستند.

### رویکردهای مبتنی بر خوشه‌بندی

مطالعات مختلفی از روش­های خوشه­بندی برای بهبود دقت فرآیند جایگذاری مقادیر جاافتاده استفاده کرده­اند. در روش جایگذاری با خوشه­بندی k-means ابتدا همه­ی نمونه­ها *با استفاده از خوشه‌بندی* k-means *به* k *خوشه‌ تقسیم می‌شوند.* پس از تشکیل خوشه­ها، از اطلاعات نمونه­های عضو خوشه­ای که نمونه­ی جاافتاده در آن قرار دارد استفاده خواهد شد تا صفات جاافتاده از نمونه­ی جاافتاده تخمین زده شوند. *کیفیت خوشه‌بندی بر این روش­ها بسیار تأثیرگذار است.*

در [72] از نسخه­ی فازی[[117]](#footnote-117) خوشه­بندی k-means برای جایگذاری مقادیر جاافتاده استفاده‌شده است. در خوشه­بندی فازی، هر نمونه یک تابع عضویت[[118]](#footnote-118) دارد که احتمال تعلق یک نمونه به یک خوشه را نشان می­دهد.

*در* [73] *با استفاده از یک خوشه‌بندی جدید به نام خوشه‌بندی همسایگان مشترک برای هدف پر کردن داده‌های جاافتاده استفاده‌شده است.* در این خوشه­بندی از تعداد همسایگان مشترک در بین k نزدیک­ترین همسایگان دو نمونه، برای محاسبه­ی فاصله­ی بین آن دو نمونه استفاده می­شود. *برای این کار از دو ماتریس برای m نمونه داده­ی ورودی استفاده می‌شود:*

*1) ماتریس فاصله:*

|  |  |
| --- | --- |
| ‏2‑6 |  |

*که در آن عنصر ، یعنی ، فاصله‌ی دو نمونه‌ی iام و j ام را نشان می‌دهد.*

*2) ماتریس همسایگی:*

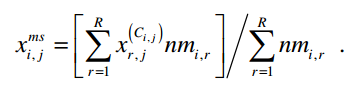
|  |  |
| --- | --- |
| ‏2‑7 |  |

*که در آن عنصر تعداد همسایه‌های مشترک در* k *نمونه‌ی نزدیک به رکوردهای i و j را نشان می‌دهد.*

*سپس برای هر مقدار جاافتاده یک لیست مرتب از همه‌ی همسایگان آن نمونه که صفت j ام آن‌ها معلوم است به نام ساخته می‌شود. این لیست بر اساس مقدار کلید رابطه‌ی* ‏2‑8 *مرتب‌شده است:*

|  |  |
| --- | --- |
| ‏2‑8 |  |

*که در آن t تعداد کل نمونه‌ها است.*

*سپس برای هریک از مقادیر جاافتاده‌ی یک خوشه مثل شامل R عنصرِ اول ساخته می­شود. در انتها، مقدار جاافتاده توسط مقادیر نمونه‌های موجود در آن خوشه بسته به نوع عددی یا نمادینِ صفت جاافتاده، تخمین زده می­شوند. نحوه توزیع داده‌ها بر تعداد همسایگان مشترک (ممکن است ماتریس همسایگان مشترک* تُنُک[[119]](#footnote-119) *شود) در این روش تاثیرگذار است.*

### رویکردهای ترکیبی

روش­های ترکیبی نیز جایگاه خاصی برای پر کردن مقادیر جاافتاده دارند. در [74] یک رویکرد دومرحله‌ای برای پر کردن مقادیر جاافتاده ارائه‌شده است. این رویکرد شامل یک تخمین اولیه از مقادیر جاافتاده توسط نزدیک‌ترین مرکز خوشه در روش خوشه‌بندی k-means و یک گام پالایش[[120]](#footnote-120) و بهبود مقدار تخمین زده‌شده از گام قبلی، توسط شبکه‌ی MLP است. در [75] در گام اول به‌جای استفاده از خوشه‌بندی k-means معمول، از خوشه‌بندی فازی c-means استفاده‌شده است. نیاز به آموزش یک شبکه‌ی MLP برای هر مقدار جاافتاده و آموزش مجدد روی کل مجموعه آموزشی، تعیین ساختار (به­هم­بندی[[121]](#footnote-121) و وزن‌ها) بهینه‌ی شبکه‌ی MLP برای هر بار آموزش برای پر کردن هر مقدار جاافتاده از نقاط ضعف این روش است.

*در* [76] *یک رویکرد ترکیبی از قوانین انجمنی و* k *نزدیک‌ترین همسایگان برای پر کردن مقادیر جاافتاده ارائه‌شده است. ابتدا قوانین حاکم بر مجموعه‌داده استخراج می‌شوند و سپس برای هر مقدار جاافتاده مجموعه قوانینی که سمت چپ آن قوانین با صفات رکورد مطابق است، انتخاب می‌شوند. اگر مجموعه‌ی قوانین انتخاب‌شده تهی باشد (هیچ قانونی نتواند با رکورد دارای مقدار جاافتاده تطبیق یابد)، به روش نزدیک‌ترین همسایه آن مقدار جاافتاده پر خواهد شد و در غیر این ‌صورت با استفاده از مجموعه قوانین انتخاب‌شده جایگزینی انجام خواهد شد. برای صفات نمادین مقداری که بیش‌ترین تکرار را در سمت راست قوانین انتخاب‌شده دارد، انتخاب می‌شود و برای صفات پیوسته مقدار میانه برای پر کردن مقادیر جاافتاده انتخاب می‌شود. در این روش فرض شده است که تعداد قوانین تولیدشده کم است.* از نقاط قوت این روش می‌توان به بهره­گیری از قدرت ترکیب دو روش متفاوت اشاره کرد. درواقع، در این روش سعی در استفاده از ارتباطات و فواصل موجود برای پر کردن مقادیر جاافتاده شده است. این روش قابلیت استفاده برای داده‌های عددی را به‌طور مستقیم ندارد (به دلیل استفاده از قوانین ‌انجمنی).

[55] روش­های k-means و نزدیک­ترین همسایگان را برای پر کردن مقادیر جاافتاده درون یک مجموعه‌داده ترکیب کرده ­است. در [77] ابتدا داده‌ها خوشه‌بندی شده و سپس هر مقدار جاافتاده با میانگین دو مقدار جایگزین می‌شود. مقدار اول فاصله‌ی نمونه‌ی جاافتاده تا مرکز خوشه‌ای است که نمونه‌ی جاافتاده به آن تعلق دارد. مقدار دوم برابر با مقدار صفتی از مرکز خوشه است که متناظرا در نمونه‌ی جاافتاده، نامعلوم است.

[78] روش­های برنامه­نویسی پویا[[122]](#footnote-122)، شبکه­های عصبی و الگوریتم ژنتیک[[123]](#footnote-123) را برای هدف جایگذاری مقادیر جاافتاده ترکیب کرده ­است. [79] روش­های c-means فازی، رگرسیون بردار پشتیبان[[124]](#footnote-124)(SVR) و الگوریتم ژنتیک را ترکیب کرده است.

### دیگر روش‌ها

روش­های بسیار دیگری مانند الگوریتم­های ژنتیک، شبکه­های بِیزی[[125]](#footnote-125) [80,81,82] ،k درخت تصمیم [83] و افراز بندی بازگشتی[[126]](#footnote-126) [84] برای هدف پر کردن مقادیر جاافتاده مطرح‌شده‌اند. در [24] و [30] به کمک جمع‌سپاری[[127]](#footnote-127) مقادیر جاافتاده پر می‌شوند. مشکل اصلی این روش­های مبتنی بر جمع­سپاری، به‌طور کامل‌ خودکار نبودن و تأثیرگذاری عامل انسانی در دقت نهایی داده‌های پرشده است.

تکنیک­های جایگذاری مقادیر جاافتاده را می‌توان بر اساس دفعات جایگذاری[[128]](#footnote-128) به دو دسته­ی اصلی تقسیم کرد: جایگذاری تکی[[129]](#footnote-129) و جایگذاری چندگانه[[130]](#footnote-130) [68]. در جایگذاری تکی برای هر مقدار جاافتاده فقط یک تک مقدار تخمین­ زده می­شود و همان مقدار جایگزین خواهد شد. جایگذاری چندگانه یک رویکرد کلی پیشنهادی توسط [85] و [86] است. ایده­ی اصلی در این رویکرد این است که توسط چندین مدلِ مختلفِ جایگذاری، مجموعه­داده‌های کامل مختلفی را برای مجموعه­داده­ی جاافتاده تولید کرده، سپس نتایج مختلف را به‌عنوان نتیجه نهایی ترکیب کنیم. این حالت منجر به تولید نتایجی با انحراف کمتر و دقت بیشتر خواهد شد. رویکرد جایگذاری چندگانه تقریبا با هر روش جایگذاری قابل‌اجرا است [37]. [85] نشان داد که جایگذاری چندگانه بهتر از روش جایگذاری با مقدار میانگین عمل می­کند.

[27] تأثیر روش­های مختلف جایگذاری مقادیر جاافتاده را بر عملکرد دسته­بندهای[[131]](#footnote-131) مختلف بررسی کرده است و عملکرد یک روش جایگذاری چندگانه و پنج روش جایگذاری تکی را بر روی شش روش مختلف دسته­بندی ارزیابی نموده است.

بنا به [87]، تک روشی برای جایگذاری مقادیر جاافتاده وجود ندارد که به تمام دیگر روش­ها برتری داشته باشد. علت این امر این است که عملکرد تمام روش­های موجود تحت تأثیر ویژگی­های مجموعه­داده‌های آزمایش‌شده و نوع مقادیر جاافتاده قرار می­گیرد.

[88] درصد جاافتادگی، نحوه توزیع مقادیر جاافتاده و ویژگی­های مجموعه‌داده‌ها (مثل تعداد نمونه­ها و تعداد صفات) را به‌عنوان عوامل مؤثر در کیفیت عملکرد رویکردهای جایگذاری مقادیر جاافتاده ذکر کرده است.

## ایده‌‌ی کلی پیشنهادی

*در اکثر روش­های موجود جایگذاری مقادیر جاافتاده، رویکرد اعمال‌شده بر کل مجموعه‌داده اعمال می­شود. عملکرد بعضی از روش­ها مثل رگرسیون و روش حداکثرسازی امید در داده­هایی که روابط قویِ بین صفات در آن­ها حاکم است، بهتر خواهد بود. بنابراین مشکلی که در این روش­ها وجود دارد، این است که ممکن است همبستگی بین دو یا چند صفت بر روی کل مجموعه‌داده برقرار نباشد و این روش­ها ضعیف عمل کنند. این مشکل را می‌توان با افراز بندی مجموعه‌داده‌ها و اعمال مدل­­های جایگذاری مناسب در هر افراز تا حدی از بین برد. برای مثال مجموعه­داده­ی رابطه­ای نمایش داده‌شده در* شکل ‏2‑1 *را نظر بگیرید. همان­طور که در شکل سمت چپ پیداست، اگرچه رابطه‌ی بین صفت قد و سن افراد تا یک سن خاص متناسب با یکدیگر است، ولی از یک سن به بعد الگوی رفتاری روابط بین این دو صفت متفاوت می­شود. شکل سمت راست نشان می­دهد که با یک افراز بندی مناسب می‌توان رفتار تغییر در مقدار دو صفت ذکرشده را به دو نوع مختلف تقسیم کرد و بنابراین همبستگی­های قوی­تری بین دو صفت در هریک از افرازها یافت. درواقع، افرازهای مختلف یک مجموعه‌داده، الگوهای رفتاری متفاوتی را در بین صفات خود ارائه داده­اند. بر همین اساس، ایده­ی پایه برای رویکردهایی که در فصل­های بعدی ارائه می­شود، افراز بندی مناسب داده­های درون یک مجموعه‌داده خواهد بود.*

*(فکر کنم قبلا گفته بودم تو فصل 1 بیاریدش و شما گفتید میخواهید بعد از خوندن کارهای پیشین بیاد. اگه نظرتون اینه شاید بد نباشه به جمع بندی این فصل اضافه بشه. هرچند که کمی طولانی میشه)*



شکل ‏2‑1 یک مثال از این‌که وابستگی قوی بین دو صفت روی کل مجموعه‌داده برقرار نیست (جدول سمت چپ) ولی روی افراز‌هایی از آن برقرار است (جدول سمت راست) [18].

## معیار­های ارزیابی برای سنجش کیفیت جایگذاری مقادیر جاافتاده

معیارهای مختلفی برای ارزیابی کیفیت روش­های مختلف پر کردن مقادیر جاافتاده وجود دارد. سه معیار شناخته‌شده برای این هدف عبارت­اند از: ریشه میانگین مربعات خطا[[132]](#footnote-132) (RMSE)، ضریب تعیین[[133]](#footnote-133) (CoD) یا و میانگین خطای مطلق[[134]](#footnote-134) (MAE).

معروف­ترین شاخص عملکرد برای سنجیدن کیفیت تخمین مقادیر جاافتاده برای متغیرهای عددی شاخصِ ریشه میانگین مربعات خطا (RMSE) است. این شاخص نشان‌دهنده‌ی میانگین اختلافات مقادیر واقعی از مقادیر تخمین زده‌شده توسط مدل­های جایگذاریِ مقادیر جاافتاده است [20]. ضریب تعیین یا در تحلیل­ مدل­های آماری برای ارزیابی این­که یک مدل چقدر خوب خروجی­های آینده را پیش­بینی می­کند، استفاده می­شود. ضریب تعیین مقداری بین 0 و 1 دارد. میانگین خطای مطلق (MAE) برابر با میانگین اندازه­ی خطاهای مجموعه­ای از پیش­بینی­ها بدون در نظر گرفتن جهت آن­ها است [64]. مقادیر کمتر برای میانگینِ خطای مطلق و ریشه میانگین مربعات خطا و مقادیر بیش­تر برای ضریب تعیین، بیانگر کیفیت بالاتر از تخمین مقادیر است.

معادلات ‏2‑9، ‏2‑10 و ‏2‑11 به ترتیب نحوه­ی محاسبه­ی معیارهای ریشه میانگین مربعات خطا، ضریب تعیین و میانگین خطای مطلق را نشان می­دهند.

|  |  |
| --- | --- |
| ‏2‑9 |  |
| ‏2‑10 |  |
| ‏2‑11 |  |

که در آن­ها نشان­دهنده مقدار واقعی، مقدار پیش‌بینی‌شده، میانگین مقادیر مشاهده‌شده و برابر با تعداد مقادیر تخمین زده‌شده است.

## خلاصه و نتیجه‌گیری

در این فصل مروری بر رویکردهای مختلف جایگذاری مقادیر جاافتاده انجام گرفت. شکل ‏2‑2 یک دسته­بندی کلی از رویکردهای مطالعه شده برای هدف جایگذاری مقادیر جاافتاده را نشان می­دهد. همچنین معیارهای ارزیابی رایج نیز برای سنجش کیفیت رویکردهای مختلف جایگذاری مقادیر جاافتاده معرفی­ شدند. در فصل­های بعدی، رویکردهای جدید پیشنهادی مطرح خواهند شد.



شکل ‏2‑2 دسته‌بندی کلی رویکردهای تخمین مقادیر جاافتاده

یکی از پارامتر­هایی که می‌توان جهت مقایسه راهکار­ها به­کارگرفت، قابلیت پر کردن مقادیر جاافتاده‌ی تکی یا چندگانه است. مزایا، چالش‌ها و مشکلات هریک از رویکردهای مطرح‌شده به‌طور خلاصه در جدول (1) ارائه‌شده است.

**جدول (1)- دسته‌بندی راه‌کارهای پر کردن مقادیر جاافتاده**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| روش | سال | پر کردن مقادیر چندگانه | نقاط قوت | مشکلات و چالش‌ها |
| *پر کردن با میانگین مقادیر* |  | 🗸 | * سادگی و قابل‌فهم بودن پیاده‌سازی رویکرد ارائه‌شده (نه لزوماً پیچیدگی محاسباتی) | * حساس به داده‌های پرت |
| *خوشه‌بندی k-means* |  | 🗸 | * سادگی و قابل‌فهم بودن پیاده‌سازی رویکرد ارائه‌شده (نه لزوماً پیچیدگی محاسباتی) | * تأثیرگذار بودن کیفیت خوشه‌بندی * تعیین تعداد خوشه، مراکز اولیه خوشه، معیار شباهت |
| *خوشه‌بندی همسایگان مشترک [9]* | 2009 | 🗸 | * استفاده از معیار شباهت برای صفات ترکیبی * دخیل کردن معیار جدیدی برای سنجش شباهت نمونه‌ها | * تأثیرگذار بودن نحوه توزیع داده‌ها بر تعداد همسایگان مشترک |
| *قوانین انجمنی امتیازدار [10]* | 2004 | 🗸 | * دخیل کردن معیار وابستگی دو طرف قوانین انجمنی و طول قوانین | * تعیین پارامترهای قوانین انجمنی |
| *قوانین انجمنی و k نزدیک‌ترین همسایگان [16]* | 2009 | 🗸 | * بهره‌گیری از قدرت ترکیب دو روش متفاوت | * ترکیب روش مناسب برای صفات عددی با روش مناسب صفات نمادین * چالش‌های KNN + چالش‌های قوانین انجمنی |
| *قوانین انجمنی رتبه‌بندی شده [11]* | 2007 | 🗸 | * دخیل کردن طول قوانین * رتبه‌بندی قوانین | * در نظر گرفتن تنها بهترین قانون هنگام پر کردن مقادیر * چالش‌های قوانین انجمنی |
| *k نزدیک‌ترین همسایگان [12]* | 2014 | 🗴 | * سادگی و قابل‌فهم بودن پیاده‌سازی رویکرد ارائه‌شده (نه لزوماً پیچیدگی محاسباتی) * قدرت تعمیم بالا | * لزوماً بررسی فقط نزدیک‌ترین نمونه بهترین نیست * عدم پر کردن مقادیر جاافتاده‌ی چندگانه |
| *(fuzzy c-means)k-means+MLP*  *[14] و [15]* | 2014  2015 | 🗸 | * بهبود k-means تنها | * زمان‌بر بودن یادگیری هر شبکه عصبی برای هر مقدار جاافتاده * تعداد زیاد پارامترهای تنظیم شبکه عصبی * پارامترهای خوشه‌بندی |
| *DMI [5]* | 2011 | 🗸 | * توجه به این‌که وابستگی بین صفات لزوماً در کل مجموعه‌داده برقرار نیست | * مناسب بودن فقط برای مجموعه‌داده‌هایی که وابستگی بین صفات آن‌ها وجود دارد * تعیین پارامترهای ساختن درخت تصمیم |

# فصل سوم: استفاده از خوشه‌بندی و رویکردی ترکیبی برای جایگذاری مقادیر جاافتاده عددی

## مقدمه

در این فصل ‌یک رویکرد دومرحله‌ای مبتنی بر خوشه‌بندی داده‌ها و استفاده از ‌ترکیب روش k نزدیک‌ترین همسایگان وزن‌دار[[135]](#footnote-135) (Wk-NN) و رگرسیون خطی[[136]](#footnote-136) برای پرکردن مقادیر جاافتاده عددی ارائه می‌‌شود. ویژگی اصلی روش ارائه‌شده استفاده‌ی ترکیبی از همبستگی[[137]](#footnote-137) موجود بین داده‌ها و فاصله‌ی بین نمونه‌ها برای 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 فصل 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 مجموعه‌داده اصلی یافت شوند. دلیل این امر این است که شاید روابط قوی بر روی کل مجموعه‌داده یافت نشود، درحالی‌که بین زیرمجموعه‌ای از نمونه‌ها، ارتباطات قوی‌تری وجود داشته باشد. در صورت یافت نشدن میزان ارتباط کافی درون زیر مجموعه‌داده‌ها، از روش پایه نزدیک‌ترین همسایگان وزن‌دار استفاده می‌شود. در [18] اگرچه با استفاده از درخت تصمیم سعی در افزایش همبستگی نمونه­ها دارد، بااین‌حال از معیاری دقیق برای سنجش صحت فرضیه افزایش همبستگی درون هر افراز از نمونه­ها استفاده نمی­کند. روش­های مبتنی بر رگرسیون به‌تنهایی این نکته که ممکن است ارتباط بین صفات در یک مجموعه­داده قوی نباشند را در نظر نمی­گیرند. بنابراین دقت عملکرد آن­ها در این شرایط به‌شدت کاهش می­یابد. در روش­های مبتنی بر k نزدیک­ترین همسایه، اطلاعات راجع به همبستگی بین صفات به‌طورکلی نادیده گرفته می­شوند. درروش ترکیبی­ ارائه‌شده سعی شده ­است تا تمام این موارد به‌طور توأم برای بهبود کیفیت تخمین در نظر گرفته شوند.

در مرحله‌ی اول، داده‌ها خوشه‌بندی می‌شوند. در مرحله‌ی دوم، داده‌های جاافتاده درون هر خوشه با استفاده از یک روش ‌ترکیبی از k نزدیک‌ترین همسایگان وزن‌دار و رگرسیون خطی تخمین زده می­شوند. از معیار همبستگی بین صفات در هر خوشه برای تعیین روش پرکردن داده‌های جاافتاده استفاده می‌‌شود. ویژگی رو­ش ارائه‌شده استفاده‌ی ترکیبی از همبستگی موجود بین داده‌ها و فاصله‌ی بین نمونه‌ها برای پرکردن مقادیر جاافتاده است.

کیفیت پرکردن مقادیر جاافتاده با استفاده از معیار میانگین مربعات خطا سنجیده می‌شود. همچنین، تأثیر پارامترهای مختلف بر کیفیت میزان خطای داده‌های تخمین زده‌شده بررسی می­شود. عملکرد روش ارائه‌شده برای تخمین داده‌های جاافتاده بر روی پنج مجموعه‌داده نیز‌ بررسی می­شود. درنهایت، عملکرد روش ارائه‌شده با چهار روش پرکردن با مقدار میانگین، روش تخمین با شبکه عصبی پرسپترون چندلایه (MLP) [25] ، روش پرکردن با خوشه­بندی فازی c-means (FCM) [89] و روش k خوشه‌ و نزدیک‌ترین همسایه‌ی مبتنی بر دسته (CKNNI) [56] مقایسه می‌شود. نتایج به‌دست‌آمده نشان داده است که خطای تخمینِ مقادیر جاافتاده در روش ارائه‌شده کمتر از خطا در مقایسه با دیگر روش‌های مقایسه شده است.

## رویکرد پیشنهادی

در این بخش بیانی دقیق و رسمی از مراحل رویکرد پیشنهادی در این فصل ارائه می‌شود. 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 مجموعه­داده وجود دارد و بنابراین تلاش می­کنند تا روابط موجود را مدل­سازی کنند. ازجمله این روش­ها می­توان به روش­های آماری مانند رگرسیون اشاره کرد. مشکل اصلی استفاده از این روش­ها به‌تنهایی این است که ممکن است در داده­های ورودی روابط قوی وجود نداشته­ باشد. بنابراین مدل ساخته‌شده دقت پایینی را ارائه خواهد داد. به‌علاوه، در اکثر روش­های ارائه‌شده مبتنی بر روابط قوی، از معیار مشخص و دقیقی برای سنجش روابط بین صفات، مانند ضریب همبستگی، استفاده‌نشده ­است. دسته­ی 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 فصل ترکیبی از دو دسته روش ذکرشده به‌منظور بهره­وری از مزایای آن­ها ارائه می­شود. با توجه به مطالعات انجام‌شده، پیش‌ازاین، ترکیبی از این دو دسته روش ارائه نشده است. درروش 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 مجموعه­داده ورودی قوی­تر از کل مجموعه­داده باشد. در کارهای پیشین از خوشه­بندی برای گروه­بندی نمونه­های نزدیک به هم استفاده‌شده است. در این پژوهش از خوشه­بندی برای یافتن نمونه­هایی با همبستگی بالا استفاده‌شده 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 پیش‌ازاین برای این منظور استفاده‌نشده است. درنهایت، بسته به این­که آیا روابط به حد کافی قوی هستند یا خیر، از یکی از دو دسته روش بالا استفاده می­شود.

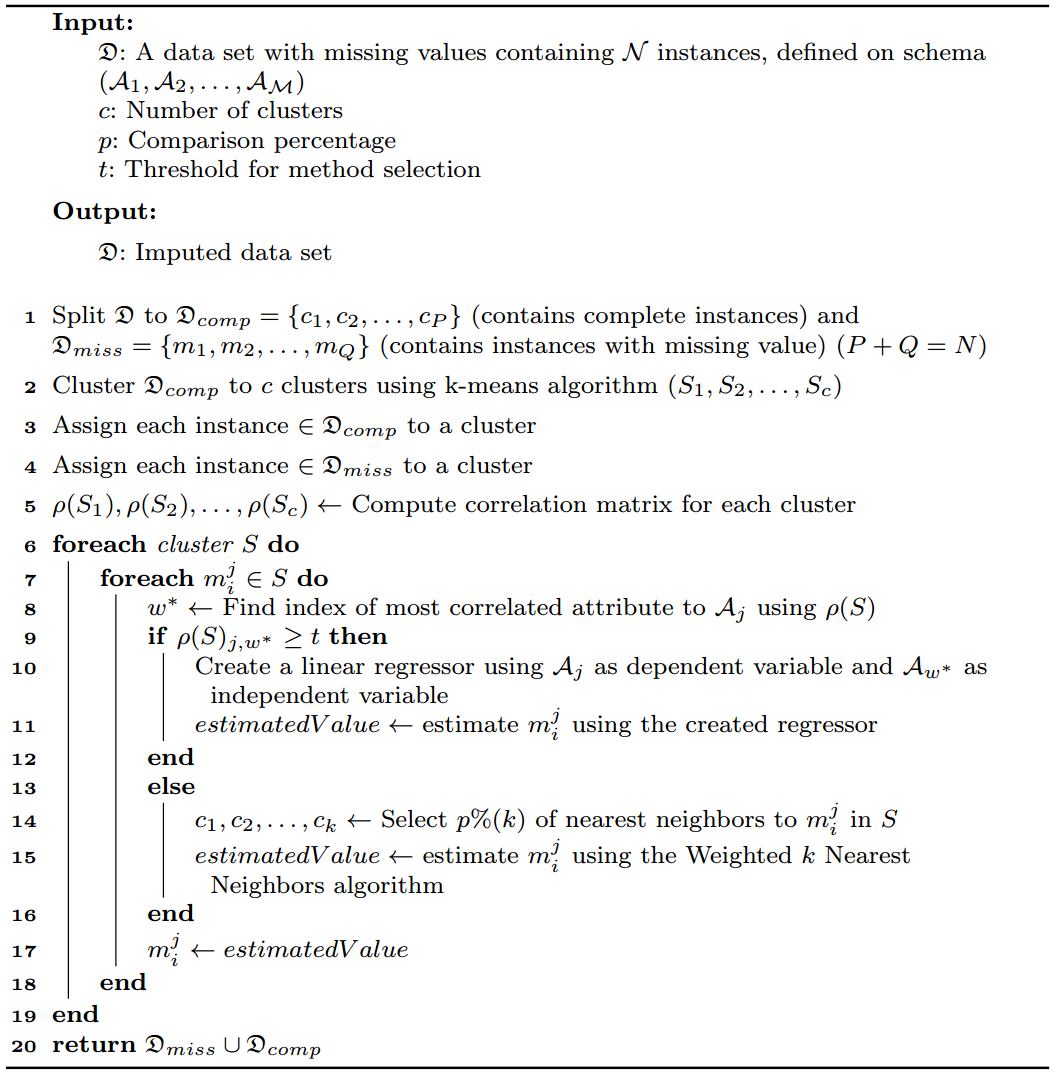
روش ارائه‌شده برای پرکردن داده‌های جاافتاده شامل دو گام اصلی است. در گام اول، داده‌های کامل (بدون مقادیر جاافتاده) خوشه‌بندی می‌‌شوند. هدف اصلی از خوشه‌بندی این است که ممکن است همبستگی قوی بر روی کل مجموعه‌داده برقرار نباشد، درحالی‌که بین زیرمجموعه‌ای از نمونه­ها، ارتباطات قوی وجود داشته باشد. بنابراین سعی می‌شود با استفاده از خوشه‌بندی، همبستگی‌های احتمالی موجود در زیرمجموعه‌های مجموعه‌داده اصلی را پیدا کرد. در گام بعدی مقادیر جاافتاده در هر خوشه با استفاده از روشی مرکب از k نزدیک‌ترین همسایگان وزن‌دار و رگرسیون خطی پر خواهند شد. از یک حد آستانه () بر روی مقادیر ماتریس همبستگیِ 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 همسایگان وزن‌دار استفاده می­شود. درواقع، ایده‌ی 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 فصل با روش خوشه‌بندی k-means انجام می‌‌شود. شکل ‏3‑1 روند کلی رویکرد پیشنهادی را نشان می­دهد. شبه کد روش پیشنهادی در شکل ‏3‑2 نمایش داده‌شده است. در ادامه‌ جزئیات مراحل مختلف رویکرد پیشنهادی شرح داده می‌شود.

فرض کنید مجموعه‌داده رابطه‌ای شامل رکورد بر روی شِمای تعریف‌شده است. هر نمونه شامل صفت است. مقدار صفت ام نمونه‌ی ام را نمایش می‌‌دهد. تابع بیانگر فاصله‌ی دو نمونه‌ی و است.

در ابتدا مجموعه‌داده اصلی () به دو مجموعه‌داده کامل () (رکوردهایی بدون مقادیر جاافتاده) و مجموعه‌داده‌های با مقادیر جاافتاده () تقسیم می‌‌شود (خط 1 از شبه کد). سپس، مجموعه‌ی با استفاده از خوشه‌بندی k-means، به تا خوشه‌ی تقسیم می‌‌شود و بنابراین مرکز خوشه به دست می‌‌آید (خط 2). هریک از عناصر با توجه به میزان نزدیکی به مراکز خوشه به ‌یک خوشه نسبت داده می‌‌شوند (خط 3). هریک از عناصر نیز با توجه به میزان نزدیکی به مراکز خوشه به‌ یک خوشه منتسب می‌‌شود (خط 4). قابل‌ذکر است که فاصله‌ی هر نمونه‌ی جاافتاده بدون در نظر گرفتن صفات جاافتاده محاسبه می‌‌شود. سپس ضریب همبستگی پیرسون[[138]](#footnote-138) برای جفت صفات که است، برای نمونه‌های هر خوشه محاسبه می‌شود و نتایج در یک ماتریس همبستگی برای آن خوشه ذخیره می‌شود (خط 5).



شکل ‏3‑1 روند کلی رویکرد پیشنهادی برای پرکردن مقادیر جاافتاده با استفاده از خوشه‌بندی و رویکردی ترکیبی



شکل ‏3‑2 شبه کدِ رویکرد پیشنهادی برای پرکردن مقادیر جاافتاده با استفاده از خوشه‌بندی و رویکردی ترکیبی

برای هر خوشه‌ مانند، ماتریس همبستگی به‌صورت ‏2‑11 خواهد بود.

|  |  |
| --- | --- |
| ‏3‑1 |  |

که در آن هر عنصر از ‏3‑2 محاسبه می‌شود.

|  |  |
| --- | --- |
| ‏3‑2 |  |

که در آن برابر با کوواریانس بین مقادیر صفت و صفت است و و به ترتیب برابر با مقادیر واریانس صفت و صفت هستند. نماد 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؛ بنابراین .

حال برای هر مقدار جاافتاده مثل اندیس همبسته‌ترین[[139]](#footnote-139) صفت به صفتِ جاافتاده طبق ‏3‑3 از طریق ماتریس همبستگیِ افراز مربوطه پیدا می‌شود (خط 8).

|  |  |
| --- | --- |
| ‏3‑3 |  |

اگر باشد، که آن t حد آستانه‌ی از پیش تعریف‌شده است، آنگاه‌ یک رگرسور خطی با صفت مستقلِ و صفت وابسته به‌صورت ‏3‑4ساخته می‌شود (خط 10).

|  |  |
| --- | --- |
| ‏3‑4 |  |

که در آن و ضرایب رگرسور خطی هستند. در این حالت مقدار جاافتاده از ‏3‑5 محاسبه می‌شود (خط 11). این ضرایب از طریق برازش[[140]](#footnote-140) بهترین خط بر روی داده­ها به کمک روش حداقل مربعات[[141]](#footnote-141) محاسبه می­شوند.

|  |  |
| --- | --- |
| ‏3‑5 |  |

اگر باشد،آنگاه مقدار جاافتاده با استفاده از روش k نزدیک‌ترین همسایگان وزن‌دار محاسبه می‌شود. در این حالت، فاصله‌ی نمونه جاافتاده با تمام نمونه‌های کامل موجود در خوشه *‌* محاسبه ‌می‌‌شود و این فواصل به ترتیب صعودی مرتب می‌‌شوند (یعنی، که در آن اندازه‌ی خوشه *‌* است). آنگاه درصد از نمونه‌های کامل در خوشه­ی نمونه‌ی جاافتاده‌ به‌عنوان k نزدیک‌ترین همسایگان کامل به نمونه‌ی جاافتاده انتخاب می‌‌شوند () و درنهایت مقدار جاافتاده از ‏3‑6 محاسبه می‌شود (خطوط 14 و 15).

|  |  |
| --- | --- |
| ‏3‑6 |  |

که در آن از ‏3‑7 به دست می‌آید.

|  |  |
| --- | --- |
| ‏3‑7 |  |

درواقع ‏3‑6 رابطه‌ی میانگین وزن‌دار است که وزن‌ها در آن معکوس فاصله‌ی دو نمونه هستند. درنهایت مقدار محاسبه‌شده‌ی به‌جای مقدار جاافتاده جایگزین می‌شود و پس از پرکردن تمام مقادیر جاافتاده مجموعه‌داده کامل بدون مقادیر جاافتاده به دست می‌آید (خط 18).

## تحلیل پیچیدگی زمانی رویکرد پیشنهادی

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

فرض کنید مجموعه­داده ورودی شامل نمونه است. از این نمونه P تای آن­ها کامل و Q تای آن­ها شامل مقادیر جاافتاده است (P+Q=). همچنین فرض کنید هر نمونه شامل ویژگی است. روش مطرح‌شده شامل سه گام اصلی است. در گام اول یک خوشه­بندی k-means فقط بر روی داده­های کامل انجام می­شود. مرتبه زمانی این خوشه­بندی برابر با است که در آن تعداد خوشه­ها و برابر با حداکثر تعداد تکرار در الگوریتم k-means است. در مرحله دوم به تعداد خوشه­ها، ماتریس همبستگی محاسبه می­شود. هر ماتریس شامل عنصر است. مرتبه زمانی محاسبه هر عنصر در بدترین حالت برابر با است. از طرفی به تعداد خوشه­ها، یعنی بار، نیاز به محاسبه ماتریس همبستگی است. بنابراین هزینه کلی این گام در بدترین حالت از مرتبه­ی خواهد بود. آخرین گام، تخمین مقادیر جاافتاده است. در این مرحله به تعداد مقادیر جاافتاده (یعنی) یکی از روش­های رگرسیون خطی با پیچیدگی و یا k نزدیک­ترین همسایگان با پیچیدگی انجام می­شود. بنابراین پیچیدگی مرحله­ی دوم در بدترین حالت (حالتی که تمام تخمین­ها از طریق رگرسیون انجام شوند) برابر با خواهد بود. با توجه به این­که در بدترین حالت است، با در نظر گرفتن موارد گفته‌شده، پیچیدگی الگوریتم پیشنهادی در بدترین حالت از مرتبه­ی خواهد بود. اگر فرض کنیم که در بدترین حالت ، آنگاه پیچیدگی رویکرد پیشنهادی برابر با است.

## آزمایش‌ها و نتایج

در این بخش بررسی‌های جامعی بر روی تأثیر مقادیر مختلف الگوریتم ارائه‌شده ارائه می‌شود. همچنین دقت رویکرد پیشنهادی با چهار روش دیگر مقایسه می‌شود. از پنج مجموعه‌داده موجود در پایگاه داده‌ی UCI برای آزمایش‌ها استفاده می‌شود. جدول ‏3‑1 مشخصات این مجموعه‌داده‌ها را نمایش می‌‌دهد. این مجموعه‌داده‌ها بدون مقادیر جاافتاده می‌باشند. تمام پیاده‌سازی‌های‌ لازم با استفاده از زبان برنامه‌نویسی پایتون[[142]](#footnote-142) انجام‌شده است.

جدول ‏3‑1 مشخصات مجموعه‌داده‌های استفاده‌شده

|  |  |  |
| --- | --- | --- |
| **مجموعه­داده** | **تعداد نمونه** | **تعداد ویژگی­** |
| Iris | 150 | 4 |
| Wine | 178 | 13 |
| Glass | 214 | 10 |
| Haberman | 306 | 3 |
| Wholesale customers | 440 | 8 |

### مشخصات آزمایش‌ها

در این زیر بخش جزئیات آزمایش‌های انجام‌شده بیان می‌شود. برای بررسی تأثیر متغیرهای رویکرد پیشنهادی، ابتدا به تعداد 10 درصدِ کل نمونه‌های هر مجموعه‌داده، مقادیر جاافتاده به‌طور تصادفی در آن مجموعه‌داده‌ ایجاد می‌شود. سپس روش­های مختلف بر روی آن مجموعه‌داده اعمال می­شوند.

از معیار فاصله اقلیدسی برای سنجش فاصله‌ی دو نمونه استفاده‌شده است. مقدار این فاصله بین دو نمونه‌ی و از ‏3‑8 محاسبه می‌شود.

|  |  |
| --- | --- |
| ‏3‑8 |  |

که در آن تعداد صفات هر نمونه است.

پیش‌پردازش نرمال‌سازی داده‌ها به بازه‌ی 0 تا 1 به روش کمینه-بیشینه[[143]](#footnote-143) [1] با استفاده از ‏3‑9 برای هر مقدار در مجموعه‌داده صورت گرفته است.

|  |  |
| --- | --- |
| ‏3‑9 |  |

که در آن و به ترتیب مقادیر کمینه و بیشینه برای صفت ام هستند. نرمال‌سازی از غلبه‌ی صفاتی با بازه‌ی مقادیر کوچک بر صفاتی با بازه‌ی مقادیر بزرگ در محاسبه‌ی فاصله جلوگیری می‌کند.

هریک از آزمایش‌های انجام‌شده در ادامه،‌ 5 بار به‌صورت تصادفی انجام‌شده است و میانگین خطای این 5 بار به‌عنوان نتیجه‌ی نهایی، گزارش‌شده است. برای محاسبه‌ی خطا از معیار میانگین مربعات خطا مطابق معادله­ی ‏2‑9 استفاده می‌شود.

از خوشه‌بندی k-means برای خوشه‌بندی داده‌ها در این پژوهش استفاده‌شده است. اگرچه این روش یکی از قدیمی‌ترین روش‌های خوشه‌بندی است، بااین‌حال هنوز هم یکی از پرکاربردترین روش­های خوشه‌بندی است [90]. حداکثر تعداد 100 تکرار برای روش خوشه‌بندی k-means تنظیم‌شده است. در ادامه تأثیر متغیرهای مختلف رویکرد ارائه‌شده ‌یعنی تعداد خوشه‌ها ()، درصد انتخاب نزدیک‌ترین همسایه‌ها () و حد آستانه برای تعیین روش تخمین () بررسی می‌گردند.

### بررسی تأثیر تعداد خوشه‌ها (c) بر کیفیت تخمین داده‌های جاافتاده

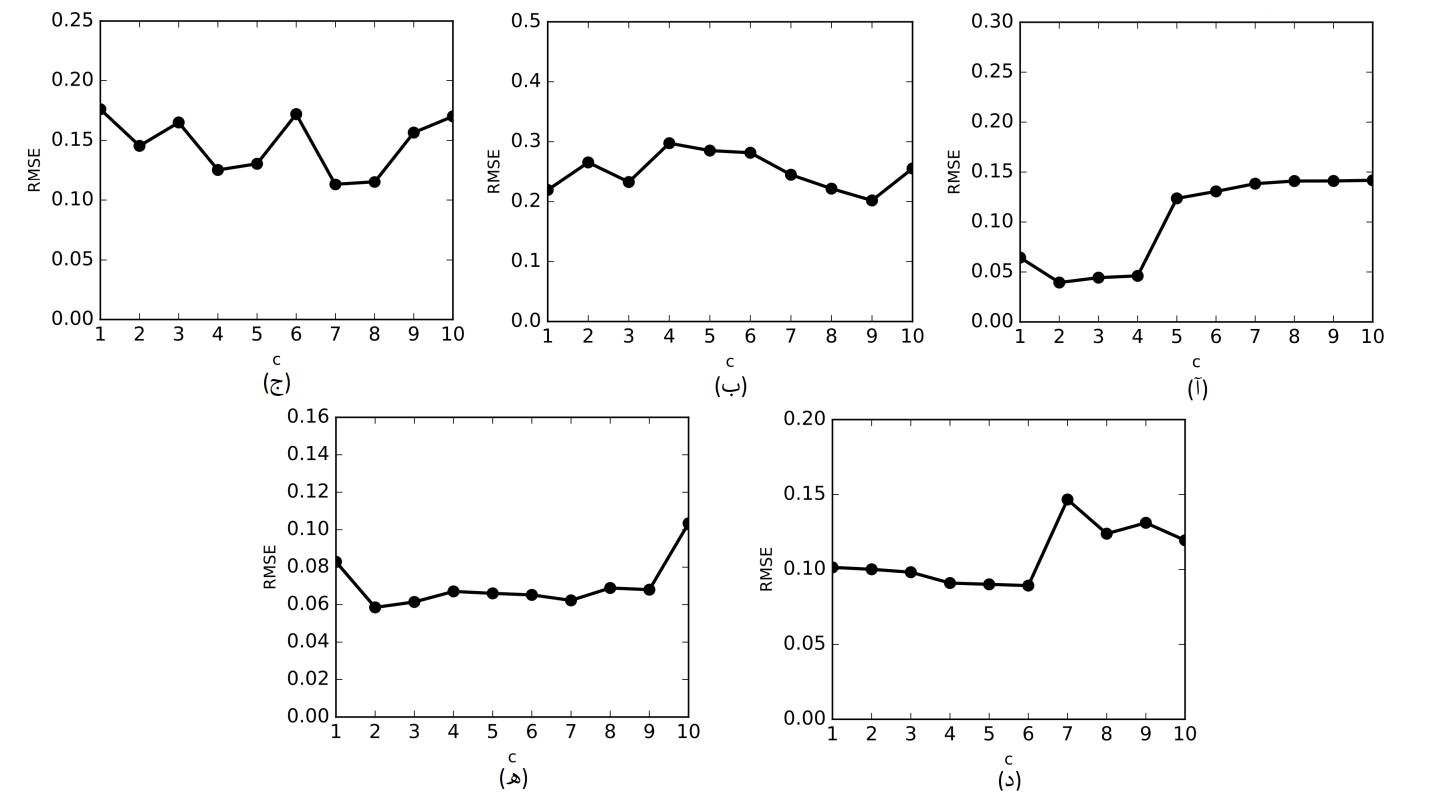
در این قسمت نتایج آزمایش‌هایی برای بررسی تأثیر مقدار c در میزان خطا بررسی می‌شود. ابتدا به تعداد 10% کل نمونه‌ها، مقدار جاافتاده به‌صورت تصادفی در هر مجموعه‌داده ایجادشده است. سپس مقدار c بین 1 تا 10 برای مقادیر ثابت p و t، تغییر می‌یابند و نتایج ثبت می‌شوند. شکل ‏3‑3 مقدار خطای RMSE را برای مقادیر مختلف c در مجموعه‌داده‌های مختلف نمایش می­دهد. همان‌طور که از نمودارها پیداست، برای مجموعه‌داده‌های مختلف خوشه‌بندی داده‌ها مقدار خطای تخمین را کاهش داده است. درواقع مقدار خطا در یک مقدار بهینه‌شده است. بنابراین فرضیه­ی کاهش خطای تخمین با خوشه‌بندی داده‌ها با توجه به نمودارها قابل‌درک است. مقدار c بهینه برای مجموعه‌داده‌های مختلف، متفاوت است. آزمایش‌ها نشان می‌دهند، افزایش مقدار c از یک حد به بعد، به‌تدریج باعث افزایش مقدار خطای تخمین می‌شود.

### بررسی تأثیر درصد مقایسات (p) برای پرکردن مقدار جاافتاده در هر خوشه

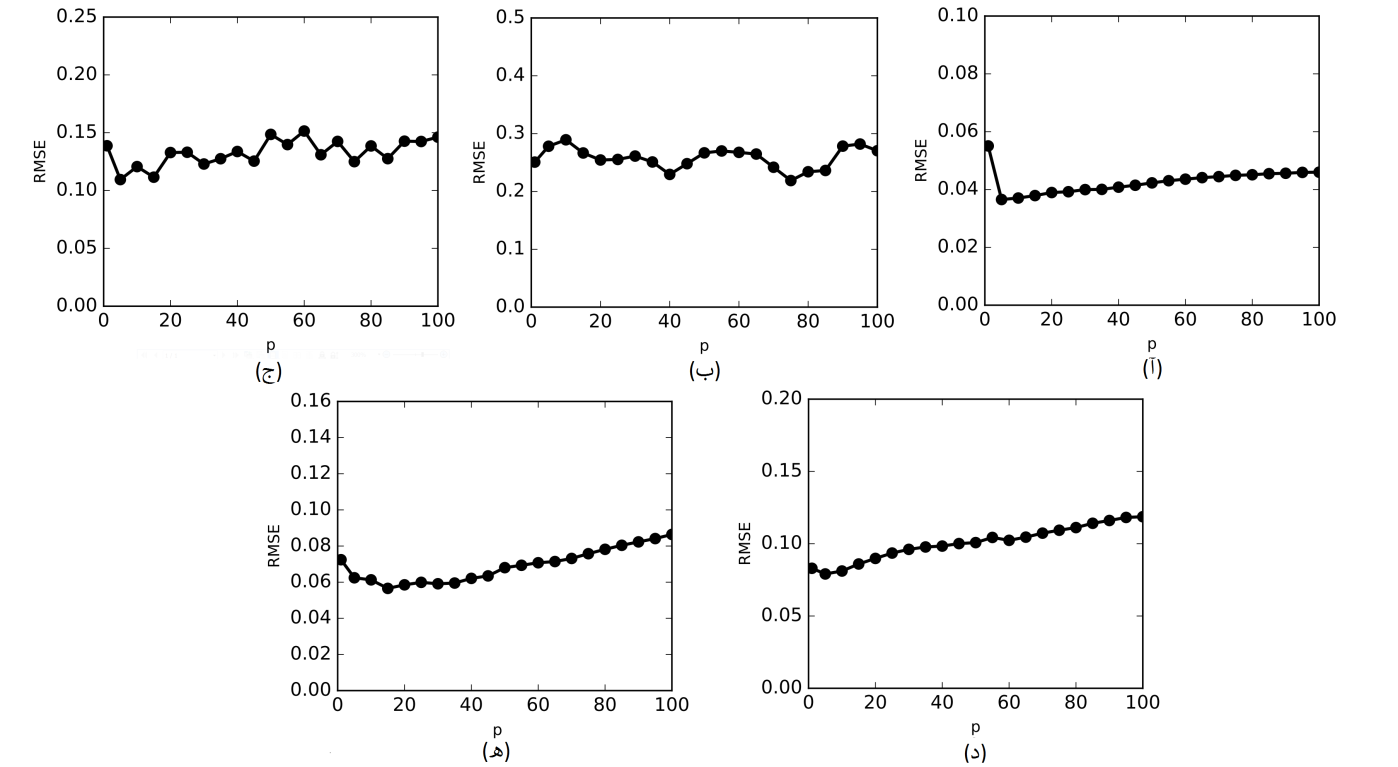
در این قسمت نتایج آزمایش‌هایی برای بررسی تأثیر مقدار p در میزان خطا بررسی می‌شود. ابتدا به تعداد 10% کل نمونه‌ها، مقدار جاافتاده به‌صورت تصادفی در هر مجموعه‌داده ایجادشده است. سپس مقدار p بین 1 تا 100 درصد با گام 5 درصد، برای مقادیر ثابت c و t، تغییر می‌یابند و نتایج ثبت می‌شوند. شکل ‏3‑4 مقدار خطای RMSE را برای مقادیر مختلف p برای مجموعه‌داده‌های مختلف نمایش می‌دهد. همان‌طور که از نمودارها پیداست، برای مجموعه‌داده‌های Iris، Glass و Wholesale Customers پس از کاهش مقدار خطا در مقادیر کوچک p، با افزایش مقدار p، مقدار خطا نیز یک سیر تقریباً صعودیِ یکنوا داشته است. برای مجموعه‌داده‌ Wine نیز اگرچه نوسان برای مقدار خطا وجود داشته است، بااین‌حال پس از یک مقدار بهینه (5%p=)، مقدار خطا یک سیر به‌طورکلی صعودی را داشته ‌است. برای مجموعه‌داده Haberman مقدار بهینه­ی p در مقادیر بالاتر حاصل‌شده است. درواقع نمودارها این مفهوم را بیان می‌کنند که مقدار بهینه­یk در روش k نزدیک‌ترین همسایگان نه بسیار کوچک است، نه بسیار نزدیک به تعداد کل نمونه‌های موجود در هر خوشه. مقدار بهینه‌ی p برای هر مجموعه‌داده را می‌توان طی آزمایش‌ها تعیین نمود.

### بررسی تأثیر حد آستانه (t) برای تعیین روش تخمین

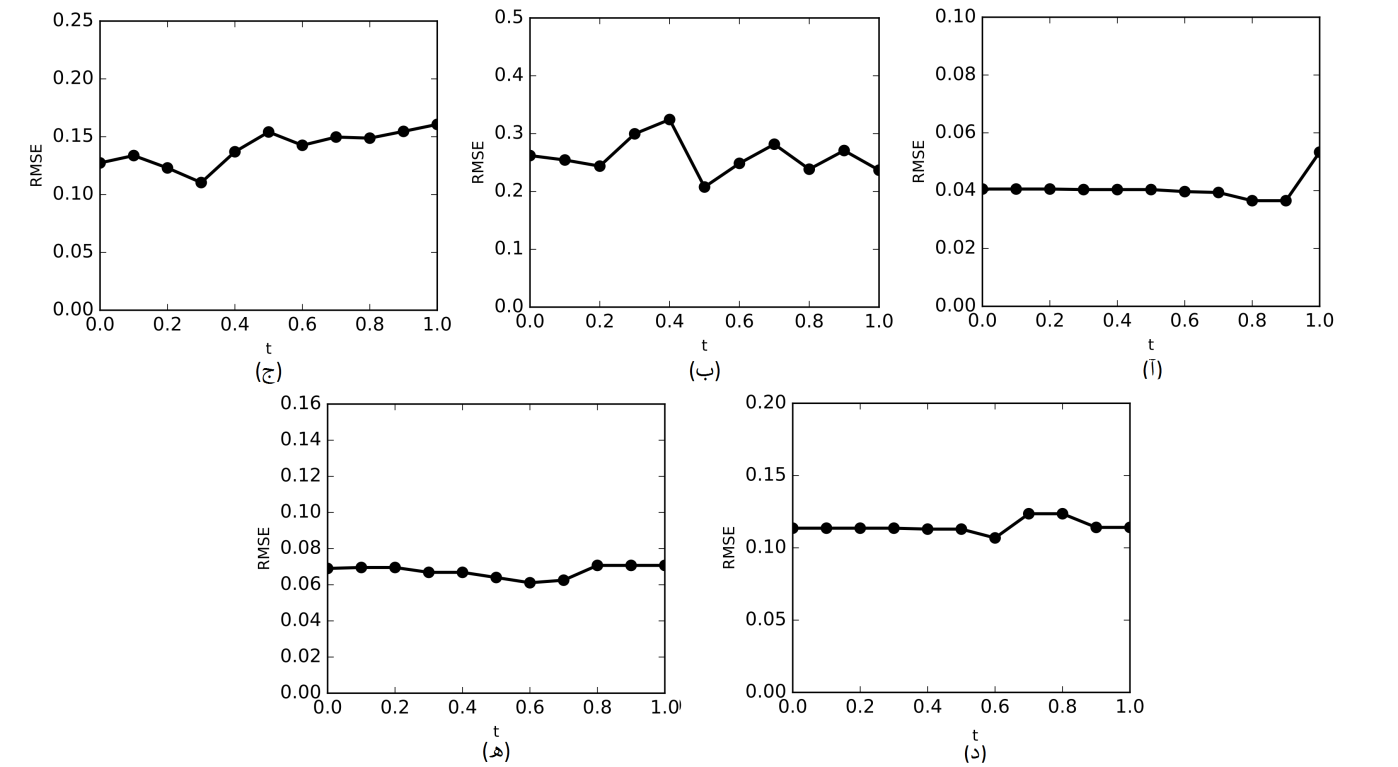
در این قسمت نتایج آزمایش‌هایی برای بررسی تأثیر مقدار t در میزان خطا بررسی می‌شود. ابتدا به تعداد 10% کل نمونه‌ها، مقدار جاافتاده به‌صورت تصادفی در هر مجموعه‌داده ایجادشده است. سپس مقدار t بین 0 تا 1 با گام 1/0 برای مقادیر ثابت p و c، تغییر می‌یابند و نتایج ثبت می‌شوند. شکل ‏3‑5 مقدار خطای RMSE را برای مقادیر مختلف t برای مجموعه‌داده‌های مختلف نمایش می‌دهد. همان‌طور که از نمودارها پیداست، برای مجموعه‌داده‌های مختلف یک مقدار بهینه‌ی 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 به‌تنهایی عمل کند. همان‌طور که پیداست، در که بیانگر روش k نزدیک‌ترین همسایگان وزن‌دار به‌تنهایی و در که بیانگر روش رگرسیون خطی به‌تنهایی است، مقادیر خطا بیش از مقدار کمینه برای خطاست. مختلف بودن مقادیر بهینه برای t به این معناست که برای مجموعه‌داده‌های مختلف یکی از روش‌ها بر دیگری غلبه دارد. درواقع میزان همبستگی درون نمونه‌های افرازهای تشکیل‌شده از مجموعه‌داده‌های مختلف، متفاوت است.



شکل ‏3‑3 تاثیر تعداد خوشه‌ها (c) بر روی مقدار RMSE در مجموعه‌داده‌های (آ) Wholesale Customers (ب) Haberman (ج) Wine (د) Glass (ه) Iris



شکل ‏3‑4 تاثیر درصد مقایسات (p) بر روی مقدار RMSE در مجموعه‌داده‌های (آ) Wholesale Customers (ب) Haberman (ج) Wine (د) Glass (ه) Iris



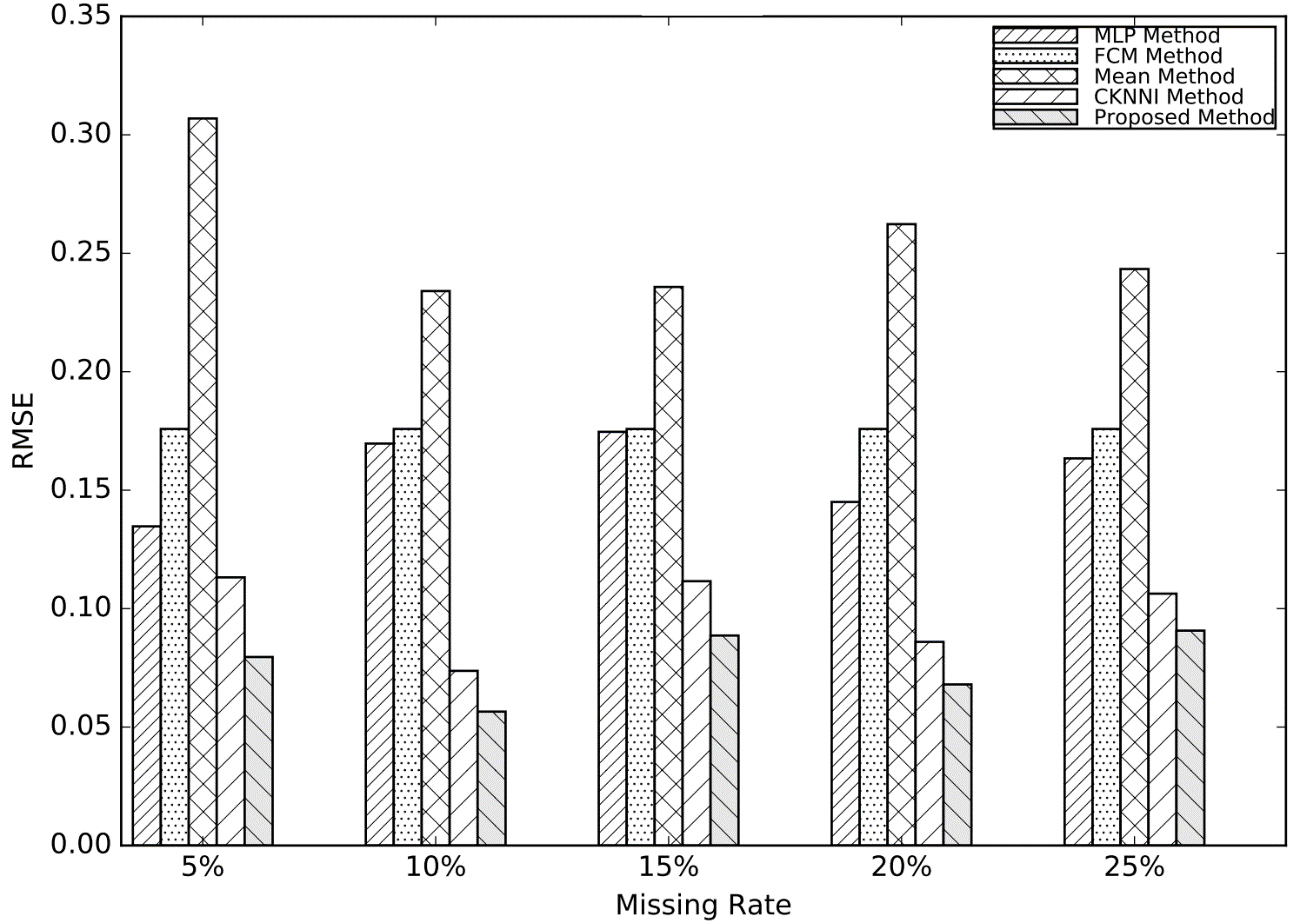
شکل ‏3‑5 تاثیر حدآستانه (t) بر روی مقدار RMSE در مجموعه‌داده‌های (آ) Wholesale Customers (ب) Haberman (ج) Wine (د) Glass (ه) Iris

### مقایسه با روش پرکردن با مقدار میانگین و CKNNI

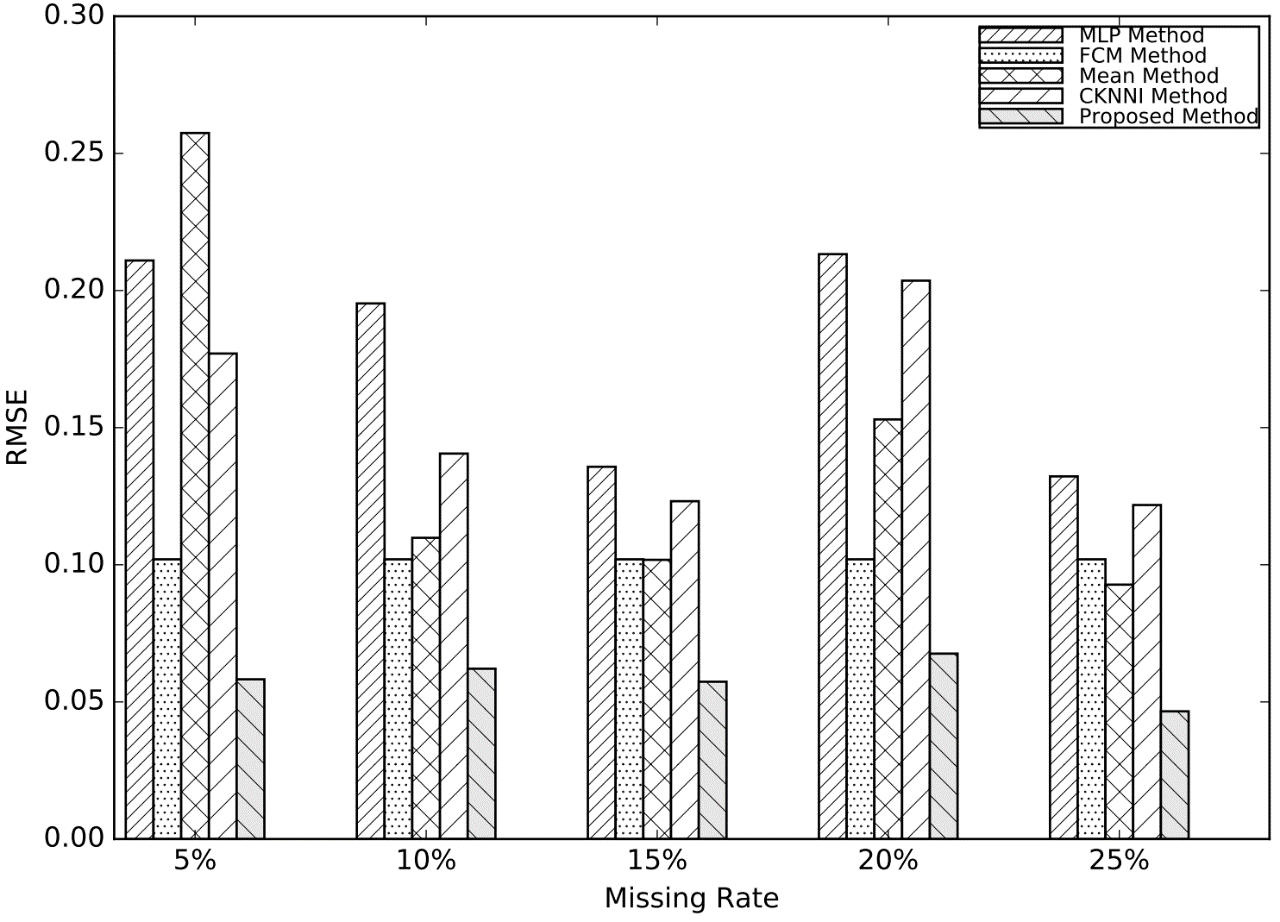
در این قسمت میزان خطای تخمین داده‌ها در مقایسه با چهار روش پرکردن با مقدار میانگین، روش تخمین با شبکه عصبی پرسپترون چندلایه (MLP) [25] ، روش پرکردن با خوشه­بندی فازی c-means (FCM) [89] و روش k خوشه‌ و نزدیک‌ترین همسایه‌ی مبتنی بر دسته (CKNNI) [56] مقایسه می‌شود. ابتدا به تعداد 5%، 10%، 15%، 20% و 25% کل نمونه‌ها، مقدار جاافتاده به‌صورت تصادفی در هر مجموعه‌داده ایجادشده است. سپس مقادیر جاافتاده با چهار روش ذکرشده و روش پیشنهادی در این فصل، پرشده‌اند. مقدار خطای RMSE برای این روش‌ها محاسبه‌شده و نتایج برای مجموعه‌داده‌های مختلف ثبت شدند. شکل ‏3‑6 تا شکل ‏3‑10 مقدار خطای RMSE محاسبه‌شده برای مجموعه‌داده‌های مختلف برای درصدهای مختلف مقادیر جاافتاده تولیدشده و برای سه روش ذکرشده را نشان ‌می‌دهد. (شکلها خیلی دورن. میگم بیارشون قبل از بخش 5-3) همان‌طور که از شکل‌ها پیداست، برای همه مجموعه‌داده‌ها به‌غیراز Haberman رویکرد پیشنهادی خطای تخمین کمتری نسبت به چهار روش دیگر داشته­­ است. اگرچه در مجموعه­داده Haberman دقت تخمین روش پیشنهادی نسبت به دیگرها کمتر است، بااین‌حال روش پیشنهادی عملکرد بسیار نزدیکی نسبت به بهترین روش در 5%، 20% و 25% جاافتادگی داشته ­است. علت ضعف نسبی روش پیشنهادی در این مجموعه­داده احتمالاً تعداد کم صفات پیش­بینی کننده در این مجموعه­داده است. به‌طورکلی با توجه به نتایج موجود می­توان مشاهده کرد که خطای روش پیشنهادی در اکثر موارد (به‌غیراز مجموعه­داده Haberman) از دیگر روش­های مقایسه شده کمتر است.

## خلاصه و نتیجه‌گیری

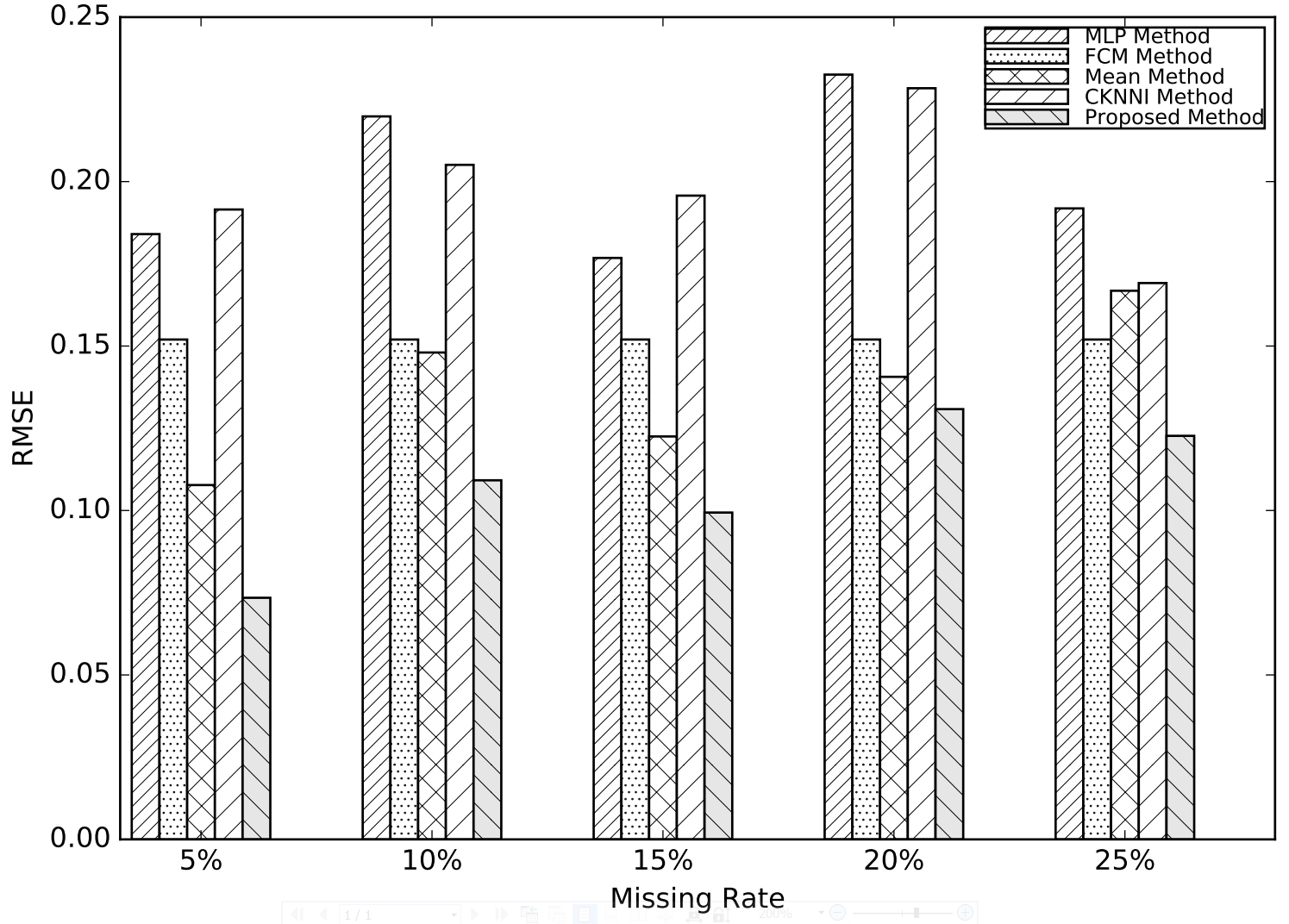
در اين فصل روشی جدید و دومرحله‌ای برای تخمین داده‌های جاافتاده ارائه شد. روش پیشنهادی شامل یک مرحله خوشه‌بندی و ترکیب دو روش پایه برای تخمین مقادیر جاافتاده است. تأکید روش پیشنهادی بر استفاده‌ از همبستگی‌های موجود بر مجموعه‌داده‌ها تا حد ممکن، برای تخمین مقادیر جاافتاده است. نتایج بررسی‌شده بر روی پنج مجموعه‌داده 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 عملیاتِ زمان­برِ جستجوی کامل برای انتخاب بهترین ترکیب پارامترها قابل برطرف شدن است.



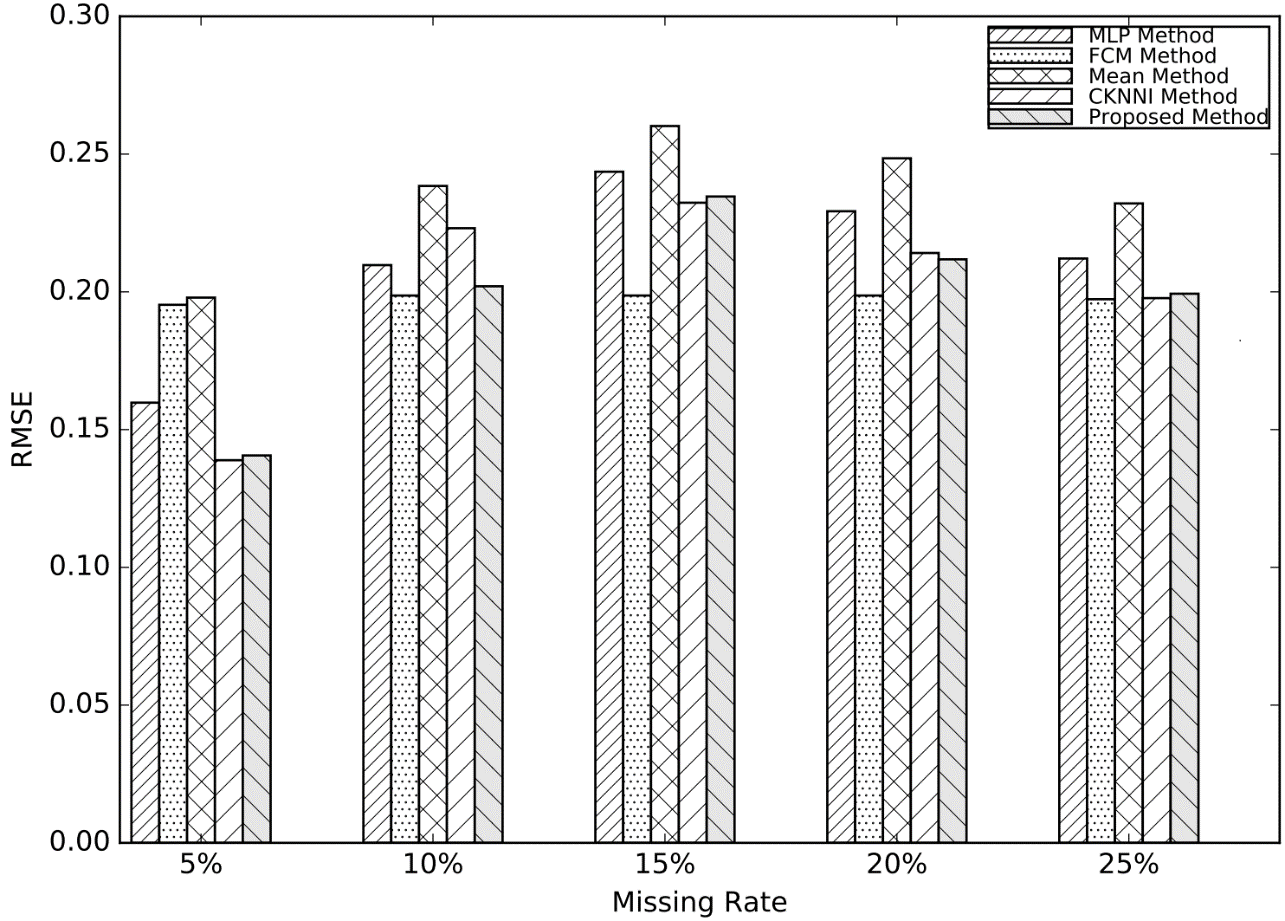
شکل ‏3‑6 مقایسه مقدار RMSE برای رویکرد ارائه‌شده و دیگر روش‌ها در مجموعه‌داده Iris



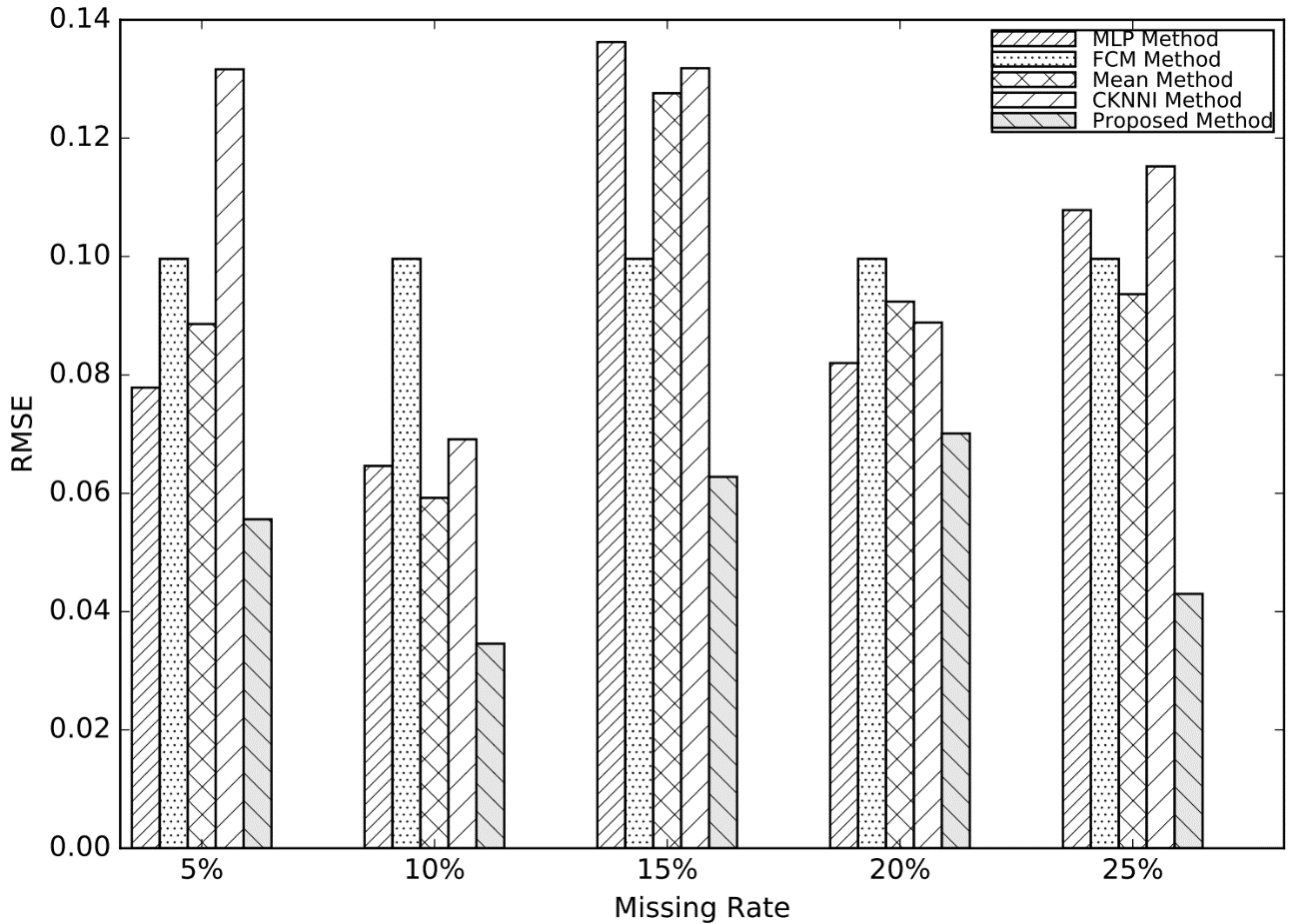
شکل ‏3‑7 مقایسه مقدار RMSE برای رویکرد ارائه‌شده و دیگر روش‌ها در مجموعه‌داده Glass



شکل ‏3‑8 مقایسه مقدار RMSE برای رویکرد ارائه‌شده و دیگر روش‌ها در مجموعه‌داده Wine



شکل ‏3‑9 مقایسه مقدار RMSE برای رویکرد ارائه‌شده و دیگر روش‌ها در مجموعه‌داده Haberman



شکل ‏3‑10 مقایسه مقدار RMSE برای رویکرد ارائه‌شده و دیگر روش‌ها در مجموعه‌داده Wholesale Customers

# فصل چهارم: جایگذاری مقادیر جاافتاده با استفاده از روش­های مبتنی بر همبستگی

## مقدمه

در این فصل چندین رویکرد جدید به نام روش­های جایگذاری مقادیر جاافتاده مبتنی بر همبستگی[[144]](#footnote-144) (CMIM) برای تخمین مقادیر جاافتاده عددی مطرح می­شود. تمامی این روش­ها مبتنی بر تلاش برای حداکثر کردن همبستگی خطی بین صفت جاافتاده و دیگر صفات معلوم هستند. این حداکثرسازی با استفاده از انتخاب زیرمجموعه­هایی از مجموعه‌داده­ ورودی که همبستگی قوی در آن­ها وجود دارد، انجام می­شود. انتخاب زیرمجموعه­های مطلوب با استفاده از رویکردهایی مبتنی بر انتخاب روبه‌جلو[[145]](#footnote-145) صورت می­گیرد. همچنین از روشی مبتنی بر جنگل تصادفی[[146]](#footnote-146)[91] برای پیدا کردن ترتیب پرکردن صفات جاافتاده استفاده می­شود. هنگامی‌که اولویت صفات جاافتاده برای پُرشدن یافته شد، مراحل کلی برای پرکردن مقادیر جاافتاده در هر نمونه عبارتنداز: انتخاب یک مجموعه پایه[[147]](#footnote-147) از نمونه­های کامل ، ساختن مجموعه‌داده‌هایی با روابط قوی بین صفات، با بسط دادن مجموعه پایه و درنهایت، تخمین مقادیر جاافتاده با اعمال یک یا چند مدل رگرسیون خطی بر مجموعه‌داده‌های ساخته‌شده. (جمله تغییر کرد ببینید مناسبه)

عملکرد رویکردهای ارائه‌شده بر روی پنج مجموعه­داده­ از دنیای واقعی در زمینه­های مختلف سنجیده می­شود. از سه معیار شناخته‌شده یعنی ریشه میانگین مربعات خطا، ضریب تعیین و میانگین خطای مطلق برای سنجش کیفیت جایگذاری مقادیر جاافتاده استفاده‌شده است. عملکرد رویکردهای ارائه‌شده با سه رویکرد دیگر در زمینه­ی تخمین مقادیر جاافتاده شامل جایگذاری با مقدار میانگین، جایگذاری با استفاده از نزدیک­ترین همسایگان و روشی مبتنی بر رگرسیون به نام الگوریتم جایگذاری با رگرسیون افزایشی صفات[[148]](#footnote-148) (IARI) [37] مقایسه شده­اند. نتایج آزمایش نشان می­دهد که در اکثر موارد، رویکردهای ارائه‌شده بهتر از دیگر روش­های مقایسه شده عمل می­کنند.

همبستگی بین صفات در یک مجموعه­داده از ویژگی­های طبیعی آن مجموعه­داده است. اغلب ممکن است زیرمجموعه­هایی درون یک مجموعه­داده یافت که روابط قوی­تری نسبت به روابط حاکم بر کل مجموعه­داده در آن­ها وجود دارد. در کاربردهای یادگیری ماشین یا داده­کاوی معمولاً صفات با همبستگی بالا به‌عنوان صفات افزونه[[149]](#footnote-149) در نظر گرفته می­شوند و فقط یکی از این صفات در تحلیل­ها به کار می­رود. اگرچه این روابط، ویژگی­های بسیاری مناسب و کمک‌کننده برای هدف تخمین مقادیر جاافتاده هستند؛ چراکه مقادیر نامعلوم می‌توانند با استفاده از این روابط تخمین زده شوند.

روش­های بسیار زیادی مبتنی بر رگرسیون برای هدف تخمین مقادیر جاافتاده وجود دارند. اگرچه اکثر این روش­ها مانند [37]، مدل­های رگرسیون را بر روی کل مجموعه­داده اعمال می­کنند. اعمال مدل­های رگرسیون بر افرازهایی از داده‌ها با همبستگی بالا می‌تواند منجر به تولید نتایج بهتری در تخمین داده‌ها توسط این مدل شود. بنابراین، عملکرد این روش­ها با توجه به این نکته می‌تواند بهبود یابد.

تحقیقات کمی به‌طور مستقیم از همبستگی­های موجود در مجموعه‌داده‌ها ­برای هدف تخمین مقادیر جاافتاده استفاده کرده­اند. [18] اگرچه بخش­هایی با همبستگی بالا را توسط درخت تصمیم پیدا می­کند، بااین‌حال معیار دقیقی را برای سنجش میزان همبستگی بین صفات به کار نبرده ­است. بنابراین ضروری به نظر می­رسد که مقدار همبستگی توسط معیاری دقیق برای اهداف ذکرشده اندازه­گیری کرد. بنابراین در این فصل روش­هایی معرفی می­شوند که قادر به یافتن بخش­هایی از داده‌ها با همبستگی بالا (توسط معیاری شناخته‌شده برای اندازه­گیری همبستگی قوی) هستند.

ترتیب انتخاب صفات جاافتاده برای عمل جایگذاری می‌تواند کیفیت جایگذاری را تحت تأثیر قرار دهد. بنابراین در روش­های پیشنهادی از رویکرد پیشنهادی توسط [37] برای رتبه­بندی صفات جاافتاده نیز برای بهبود دقت مقادیر تخمین زده‌شده، استفاده‌شده ­است.

رویکردهای پیشنهادی ماهیتی افزایشی[[150]](#footnote-150) دارند؛ یعنی پس از جایگذاری مقادیر جاافتاده برای هر صفت جاافتاده، از آن صفت به‌عنوان متغیرهای پیش­بینی کننده برای تخمین­های دیگر صفات استفاده خواهد شد. ماهیت افزایشی، امکان استفاده از اطلاعات موجود در مقادیر جاافتاده را فراهم می­کند. به‌علاوه، از صفت برچسب دسته­ی هر نمونه­ی جاافتاده نیز می‌توان به‌عنوان صفت پیش­بینی کننده در فرآیند تخمین استفاده کرد.

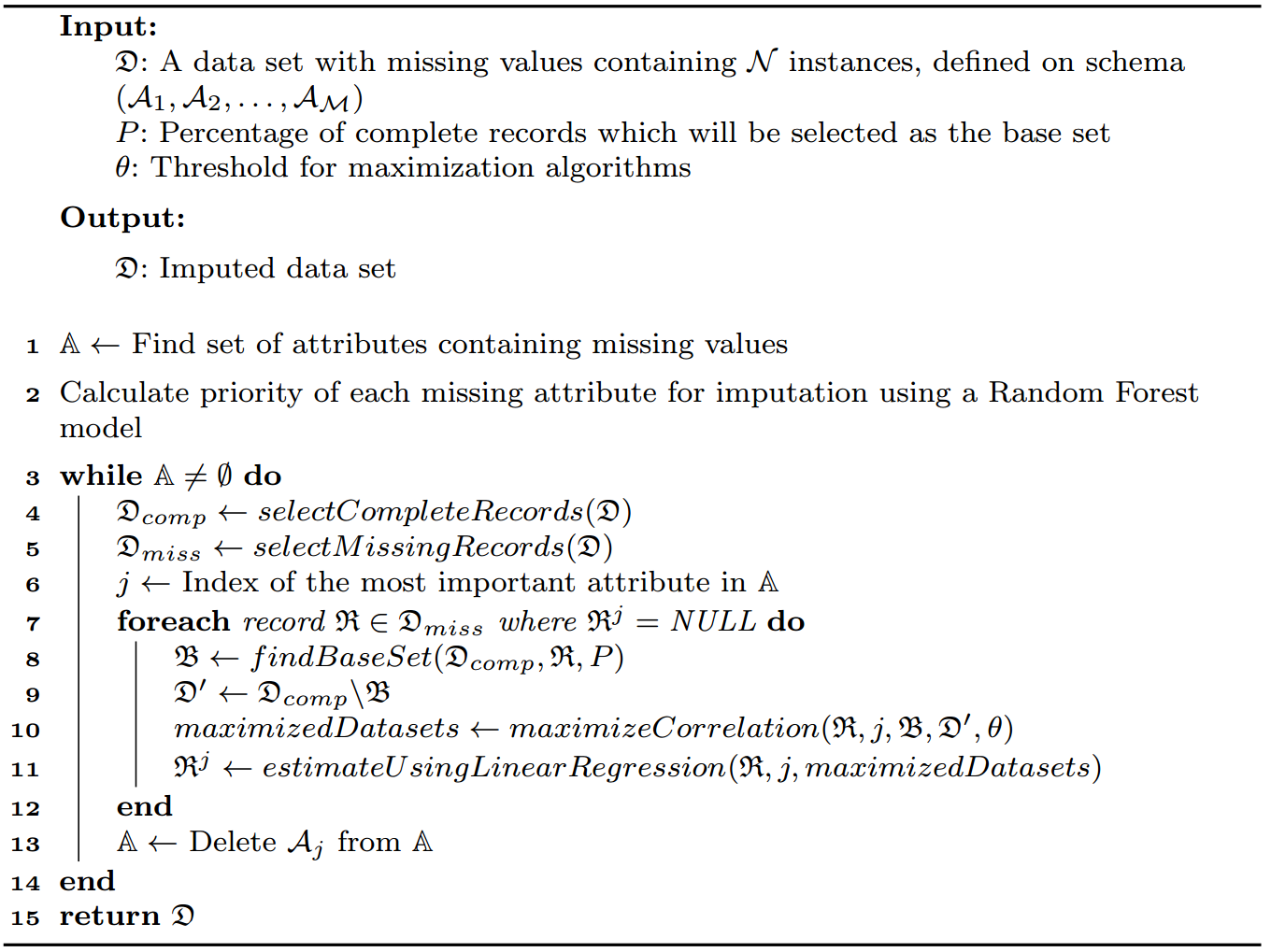
دو گام اصلی در رویکردهای پیشنهادی وجود دارد. گام حداکثرسازی[[151]](#footnote-151) و گام تخمین[[152]](#footnote-152). در گام حداکثرسازی تلاش می­شود تا زیرمجموعه­های موردنظر (که همبستگی بالا دارند) پیدا شوند. در گام تخمین، مقادیر جاافتاده با استفاده از زیرمجموعه­های یافت شده تخمین زده خواهند شد. رویکردهای مختلفی برای هریک از این دو گام مطرح می­شود. با ترکیب انواع مختلف گام­های حداکثرسازی و تخمین ده رویکرد متفاوت در این فصل پیشنهاد می­شود.

## جزئیات مراحل پیشنهادی

در این بخش و زیر بخش‌های آن، مراحل مختلف رویکردهای پیشنهادی برای جایگذاری مقادیر جاافتاده شرح داده می­شوند.

دو نکته‌ی مهم در روش‌های پیشنهادی وجود دارد. اول اینکه، ترتیب پرکردن صفات جاافتاده می‌تواند در نتیجه‌ی نهایی پرکردن مقادیر جاافتاده تأثیرگذار باشد. دوم اینکه، شناسایی بخش‌هایی از مجموعه‌داده‌ی که روابط خطی قوی بر آن‌ها حاکم است و سپس اعمال مدل‌های خطی داخل این بخش‌ها منجر به تخمین‌های دقیق‌تری از مقادیر جاافتاده می‌شود.

فرض کنید مجموعه‌داده‌ی‌‌ رابطه‌ای شامل نمونه (رکورد) است و هر نمونه مثل شامل صفت عددی است. مقدار صفتام را در رکورد ام در نشان می‌دهد. شکل ‏4‑1 رویه‌ی کلی را برای روش‌های ارائه‌شده نمایش می‌دهد.



شکل ‏4‑1 رویه کلی رویکردهای پیشنهادی برای تخمین مقادیر جاافتاده

در ابتدا، اولویت (اهمیت) هر صفت جاافتاده برای پُرشدن، توسط یک مدل جنگل تصادفی محاسبه می‌شود (خطوط 1 و 2). پس‌ازآن، مجموعه‌داده‌ی ورودی به دو مجموعه‌داده‌ی کامل () و مجموعه‌داده‌ی با مقادیر جاافتاده () تقسیم می‌شود (خطوط 4 و 5). سپس صفتی با بالاترین اولویت جهت پرشدن انتخاب می‌شود (خط 6). آنگاه نمونه‌هایی که مقدار جاافتاده در صفت ام (اندیس مهم‌ترین صفت جاافتاده) خود دارند، تک‌به‌تک برای فرآیند پر شدن انتخاب می‌شوند (خط 7). بعد از این کار، برای هر نمونه انتخاب‌شده یک مجموعه پایه از نمونه‌های کامل انتخاب می‌شود. سپس، با اعمال یک روش حداکثرسازی، بسته به نوع روش، یک یا چند زیرمجموعه از رکوردهای کامل با همبستگی بالا ساخته خواهند شد (خطوط 8 تا 10). حداکثرسازی با اعمال رویکردهایی مبتنی بر انتخاب روبه‌جلو بر مجموعه‌ی پایه () و با کمک نمونه‌های کاملِ غیر پایه () انجام می­شود. درنهایت، مقدار جاافتاده در یک رکورد با اعمال یک یا چند مدل خطی بر روی زیرمجموعه(ها)ی ساخته‌شده تخمین زده می‌شود. این کار آن‌قدر ادامه می‌یابد تا تمام صفات جاافتاده پر شوند.

همان­طور که الگوریتم پیشنهادی نشان می‌دهد، پس از پرکردن هر صفت جاافتاده، مقادیر آن صفت برای تخمین دیگر صفات جاافتاده، در تکرارهای بعدی استفاده خواهد شد. درواقع، مجموعه‌داده‌های کامل () پس از هر تکرار حلقه به‌روزرسانی می‌شود. الگوریتم کلی پیشنهادی دو پارامتر اصلی، یعنی و ، دارد که نیاز به مقداردهی دارند. در زیر بخش‌های بعدی، پارامترها، نحوه­ی انتخاب مجموعه پایه­ و گام­های حداکثرسازی و تخمین و دیگر جزئیات روش­های پیشنهادی شرح داده خواهند شد. درنهایت ترکیب‌های مختلف از گام­های حداکثرسازی و تخمین برای تولید روش­های پیشنهادی مختلف ذکر خواهند شد.

### محاسبه اولویت هر صفت جاافتاده برای هدف تخمین داده‌های جاافتاده

صفات جاافتاده در روش‌های پیشنهادی تک‌به‌تک به ترتیب اهمیتشان طبق رویکرد ارائه‌شده توسط [37] پر خواهند شد. صفتی با بیش‌ترین اهمیت ابتدا و صفتی با کمترین اهمیت در انتها پر خواهد شد. اولویت هر صفت جاافتاده برای پرشدن در سه گام محاسبه می‌شود. ابتدا، مقادیر جاافتاده در مجموعه‌داده‌ی ورودی (شامل مقادیر جاافتاده) با استفاده از یک رویکرد تخمین ساده، مثل تخمین با مقدار میانگین، پر می‌شوند. سپس، یک مدل جنگل تصادفی بر روی مجموعه‌داده‌ی پرشده ساخته می‌شود. درنهایت، مدل ساخته‌شده به نمونه‌های خارج از کیسه[[153]](#footnote-153) اعمال می‌شود تا اهمیت هر صفت اندازه‌گیری شود. جزئیات این کار در [92] شرح داده‌شده است. پس از پیدا کردن ترتیب هر صفت برای پرشدن، صفتی با بیش‌ترین اهمیت برای ادامه فرآیند پرکردن انتخاب خواهد شد. بنابراین با فرض این­که اندیس مهم‌ترین صفت جاافتاده باشد، آنگاه نمونه‌های که در صفت ام خود مقدار جاافتاده دارند، تک‌به‌تک برای ادامه در مراحل بعدیِ فرآیندِ جایگذاری انتخاب خواهند شد و درنهایت پر خواهند شد.

### انتخاب مجموعه پایه

تا این مرحله، ترتیب تخمین هریک از صفات جاافتاده تعیین شد. از این مرحله به بعد، نحوه‌ی پرکردن مقادیر جاافتاده برای هر نمونه‌ی جاافتاده شرح داده خواهد شد. در رویکردهای ارائه‌شده، هر صفت جاافتاده از یک نمونه‌ی جاافتاده با اعمال یک یا چند مدل رگرسیون خطی بر زیرمجموعه(هایی) از نمونه‌های کامل تخمین­ زده ­خواهند شد. ویژگی این زیرمجموعه‌‌های انتخاب‌شده این است که همبستگی قوی بین صفت جاافتاده و دیگر صفات وجود دارد.

در گام‌های بعدی، روش‌های حداکثرسازیِ همبستگی از رویکردهای مبتنی بر انتخاب روبه‌جلو برای پیدا کردن زیرمجموعه‌های مطلوب استفاده خواهند کرد. ازاین‌رو لازم است که یک مجموعه پایه مناسب از نمونه‌ها انتخاب شود. این مجموعه­ی پایه باید دیدِ کلیِ خوبی از روابط موجود در نمونه­هایی که در رویکرد انتخاب روبه‌جلو استفاده خواهند شد، بدهد. برای این منظور از روش معروف نزدیک‌ترین همسایگان برای تشکیل مجموعه پایه برای هر نمونه‌ی جاافتاده استفاده می‌شود. به بیان دقیق‌تر، درصد از نزدیک‌ترین نمونه‌های کامل به نمونه‌ی جاافتاده به‌عنوان نزدیک‌ترین همسایگان برای آن نمونه جاافتاده انتخاب می‌شوند. این همسایگان به‌عنوان مجموعه پایه برای آن نمونه‌ی جاافتاده انتخاب می‌شوند. بنابراین خروجی این گام برای هر نمونه‌ی جاافتاده، یک زیرمجموعه از نمونه‌های کامل خواهد بود.

فاصله‌ی بین دو نمونه‌ی و با استفاده از فرمول فاصله‌ی اقلیدسی مطابق با معادله ‏4‑1 محاسبه می‌شود.

|  |  |
| --- | --- |
| ‏4‑1 |  |

قابل‌ذکر است که در محاسبه­ی فاصله‌ی دو نمونه، فقط صفات با مقادیر معلوم در محاسبات دخیل می‌شوند.

### حداکثر کردن همبستگی بین صفات با بسط دادن مجموعه پایه

پس از انتخاب یک مجموعه پایه مناسب، الگوریتم تلاش می‌کند تا رابطه‌ی خطی بین صفت جاافتاده و دیگر صفات رکورد جاافتاده را در مجموعه پایه افزایش دهد. این افزایش با استفاده از بسط دادن مجموعه پایه توسط رکوردهای غیرپایه به نحوی انجام می‌گیرد که همبستگی بین صفت جاافتاده و دیگر صفات حداکثر شود. این کار با به‌کارگیری یک رویکرد انتخاب روبه‌جلو انجام می‌شود. ایده‌ی اصلی برای انجام چنین کاری این است که وجود روابط خطی قوی‌تر بین متغیر(های) پیش‌بینی کننده و متغیر پاسخ، منجر به تخمین بهتر متغیر پاسخ می‌شود.

رابطه خطی بین دو متغیر تصادفی را می‌توان با استفاده از ضریب همبستگی پیِرسون (PCC) اندازه‌گیری کرد. درواقع مقدار PCC میزان همبستگی دو متغیر تصادفی را نشان می‌دهد. PCC می‌تواند مقداری در بازه‌ی [1,+1-] داشته باشد. علامتِ ضریب همبستگی پیرسون، جهت رابطه­ی خطی را نشان می‌دهد. این علامت می‌تواند مثبت و یا منفی باشد. علامت مثبتِ همبستگی به این معنی است که با افزایش مقدار یک متغیر، مقدار متغیر دیگر افزایش می‌یابد و با کاهش مقدار متغیر، مقدار متغیر دیگر کاهش می‌یابد. مقدار منفی ضریب همبستگی پیرسون نشان می‌دهد که با افزایش مقدار یک متغیر مقدار متغیر دیگر کاهش می‌یابد و بالعکس [93].

با محاسبه مقدار PCC برای هر جفت از صفات یک مجموعه‌داده‌ی مثل ، یک ماتریس متقارن همبستگی با اندازه‌ی می‌توان ساخت که مدخل در این ماتریس، مقدار PCC بین صفات و را در نشان می‌دهد. در ادامه، نماد ماتریس همبستگی برای مجموعه‌داده‌ی را نشان می‌دهد. نماد اشاره به مدخل ردیف ام و ستونام از ماتریس همبستگی برای مجموعه‌داده‌ی خواهد داشت. معادلات ‏4‑2 و ‏4‑3 نحوه محاسبه ماتریس همبستگی و همبستگی بین دو متغیر تصادفی و را نشان می‌دهد.

|  |  |
| --- | --- |
| ‏4‑2 |  |

که در آن

|  |  |
| --- | --- |
| ‏4‑3 |  |

که در آن، مقدار کوواریانس بین مقادیر دو صفت و را نشان می‌دهد. و به ترتیب بیانگر مقادیر انحراف از معیار برای صفات و هستند.

ازآنجایی‌که میزان همبستگی بین صفات (قوی یا ضعیف بودن) در روش‌های ارائه‌شده مهم است، نه جهت همبستگی (همبستگی مثبت یا منفی)، از قدر مطلق مقدار ضریب همبستگی در محاسبات ماتریس همبستگی استفاده می‌شود. بنابراین برای هر مقدار و خواهیم داشت: .

با داشتن یک رکورد جاافتاده مثل که در آن مقدار جاافتاده است (صفتام آن رکورد مقدار تهی دارد)، الگوریتم­های پیشنهادی تلاش خواهند کرد تا همبستگی بین صفت جاافتاده و دیگر صفات را با اعمال یک روش انتخاب روبه‌جلو بر روی مجموعه پایه افزایش دهند.

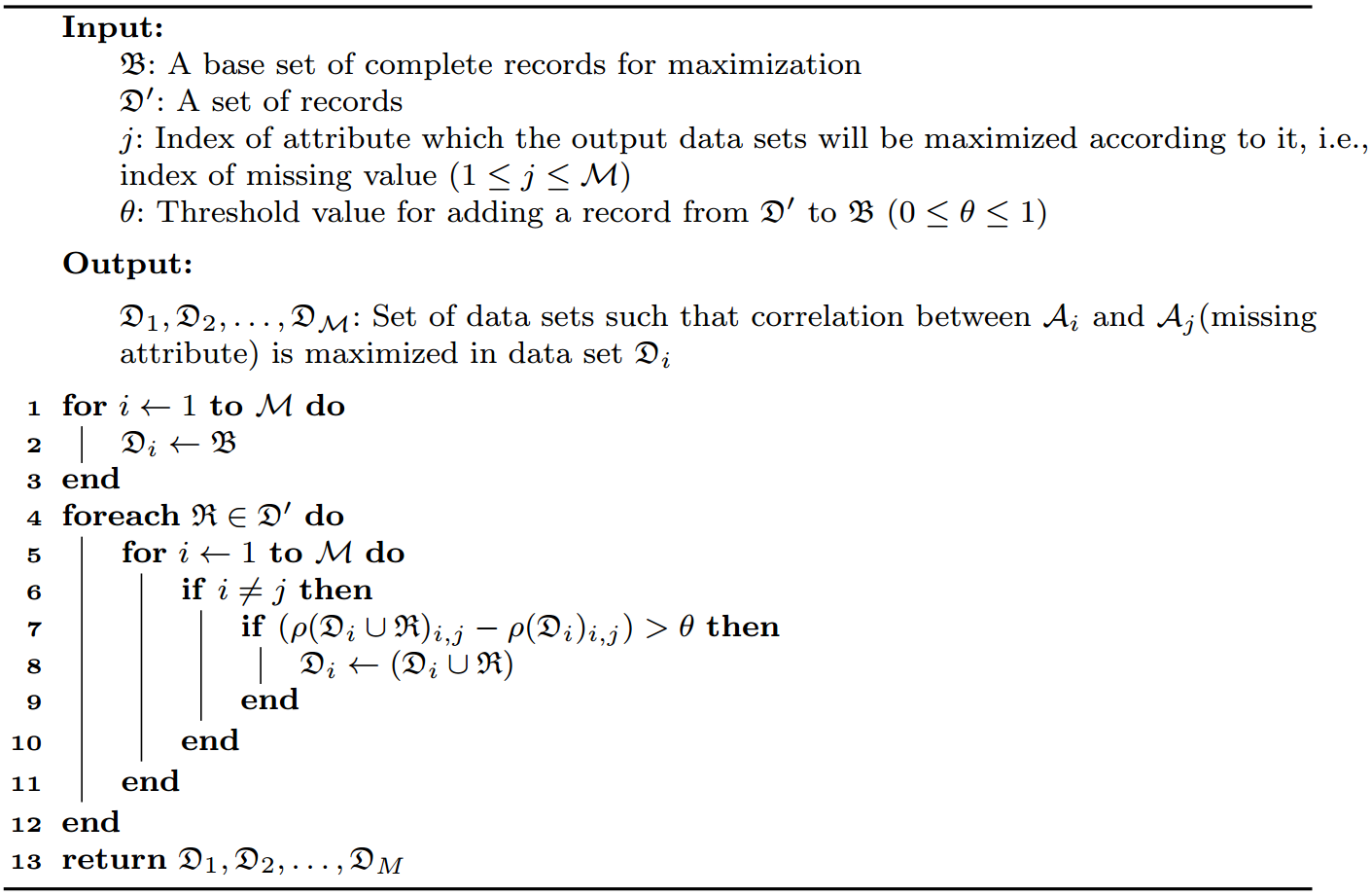
یک رویکرد کلی برای انتخاب یک ترکیب تایی[[154]](#footnote-154) از (زیرمجموعه­ی بدون تکرار عضوی از ) که مقدار ضریب همبستگی پیرسون در آن زیرمجموعه بین صفت جاافتاده و دیگر صفات حداکثر است، انجام جستجوی کامل[[155]](#footnote-155) است. یعنی مقدار ضریب همبستگی را برای تمام زیرمجموعه‌های تایی از محاسبه کرده و زیرمجموعه‌ای با بیش‌ترین مقدار PCC را انتخاب کنیم. برای این کار ( اندازه‌ی را نشان می‌دهد) انتخاب مختلف وجود خواهد داشت. بنابراین این رویکرد یک مشکل بدیهی دارد. این رویکرد بسیار زمان‌بر است و به‌ویژه زمانی که اندازه‌ی بسیار بزرگ است، عملی نخواهد بود. بنابراین نحوه‌ی انتخاب زیرمجموعه‌های مطلوب باید هوشمندانه‌تر باشد. پیشنهاد ما استفاده از رویکردی مبتنی بر انتخاب روبه‌جلو برای یافتن زیرمجموعه‌های مطلوب است. یعنی مجموعه‌ی پایه را با استفاده از رویکرد انتخاب روبه‌جلو بسط دهیم تا زیرمجموعه‌ی مطلوب حاصل شود. بدیهی است این رویکرد یافتن بهترین زیرمجموعه­ی ممکن را ضمانت نمی­کند، بااین‌حال از رویکرد جستجوی کامل منطقی‌تر به نظر می‌رسد. دو نوع الگوریتم حداکثرسازی همبستگی در این فصل ارائه‌شده است. در زیر بخش‌های بعدی این الگوریتم‌ها شرح داده خواهند شد.

#### حداکثر کردن همبستگی هر یک از صفات معلوم نسبت به صفت جاافتاده

الگوریتمِ اول سعی می‌کند تا مجموعه پایه را با در نظر گرفتن هر صفت غیر جاافتاده (معلوم) به‌طور جداگانه بسط دهد. ورودی‌های این الگوریتم عبارت‌اند از: اندیس یک صفت مثل که صفات معلوم در مجموعه‌داده‌های خروجی نسبت به آن صفت حداکثر می‌شوند، یک مجموعه از رکوردهای پایه، یک مجموعه دیگر از رکوردها که کاندیدای اضافه شدن به مجموعه پایه هستند و یک پارامترِ حد آستانه () که توسط کاربر تعریف می‌شود. خروجی الگوریتم، مجموعه‌داده‌ی خواهد بود؛ به‌نحوی‌که همبستگی بین صفت و صفت جاافتاده­ی در حداکثر شده است. جزئیات این الگوریتم در شکل ‏4‑2 نمایش داده شده است.

در ابتدا، مقادیر تمام مجموعه‌داده‌های برابر با مجموعه پایه قرار داده می­شوند (خطوط 1 تا 3). سپس الگوریتم به‌طور موقت نمونه‌های غیرپایه () را تک‌تک به هر اضافه می‌کند. پس از هر اضافه کردن، الگوریتم بررسی می‌کند که آیا مقدار همبستگی بین صفات و در افزایش‌یافته است یا خیر (خط 7). اگر مقدار PCC پس از اضافه کردن نمونه‌ی جدید بیش از مقدار حد آستانه یعنی افزایش‌یافته بود، آن نمونه به‌طور دائم به اضافه خواهد شد (خط 8). در غیر این صورت از آن نمونه چشم‌پوشی می‌شود و الگوریتم به سراغ نمونه‌ی بعدی می­رود. این کار آن‌قدر ادامه می‌یابد تا تمام نمونه‌های غیرپایه بررسی شوند.

نسخه­ا‌ی کمی متفاوت نیز برای این الگوریتم وجود دارد. در نسخه‌ی جایگزین، الگوریتم مقدار PCC اضافه‌شده را با مقدار اولیه PCC برای مجموعه پایه مقایسه می‌کند (به‌جای مقایسه با مقدار PCC در فعلی). برای این منظور عبارت در خط 7 از الگوریتم شکل ‏4‑2 در شرط جایگزین می‌شود.



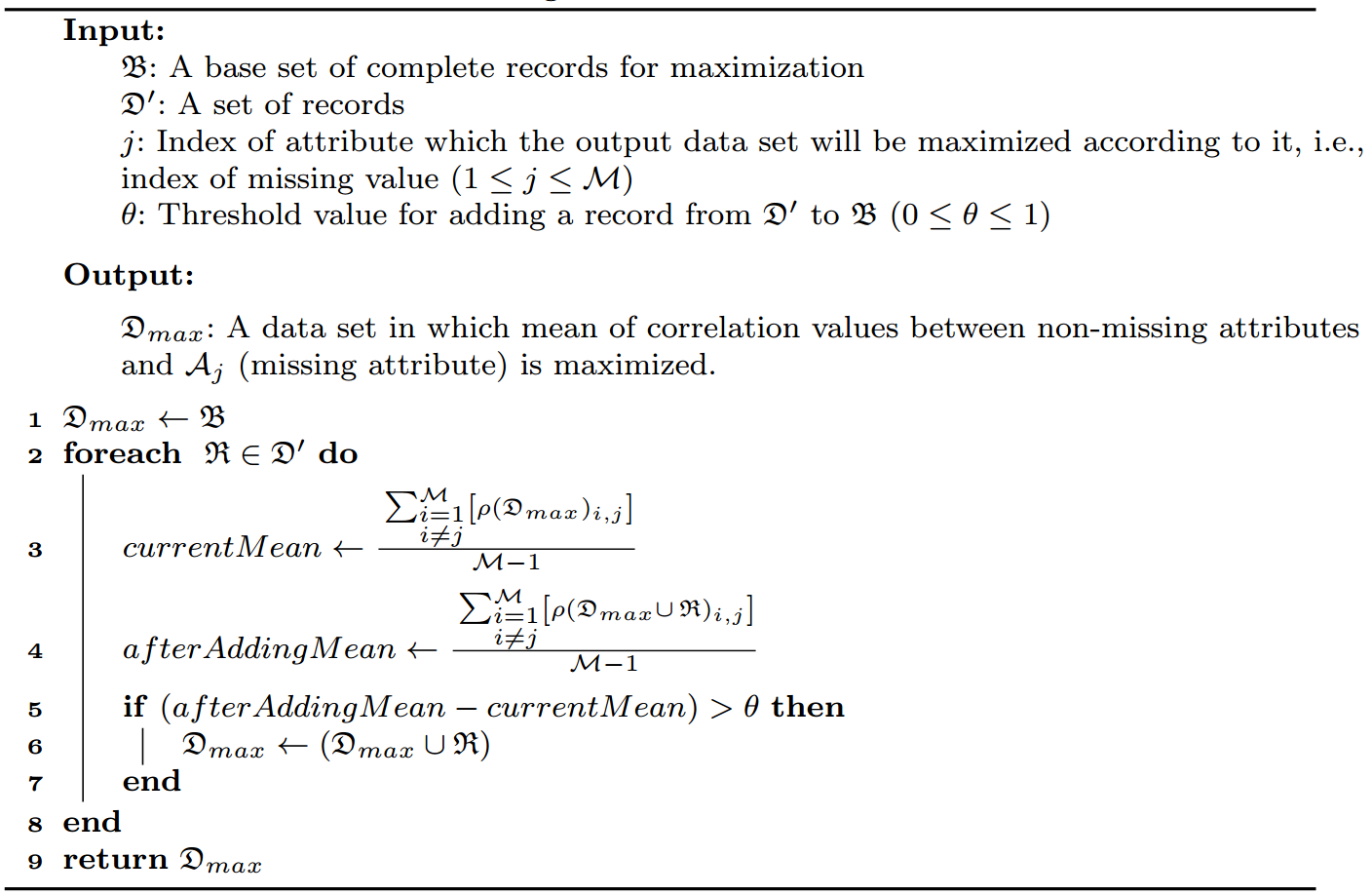
شکل ‏4‑2 الگوریتم حداکثرسازی همبستگی هریک از صفات نسبت صفات جاافتاده

#### حداکثر کردن میانگین مقادیر همبستگی بین صفات معلوم و صفت جاافتاده

این الگوریتم سعی می­کند تا میانگین مقادیر همبستگی بین صفات غیر جاافتاده و صفت جاافتاده را افزایش دهد. ورودی‌های این الگوریتم همانند الگوریتم قبلی هستند، اما خروجی تولیدشده متفاوت خواهد بود. خروجی، یک تک مجموعه‌داده‌ی () خواهد بود که در آن میانگین مقادیر همبستگی بین صفات غیر جاافتاده و صفت جاافتاده حداکثر شده است. مراحل این روش در شکل ‏4‑3 نمایش داده شده‌اند.

در ابتدا مقدار اولیه برابر با مجموعه پایه قرار می‌گیرد (خط 1). سپس الگوریتم به‌طور موقت نمونه‌های غیرپایه () را تک‌تک به اضافه می‌کند. پس از هر اضافه کردن، الگوریتم بررسی می‌کند که آیا میانگین مقادیر همبستگی بین صفات معلوم و صفت افزایش‌یافته ‌است یا خیر (خط 7). اگر مقدار PCC پس از اضافه کردن نمونه‌ی جدید بیش از مقدار حد آستانه یعنی افزایش‌یافته بود، آن نمونه به‌طور دائم به اضافه خواهد شد (خط 8). در غیر این صورت، از آن نمونه چشم‌پوشی می‌شود و الگوریتم به سراغ نمونه‌ی بعدی می‌رود. این کار آن‌قدر ادامه می‌یابد تا تمام نمونه‌های غیرپایه بررسی شوند.

همانند الگوریتم قبلی، نسخه‌ی متفاوتی نیز برای این الگوریتم وجود دارد. در نسخه‌ی جایگزین، الگوریتم مقدار PCC اضافه‌شده را با مقدار اولیه PCC برای مجموعه پایه مقایسه می‌کند (به‌جای مقایسه با مقدار PCC در فعلی). برای این منظور و به ترتیب در نمادهای سیگما در خطوط 3 و 4 در الگوریتم شکل ‏4‑3 جایگزین می‌شوند.



شکل ‏4‑3 الگوریتم حداکثرسازی میانگین مقادیر همبستگی بین صفات غیر جاافتاده نسبت به صفت جاافتاده

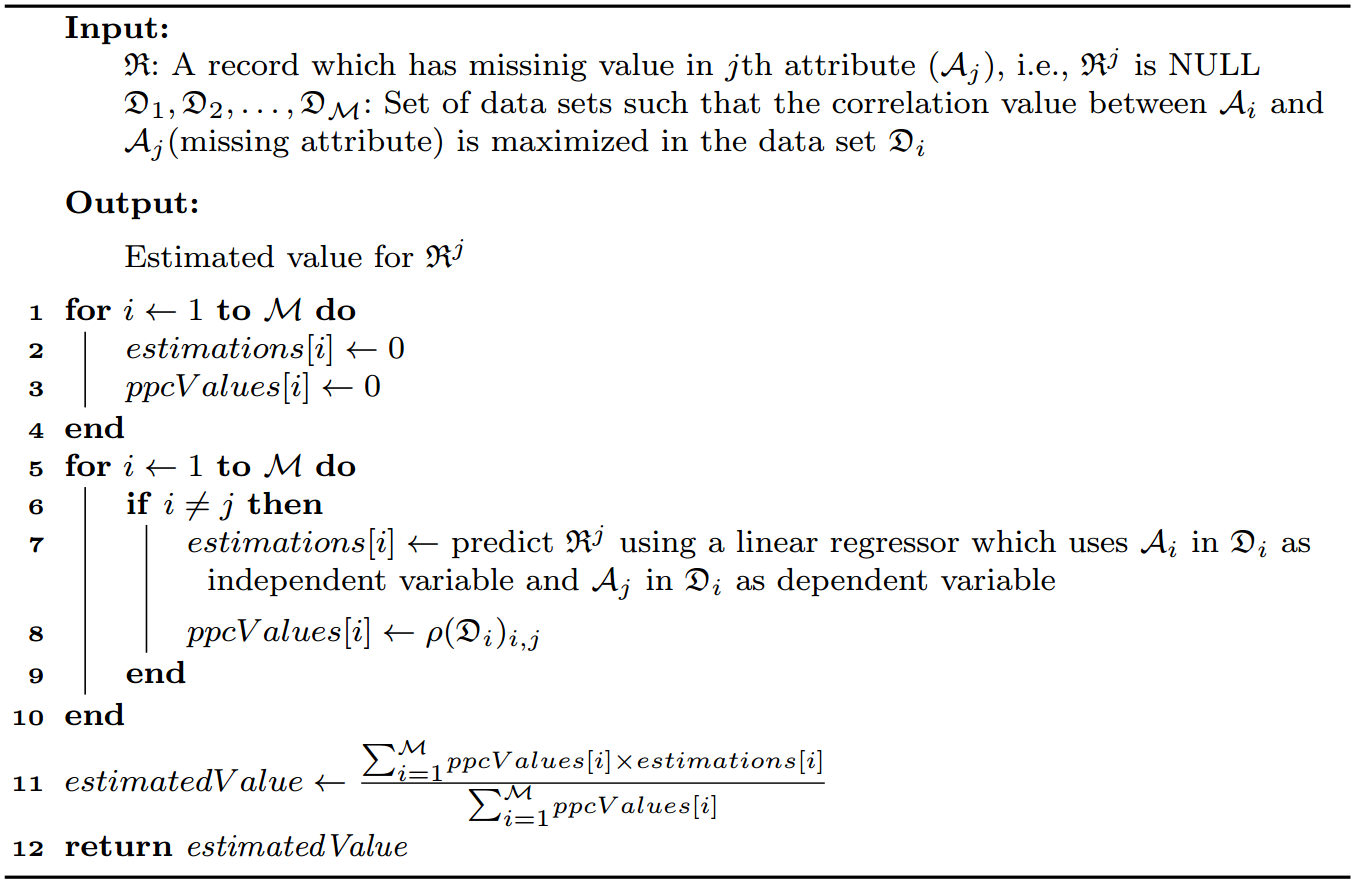
### تخمین مقادیر جاافتاده با استفاده از زیرمجموعه‌های یافته شده

پس از پیدا کردن زیرمجموعه‌های مطلوب از داده‌های کامل، آخرین گام، تخمین صفت جاافتاده در نمونه­ی جاافتاده است. بسته به نوعِ روش حداکثرسازیِ همبستگی در گام قبلی، رویکرد تخمین مقدار جاافتاده متفاوت خواهد بود. در رویکرد اول، تعداد رگرسور خطی ساده (تک صفتی) بر روی مجموعه‌داده‌ی ساخته‌شده اعمال خواهد شد، به‌نحوی‌که در مجموعه‌داده‌ی صفت به‌عنوان متغیر مستقل و به‌عنوان متغیر وابسته در ساخت مدل رگرسیون استفاده خواهد شد. هر مدل خطی رگرسیون در این حالت مطابق با معادله‌ی ‏4‑4 خواهد بود.

|  |  |
| --- | --- |
| ‏4‑4 |  |

که در آن نشان‌دهنده متغیر وابسته، بیانگر متغیر مستقل، برابر با خطای مدل و و ضرایب مدل رگرسیون ساخته‌شده هستند.

الگوریتم پیشنهادی هر مقدار جاافتاده را با محاسبه میانگین وزن­دار مقدار تخمین زده‌شده برای آن صفت جاافتاده جایگزین خواهد کرد. وزن مقدار تخمین زده‌شده‌ی ام برابر با این خواهد بود که تا چه میزان دو صفت و توانسته‌اند در نسبت به یکدیگر به‌صورت خطی مرتبط شوند. یعنی وزن مقدار تخمین زده‌ شده‌ی ام برابر با خواهد بود. هرچقدر که این همبستگی قوی­تر باشد، آن مقدار تخمین زده‌شده تأثیر بیش‌تری بر روی مقدار نهایی تأثیر خواهد گذاشت. شکل ‏4‑4 گام‌های توضیح داده‌شده را نمایش می‌دهد.



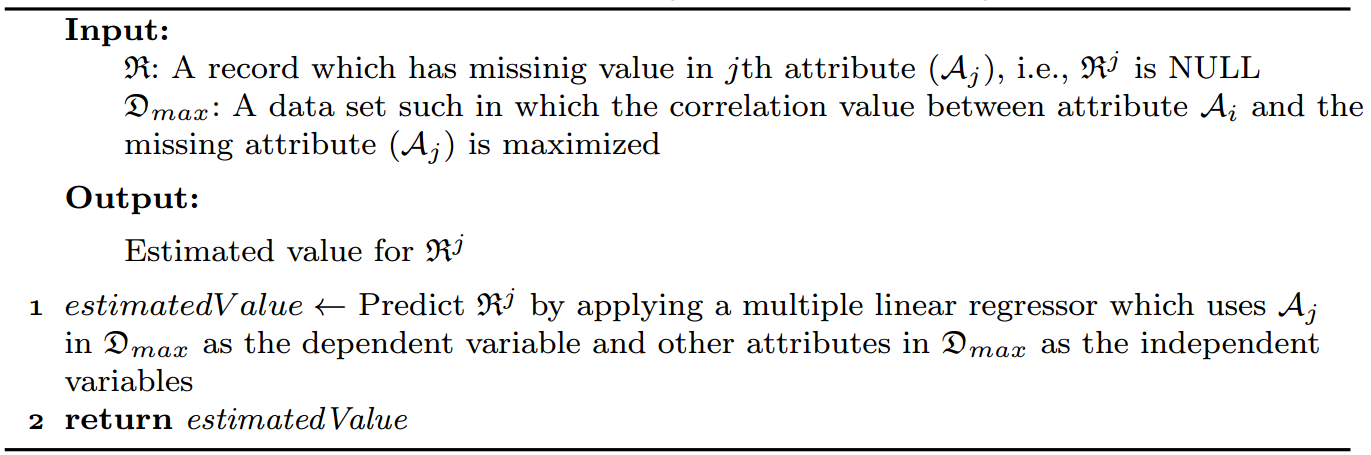
شکل ‏4‑4 الگوریتم اول برای گام تخمین

در دومین رویکردِ تخمین، فقط یک مدل رگرسیون خطی چند صفتی بر روی خروجی گام حداکثرسازی اعمال خواهد شد. از صفت جاافتاده به‌عنوان متغیر وا‌بسته و از بقیه صفات (غیر جاافتاده) به‌عنوان متغیرهای مستقل برای ساخت مدل خطی استفاده خواهد شد. یک مدل رگرسیون خطی چند صفتی مطابق با معادله­ی ‏4‑5 خواهد بود.

|  |  |
| --- | --- |
| ‏4‑5 |  |

که در آن نشان‌دهنده متغیر وابسته، بیانگر متغیرهای مستقل، برابر با خطای مدل و ضرایب مدل رگرسیون ساخته‌شده هستند. شکل ‏4‑5 گام­های توضیح داده‌شده را نمایش می‌دهد.

نکته‌ی مهمی که وجود دارد این است که هنگام اعمال مدل تخمین زننده، اگر مقادیر جاافتاده‌ی دیگری در یک نمونه‌ی جاافتاده وجود داشته باشد (غیر از اندیس)، این مقادیر ابتدا با یک روش ساده­ی تخمین مثل جایگذاری با مقدار میانگین به‌طور موقت پر می‌شوند و سپس نمونه‌ی جاافتاده به مدل رگرسیون اعمال می‌شود.



شکل ‏4‑5 الگوریتم دوم برای گام تخمین

### ترکیب گام‌های حداکثرسازی و تخمین

با ترکیب روش‌های مختلف حداکثرسازی و تخمین ارائه‌شده، روش‌های متفاوتی مبتنی بر همبستگی برای تخمین مقادیر جاافتاده می‌توان ارائه داد. جدول ‏4‑1 حالات مختلف ممکن برای تولید روش‌های مختلف را نشان می‌دهد. ستون اول شناسه‌ی روش را نشان می‌دهد. ستون دوم نحوه­ی حداکثرسازی همبستگی و ستون سوم نحوه‌ی تخمین نهایی مقدار جاافتاده را نمایش می‌دهد. ستون آخر نشان می‌دهد که آیا از برچسب کلاس هر نمونه­ی جاافتاده مانند دیگر صفات در فرآیند پرکردن مقادیر جاافتاده استفاده خواهد شد یا خیر.

جدول ‏4‑1 ترکیب‌های مختلف برای گام‌های حداکثرسازی و تخمین

|  |  |  |  |
| --- | --- | --- | --- |
| **شناسه**  **روش** | **روش حداکثرسازی** | **روش تخمین** | **استفاده از**  **برچسب دسته** |
| **1** | حداکثر کردن همبستگی هریک از صفات | تک مدل خطی چند صفته | ✔ |
| **2** | حداکثر کردن همبستگی هریک از صفات | تک مدل خطی چند صفته | ✖ |
| **3** | حداکثر کردن همبستگی هریک از صفات | نظرخواهی وزن­دار از چندین مدل خطی تک صفته | ✖ |
| **4** | حداکثر کردن میانگین مقادیر همبستگی تمام صفات | تک مدل خطی چند صفته | ✔ |
| **5** | حداکثر کردن میانگین مقادیر همبستگی تمام صفات | تک مدل خطی چند صفته | ✖ |
| **6** | حداکثر کردن همبستگی هریک از صفات (نسخه جایگزین) | تک مدل خطی چند صفته | ✔ |
| **7** | حداکثر کردن همبستگی هریک از صفات (نسخه جایگزین) | تک مدل خطی چند صفته | ✖ |
| **8** | حداکثر کردن همبستگی هریک از صفات (نسخه جایگزین) | نظرخواهی وزن­دار از چندین مدل خطی تک صفته | ✖ |
| **9** | حداکثر کردن میانگین مقادیر همبستگی تمام صفات (نسخه جایگزین) | تک مدل خطی چند صفته | ✔ |
| **10** | حداکثر کردن میانگین مقادیر همبستگی تمام صفات (نسخه جایگزین) | تک مدل خطی چند صفته | ✖ |

روش‌های 1، 2 و 3 از الگوریتم ارائه‌شده در شکل ‏4‑2 برای گام حداکثرسازی استفاده می‌کنند. روش‌های 6، 7 و 8 از نسخه‌ی جایگزینِ این الگوریتم که در بخش ‏4-2-3-1 توضیح داده شد، برای این منظور استفاده می‌نمایند. روش‌های 4 و 5 از الگوریتم ارائه‌شده در شکل ‏4‑3 و روش‌های 9 و 10 از نسخه‌ی جایگزین این الگوریتم که در بخش ‏4-2-3-2 توضیح داده شد، برای گام حداکثرسازی استفاده می‌نمایند.

در روش‌های 1، 2، 6 و 7 پس از تولید شدن خروجی گام حداکثرسازی () مجموعه‌داده‌ی که در آن

|  |  |
| --- | --- |
| ‏4‑6 |  |

است، انتخاب شده و سپس به الگوریتمِ شکل ‏4‑5 برای گام تخمین داده می‌شود. اندیس ، اندیس صفت جاافتاده است.

خروجی‌های گام حداکثرسازی در روش‌‌های 3 و 8 برای گام تخمین به‌طور مستقیم به الگوریتمِ شکل ‏4‑4 داده می‌شوند. خروجی‌ گام حداکثرسازی در روش‌های 4، 5، 9 و 10، یعنی ، برای گام تخمین به‌طور مستقیم به الگوریتمِ شکل ‏4‑5 داده می‌شود.

## آزمایش‌های تجربی

در این قسمت، روش‌های مورد مقایسه با روش‌های پیشنهادی، مجموعه‌داده‌های استفاده‌شده برای آزمایش‌ها و معیارهای ارزیابی کارایی رویکردهای متفاوت بررسی می‌شوند. علاوه بر این، آزمایش‌هایی برای بررسی تأثیر پارامتر‌های و بر روی نتایج تخمین نهایی انجام می‌شود. عملکرد رویکردهای پیشنهادی با سه روش «تخمین با استفاده از مقدار میانگین»، «تخمین با استفاده از نزدیک‌ترین همسایگان» و «جایگذاری با استفاده از رگرسیون افزایشی صفات» (IARI) [37] مقایسه شده است.

به‌منظور حذف عامل تصادفی بودن در آزمایش­ها، هر آزمایش 10 بار تکرار شده است. میانگین این 10 بار تکرار به‌عنوان نتیجه نهایی در این پژوهش گزارش‌شده‌اند. کدهای مورد استفاده برای رویکردهای مختلف همگی با زبان برنامه­نویسی پایتون و با کمک بعضی از بسته­های یادگیری ماشین از مخزن Scikit-Learn [94] پیاده­سازی شده­اند. در پیاده­سازی­ها از رگرسیون خطی خط‌الرأس[[156]](#footnote-156) [95] استفاده‌شده است. در این رگرسیون از تابع حداقل مربعات خطی[[157]](#footnote-157) به‌عنوان تابع هزینه­[[158]](#footnote-158) استفاده‌شده است. همچنین از تنظیم‌کردنِ[[159]](#footnote-159) با پارامتر تنظیمِ[[160]](#footnote-160) برای جلوگیری از مشکل بیش­برازش[[161]](#footnote-161) در ساخت این مدل رگرسیون بهره گرفته‌شده است.

### مجموعه‌داده‌های مورد آزمایش

روش­های مختلف جایگذاریِ مقادیر جاافتاده بر روی پنج مجموعه‌داده‌ی از دنیای واقعی موجود در مخزن داده­ی یادگیری ماشین UCI مطابق با جدول ‏3‑1، اعمال شدند. این مجموعه‌داده‌ها بدون مقادیر جاافتاده می‌باشند. در این مجموعه‌داده­های کامل، ابتدا مقادیر جاافتاده به‌طور مصنوعی ایجاد شدند. سپس مقادیر جاافتاده با استفاده از روش­های مختلف جایگذاری پُر شدند. از آن­جایی که مقادیر اصلی برای مقادیر جاافتاده­ی ساخته شده معلوم است، بنابراین عملکرد رویکردهای مختلف با دانستن مقادیر واقعی قابل ارزیابی است.

### پیش‌پردازش داده‌ها و نحوه ایجاد مقادیر جاافتاده

برای این­که امکان مقایسه نتایج بین رویکردهای مختلف وجود داشته باشد و مقایسه­­ها بامعنا باشند، مقادیر هر صفت از مجموعه‌داده‌های با استفاده از نرمال­سازی به روش نمره­ی z[[162]](#footnote-162) [1] نرمال شدند. در این نرمال‌سازی مقدار هر صفت به اطراف صفر متمرکز و بر انحراف معیار آن صفت تقسیم می­شود. این کار با محاسبه و جایگذاری هر مقدار با مقدار

|  |  |
| --- | --- |
| ‏4‑7 |  |

که در آن و به ترتیب بیانگر مقادیر میانگین و انحراف معیار برای صفت هستند، انجام می­شود.

به‌منظور ایجاد مقادیر جاافتاده در مجموعه‌داده‌ها بعضی صفات انتخاب می­شوند و بعضی از مقادیر این صفات به مقدار تهی تغییر می­یابند. برای این منظور ابتدا نیمی از صفات به‌طور تصادفی انتخاب می­شوند. سپس در 30%، 40%، 50% و 60% از نمونه­ها مقادیر جاافتاده به‌طور تصادفی ایجاد می‌شوند؛ بنابراین هر نمونه­ی جاافتاده ممکن است بین 1 تا 2/*m* صفت جاافتاده داشته باشد. همه­ی مقادیر جاافتاده مطابق با الگوی جاافتاده به‌طور تصادفی (MAR) ایجاد شدند.

### معیارهای ارزیابی

توانایی رویکردهای جایگذاری مقادیر جاافتاده در مقایسه با دیگر روش­ها توسط سه معیار شناخته­شده­­ی ذکرشده در بخش ‏2-5، یعنی ریشه میانگین مربعات خطا (RMSE) ، ضریب تعیین () و میانگین خطای مطلق (MAE) سنجیده می­شود.

### مقایسه عملکرد رویکردهای ارائه‌شده با دیگر روش‌ها

در این بخش آزمایش­هایی به‌منظور مقایسه عملکرد رویکردهای پیشنهادی با دیگر روش­های موجود، انجام می­شود. عملکرد روش­های ارائه‌شده با روش­های جایگذاری با مقدار میانگین، جایگذاری با استفاده از نزدیک­ترین همسایگان و روشی مبتنی بر رگرسیون به نام الگوریتم جایگذاری با رگرسیون افزایشی صفات (IARI) [37] مقایسه شده­اند.

جدول ‏4‑2 تا جدول ‏4‑6 عملکرد هریک از روش­های جایگذاری مقادیر جاافتاده مبتنی بر همبستگی پیشنهادی را در مقایسه با سه روش دیگر با توجه به معیارهای RMSE، MAE و برای درصدهای مختلف جاافتادگی در مجموعه‌داده‌های مختلف نشان می­دهد. مقادیر پررنگ، رویکردی را که از دیگر رویکردهای مقایسه شده بهتر عمل می‌کند را برای هر درصد جاافتادگی نمایش می‌دهد.

از جداول ذکرشده پیداست که با افزایش درصد جاافتادگی ایجادشده در مجموعه‌داده‌­ها، خطای کلیِ تخمین در همه­ی روش­ها افزایش می­یابد. به‌عبارت‌دیگر، تعدادِ بیش­ترِ داده­های کامل، منجر به افزایش دقت در تخمین­ها­ می­شود. این پدیده منطقی است، چراکه با افزایش درصد جاافتادگی اطلاعات مفید بیش­تری از دست می­رود.

همان­طور که انتظار می­رفت، روش­هایی که از برچسب دسته­ی نمونه­ها نیز به‌عنوان متغیر پیش­بینی کننده استفاده می­کنند، نتایج بهتری در مقابل با روش­های مشابه که از برچسب دسته نمونه­ها استفاده نمی­کنند، تولید کرده­اند.

در همه­ی مجموعه‌داده‌­ها به‌غیراز مجموعه‌داده‌ی Haberman، روش IARI تقریباً در همه­ی موارد از روش جایگذاری k-NN بهتر عمل کرده است و همچنین روش جایگذاری k-NN بهتر از روش جایگذاری با مقدار میانگین عمل کرده است. از طرفی در اکثر موارد، عملکرد کلی روش­های ارائه‌شده برای هر سه معیار مقایسه شده بهتر از IARI بوده است. بنابراین می‌توان گفت که به‌طورکلی، روش­های ارائه‌شده بهتر از دیگر روش­های مقایسه شده عمل می­کنند. قابل‌ذکر است که روش IARI به‌طور پیش­فرض از برچسب دسته­ی نمونه­ها به‌عنوان متغیر پیش‌بینی کننده استفاده می­کند.

برای مجموعه‌داده‌ها‌ی Wholesale Customers و Wine، تمام روش‌های پیشنهادی در تمام حالات از دیگر روش‌ها با توجه به دو شاخص‌‌‌ ارزیابی بهتر عمل کرده‌اند. برای مجموعه‌داده‌ی Iris، تمام رویکردهای پیشنهادی به‌جز هشتمین روش، از دیگر روش­های مقایسه شده با توجه تمام شاخص­های ارزیابی بهتر عمل کرده است. برای مجموعه‌داده‌ی Glass در شرایطی که 30% یا 60% جاافتادگی در نمونه­ها وجود دارد، در اکثر روش­ها به‌جز روش­های 3 و 8، روش­های پیشنهادی (CMIM) بهتر از IARI عمل کرده است. متأسفانه فقط روش­های 9 و 10 در معیار RMSE و هنگامی‌که 40% جاافتادگی وجود دارد و روش­های ۱، ۴، ۶، ۷، ۹ و ۱۰ در معیار RMSE و و روش‌های ۱،2، ۴، 5، ۶، ۷، ۹ و ۱۰ در معیار MAE هنگامی‌که 50٪ جاافتادگی وجود دارد، توانسته­اند از روش IARI بهتر عمل کنند. در مجموعه‌داده‌ی Haberman اگرچه روش­های IARI و k-NN بدتر از روش میانگین عمل کرده­اند، بااین‌حال مشاهده می­شود که روش­های پیشنهادی کمی بهتر از روش میانگین عمل کرده­اند. ضعف نسبی روش­ها دیگر نسبت به روش میانگین احتمالاً به دلیل تعداد کم متغیرهای پیش­بینی کننده در این مجموعه‌داده‌ است.

جدول ‏4‑2 مقایسه عملکرد رویکردهای ارائه‌شده با سه روش دیگر برای مجموعه‌داده‌ی Iris



جدول ‏4‑3 مقایسه عملکرد رویکردهای ارائه‌شده با سه روش دیگر برای مجموعه‌داده‌ی Wine



جدول ‏4‑4 مقایسه عملکرد رویکردهای ارائه‌شده با سه روش دیگر برای مجموعه‌داده‌ی Glass



جدول ‏4‑5 مقایسه عملکرد رویکردهای ارائه‌شده با سه روش دیگر برای مجموعه‌داده‌ی Haberman



جدول ‏4‑6 مقایسه عملکرد رویکردهای ارائه‌شده با سه روش دیگر برای مجموعه‌داده‌ی Wholesale Customers



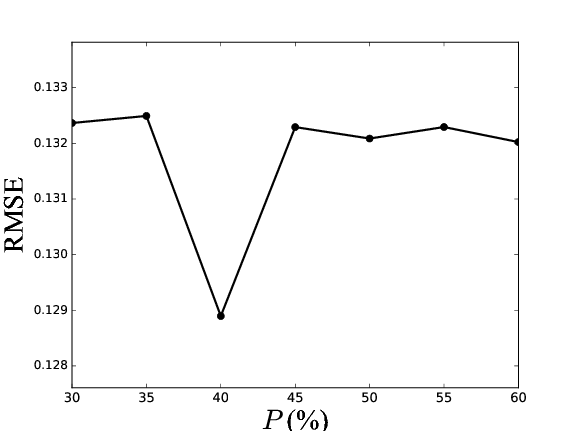
### تأثیر پارامترهای و

در این بخش تأثیر پارامترهای ورودی بر روی عملکرد رویکردهای ارائه‌شده بررسی می­شود. همان­طور که در بخش­های پیش توضیح داده ­شد، در روش­های ارائه‌شده دو پارامتر اصلی و نیاز به تنظیم دارند. پارامتر تعیین می­کند که آیا یک رکورد به مجموعه­ی پایه اضافه شود یا خیر. پارامتر ، درصدِ انتخاب از نزدیک‌ترین همسایگانِ کامل نسبت به هر نمونه‌ی جاافتاده را (به‌عنوان مجموعه‌‌‌ی پایه) تعیین می‌‌‌کند.

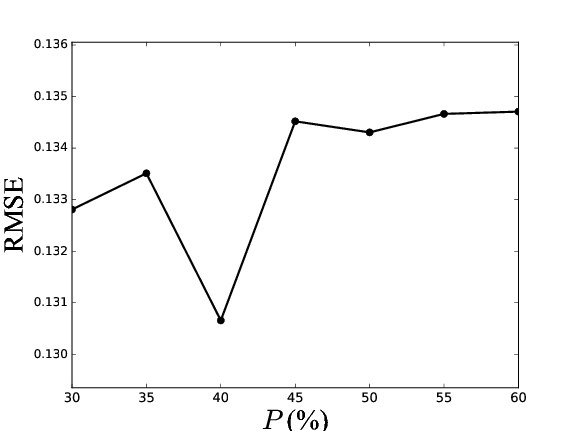
برای بررسی تأثیر پارامتر، مقدار برای یک مقدار ثابتِ بین 30 تا 60 با فواصل 5 تغییر یافتند و مقدار RMSE برای هر مقدار ثبت شد. شکل ‏4‑6 تا شکل ‏4‑15مقدار RMSE را برای مقادیر مختلف بر روی مجموعه‌داده‌ی Iris با 30% جاافتادگی نشان می­دهد. نتایج برای باقی مجموعه‌داده‌ها و دیگر معیارها به دلیل محدودیت فضا نشان داده نشده­اند.می‌توان مشاهده کرد که برای روش­هایی که فقط در استفاده یا عدم استفاده از برچسب دسته­ی نمونه­ها تفاوت دارند، یعنی جفت روش­های (1و 2)، (4و 5)، (6 و 7) و (9 و 10)، الگوهای تغییر مقدار RMSE نسبتاً یکسان است. برای روش­های 1 و 2، عملکرد بهینه در حاصل‌شده است. برای روش­های 4، 5، 8، 9 و 10 مقدار RMSE پس از رسیدن به مقدار بهینه روندی صعودی داشته است. در روش­های 6 و 7، با افزایش مقدار ، مقدار RMSE به‌طور یکنواخت کاهش‌یافته ­است. به‌طورکلی، نتایج نشان ‌داده‌اند که مقادیر خیلی پایین و خیلی بالا برای *p* مناسب نیستند. مقادیر بسیار بالا، مفهوم انتخاب روبه‌جلو را بی‌معنا می­کند. در مقابل، مقادیر بسیار پایین دید خوبی از روابط موجود در مجموعه پایه نخواهد داد.

برای بررسی تأثیر پارامتر ، مقدار برای یک مقدار ثابتِ بین 0/0 تا 0/1 با فواصل 01/0 تغییر یافتند و مقدار RMSE برای هر مقدار ثبت شد. شکل ‏4‑16 تا شکل ‏4‑25 مقدار RMSE را برای مقادیر مختلف بر روی مجموعه‌داده‌ی Iris با 30% جاافتادگی نشان می­دهد. مجدداً، به دلیل محدودیت فضا، نتایج برای باقی مجموعه‌داده‌ها و دیگر معیارها نشان داده نشده­اند. (این شکلها هم خیلی دورن)

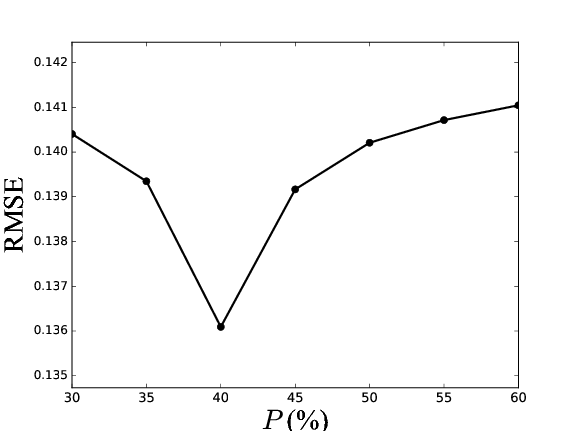
نمودار­ها ثابت ماندن مقدار RMSE برای مقادیر بالای (وقتی به 1 میل می­کند) را نشان می­دهند. این روند منطقی است، چراکه در این شرایط الگوریتم­ها به‌طور سخت­گیرانه اجازه­ی اضافه شدن رکوردهای غیرپایه را به مجموعه پایه خواهند داد. (این شکلارو اگه میتونستی دوتادوتا کنار هم بذاری به نظر من قشنگتر میشد. یعنی در واقع قرارشون میدادی تو یه جدول با 5 سزر و دوستون)



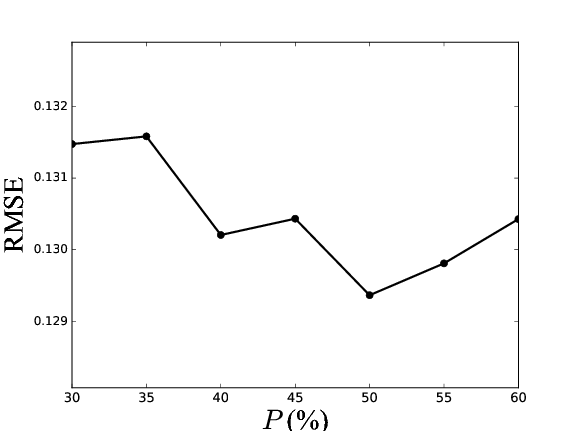
شکل ‏4‑6 تأثیر پارامتر P بر مقدار RMSE برای روش 1



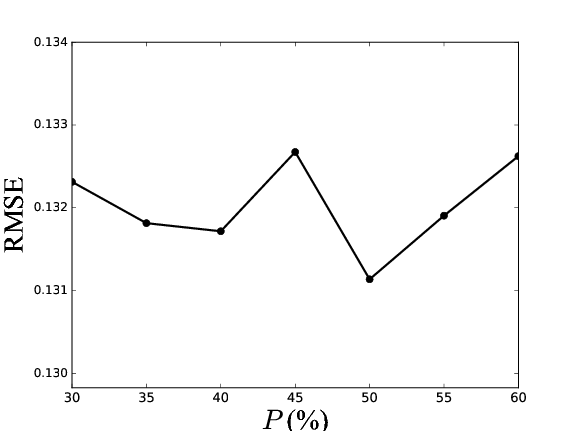
شکل ‏4‑7 تأثیر پارامتر P بر مقدار RMSE برای روش 2



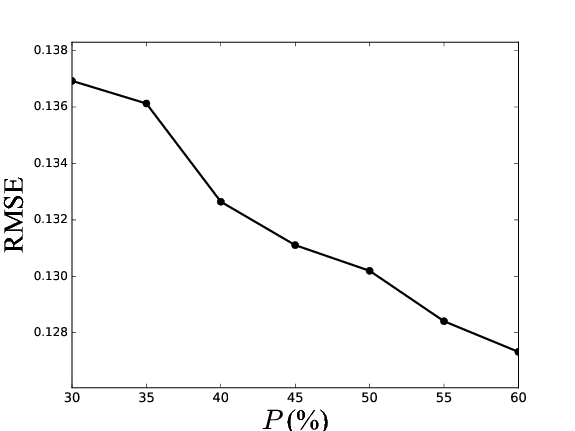
شکل ‏4‑8 تأثیر پارامتر P بر مقدار RMSE برای روش 3



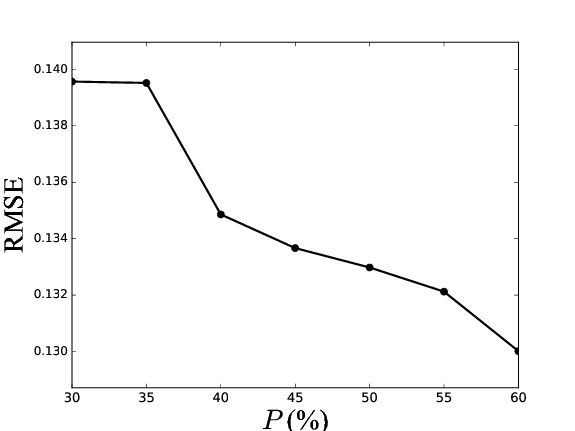
شکل ‏4‑9 تأثیر پارامتر P بر مقدار RMSE برای روش 4



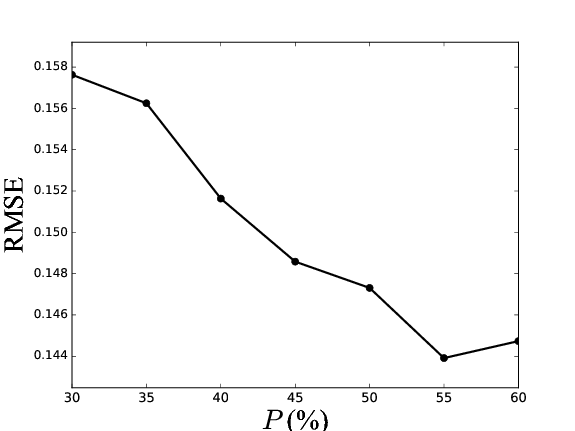
شکل ‏4‑10 تأثیر پارامتر P بر مقدار RMSE برای روش 5



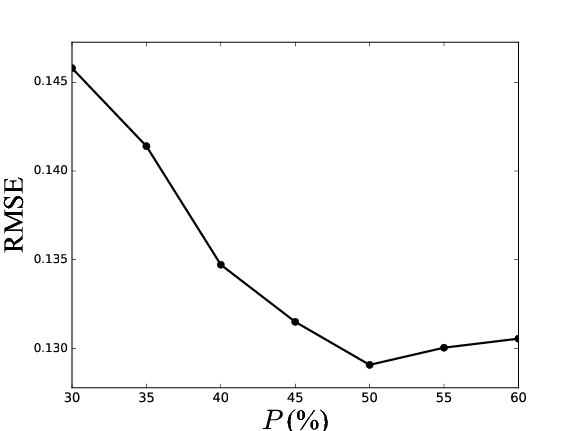
شکل ‏4‑11 تأثیر پارامتر P بر مقدار RMSE برای روش 6



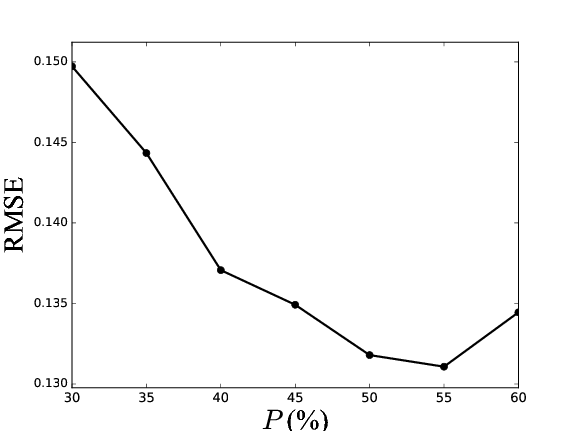
شکل ‏4‑12 تأثیر پارامتر P بر مقدار RMSE برای روش 7



شکل ‏4‑13 تأثیر پارامتر P بر مقدار RMSE برای روش 8



شکل ‏4‑14 تأثیر پارامتر P بر مقدار RMSE برای روش 9



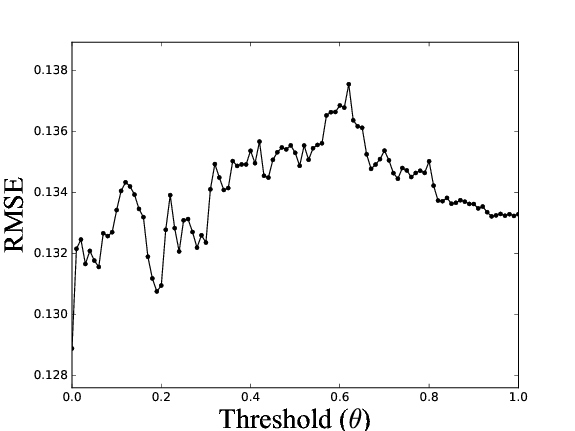
شکل ‏4‑15 تأثیر پارامتر P بر مقدار RMSE برای روش 10

مقدار کمینه سراسری برای روش­های 1، 2، 4، 5، 6 و 7 در حدود آستانه نزدیک به صفر رخ‌داده‌اند. برای روش­های 9 و 10، مقدار بهینه­ی RMSE در ظاهرشده است. مجدداً دیده می­شود، برای روش­هایی که فقط در استفاده یا عدم استفاده از برچسب دسته­ی نمونه­ها تفاوت دارند، الگوهای نوسانات برای مقدار RMSE نسبتاً یکسان است.

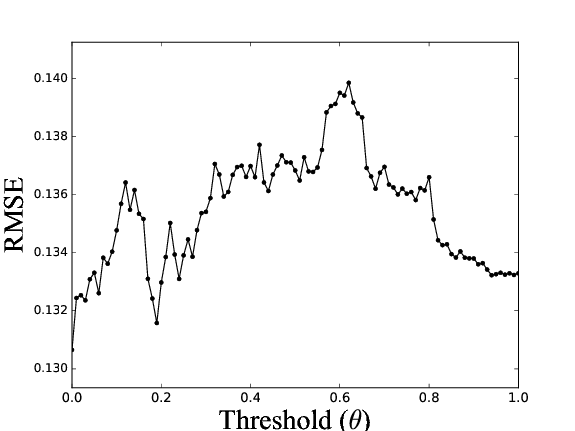
## خلاصه و نتیجه‌­گیری

در این فصل چندین رویکرد جدید برای جایگذاری مقادیر جاافتاده عددی، مبتنی بر همبستگی برای مطرح‌ شد. ایده اصلی در این روش­ها، اعمال مدل­های رگرسیون در بخش­هایی از مجموعه­داده با همبستگی بالا به‌جای کل مجموعه­داده است.

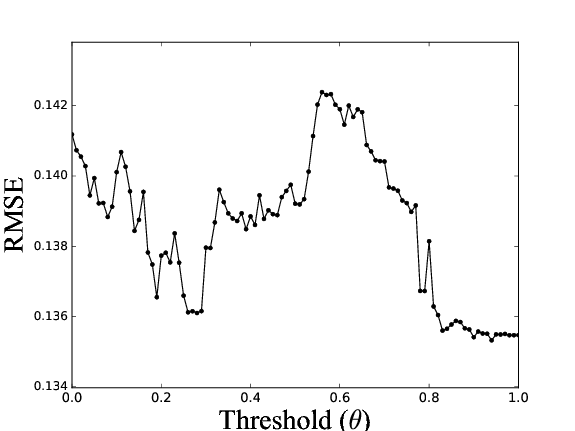
روش­های پیشنهادی با سه روش جایگذاری مقادیر جاافتاده­ی دیگر یعنی جایگذاری با مقدار میانگین، جایگذاری با k نزدیک­ترین همسایگان و IARI [37] مقایسه شدند. در آزمایش­های انجام‌شده از پنج مجموعه‌داده‌ی مختلف و سه معیار ارزیابی مختلف استفاده شد. نتایج نشان داد که در اکثر موارد، عملکرد رویکردهای پیشنهادی بهتر از روش­های مقایسه شده بود­ه است. درنهایت، تأثیر پارامترهای ورودی بر عملکردهای پیشنهادی بررسی شدند. مطالعات آینده می‌تواند شامل ارزیابی عملکرد رویکردهای پیشنهادی بر دیگر مجموعه‌داده‌ها و دیگر مکانیزم­های تولید مقادیر جاافتاده یعنی MCAR و MNAR خواهد بود.

****

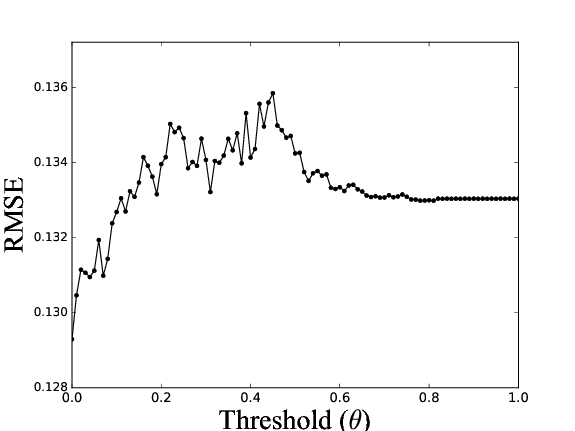
شکل ‏4‑16 تأثیر پارامتر بر مقدار RMSE برای روش 1

****

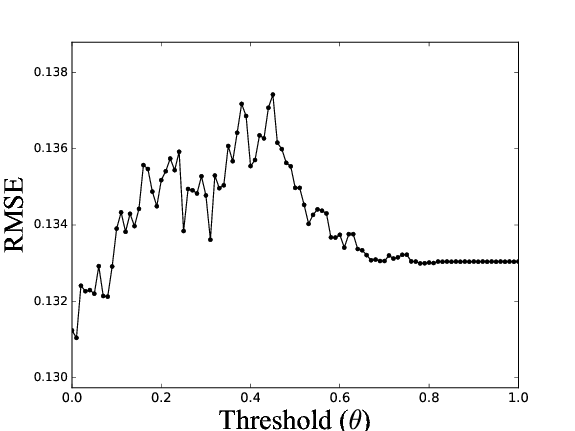
شکل ‏4‑17 تأثیر پارامتر بر مقدار RMSE برای روش 2

****

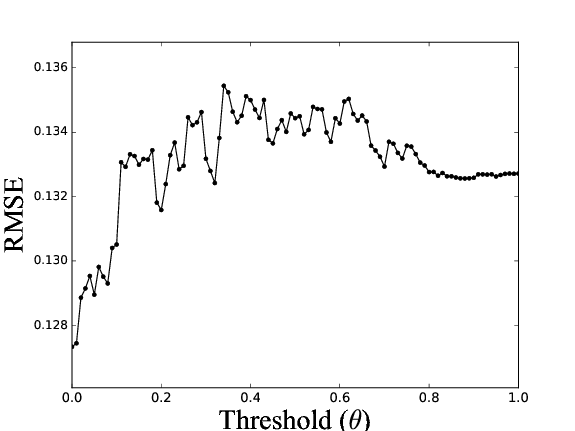
شکل ‏4‑18 تأثیر پارامتر بر مقدار RMSE برای روش 3

****

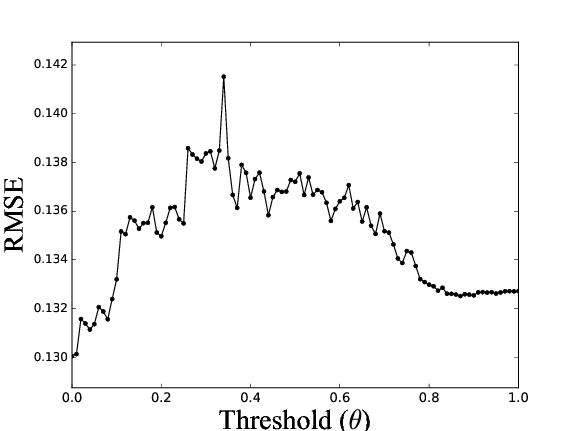
شکل ‏4‑19 تأثیر پارامتر بر مقدار RMSE برای روش 4



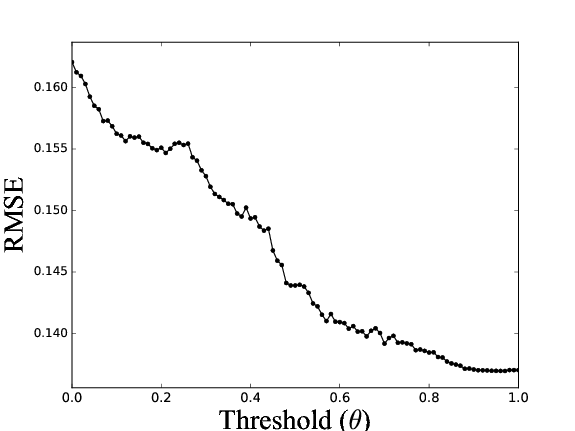
شکل ‏4‑20 تأثیر پارامتر بر مقدار RMSE برای روش 5

****

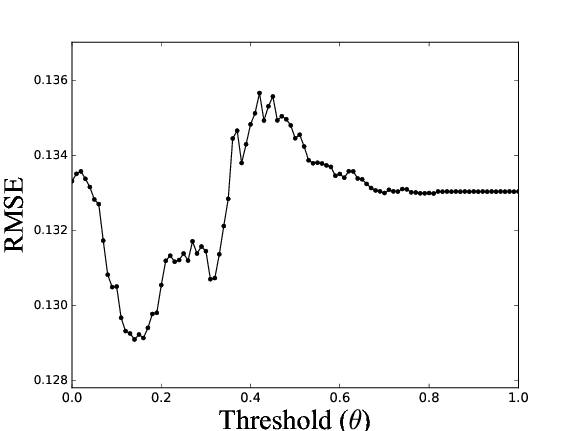
شکل ‏4‑21 تأثیر پارامتر بر مقدار RMSE برای روش6

****

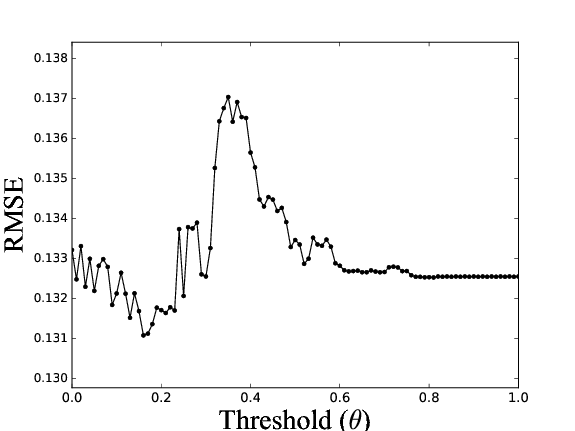
شکل ‏4‑22 تأثیر پارامتر θ بر مقدار RMSE برای روش 7

****

شکل ‏4‑23 تأثیر پارامتر بر مقدار RMSE برای روش 8

****

شکل ‏4‑24 تأثیر پارامتر بر مقدار RMSE برای روش 9

****

شکل ‏4‑25 تأثیر پارامتر بر مقدار RMSE برای روش 10

(این شکلارو هم اگه میتونستی دوتادوتا کنار هم بیاری خوب میشد)



# فصل پنجم: جایگذاری مقادیر جاافتاده با استفاده از خوشه‌بندی فازی c-means مبتنی بر گِرِی و انتخاب ویژگی مبتنی بر معیار اطلاعات متقابل

## مقدمه

در این فصل رویکردی جدید با استفاده از خوشه‌بندی c-means فازی مبتنی بر گِرِی (GFCM) و انتخاب ویژگی مبتنی بر معیار اطلاعات متقابل[[163]](#footnote-163) برای تخمین مقادیر جاافتاده عددی مطرح می­شود. در این روش، از یک نسخه‌ی جدید برای الگوریتم خوشه‌بندی فازی c-means برای هدف افراز بندی داده‌ها استفاده می‌شود. در این نسخه‌ی جدید از مباحث نظریه سیستم‌های‌ گِرِی[[164]](#footnote-164) (GST) که در شرایط با عدم قطعیت کاربرد دارند، استفاده می‌شود. به بیان دقیق‌تر، از معیار نمره‌ی رابطه‌ای گِرِی[[165]](#footnote-165) (GRG) که مزیت‌هایی نسبت به معیارهای شبیه به معیار میندُوسکی[[166]](#footnote-166) دارد، در الگوریتم پیشنهادی به‌جای معیار فاصله‌ی اقلیدسی استفاده‌شده است. به‌علاوه، یک عملیات انتخاب ویژگی مبتنی بر معیار اطلاعات متقابل در هر خوشه به‌صورت محلی برای انتخاب صفت‌هایی با ارتباط بالا با صفت جاافتاده به‌منظور افزایش دقت جایگذاری انجام می‌شود(بشکنیش به دو جمله بهتره. "یک" هم قشنگ نیست. انگلیسی میشه جملت با یک). به‌طور خلاصه، رویکرد پیشنهادی شامل این مراحل است: ابتدا اولویت صفات جاافتاده برای پُرشدن محاسبه می‌شود و یک تخمین اولیه از مقادیر جاافتاده صورت می‌گیرد. سپس، نمونه‌ها به چندین خوشه‌ی فازی تقسیم می‌شوند. آنگاه، خوشه‌هایی که یک شرط حداقلی را ارضا کنند برای ادامه‌ی فرآیند انتخاب می‌شوند. پس‌ازآن، صفاتی که نسبت صفت جاافتاده ارتباط بالایی دارند(بازنویسی)، در هر خوشه توسط یک رویکرد انتخاب ویژگی مبتنی بر اطلاعات متقابل انتخاب می‌شوند. سپس، مدل‌های رگرسیون بر صفات انتخاب شده از نمونه‌های خوشه‌های انتخاب‌شده اعمال می‌شوند و مقادیر جاافتاده تخمین زده می‌شوند. درنهایت، با استفاده از میانگین‌گیری وزن‌دار از تخمین‌های انجام‌شده، مقدار نهایی هر مقدار جاافتاده، جایگذاری خواهد شد.

عملکرد رویکردهای ارائه‌شده بر روی پنج مجموعه­داده­ از دنیای واقعی در زمینه­های مختلف سنجیده می­شود. از سه معیار شناخته‌شده ریشه میانگین مربعات خطا، ضریب تعیین و میانگین خطای مطلق، برای سنجش کیفیت جایگذاری مقادیر جاافتاده استفاده‌شده است. عملکرد رویکرد ارائه‌شده با پنج رویکرد دیگر یعنی روش جایگذاری با مقدار میانگین، جایگذاری با استفاده از نزدیک­ترین همسایگان (kNN)، جایگذاری با شبکه عصبی پرسپترون چندلایه (MLP) [25]، جایگذاری با خوشه‌بندی فازی c-means (FCM) [89] و روش جایگذاری با رگرسیون افزایشی صفات (IARI) [37] مقایسه شده­ است. نتایج آزمایش نشان می­دهد که در اکثر موارد، رویکرد ارائه‌شده، بهتر از دیگر روش­های مقایسه شده عمل می­کنند.

## مفاهیم پیش‌نیاز

در این بخش مفاهیمی که از آن‌ها در رویکرد پیشنهادی استفاده‌شده است، به‌طور خلاصه بیان می‌شوند.

### خوشه‌بندی c-means فازی

خوشه‌بندی، روشی بدون ناظر برای تقسیم داده‌ها به گروه‌هایی مبتنی بر شباهت نمونه‌ها است. خوشه‌بندی را می‌توان به دودسته‌ی خوشه‌بندی سخت[[167]](#footnote-167) و یا خوشه‌بندی فازی (نرم)[[168]](#footnote-168) تقسیم‌بندی کرد [10]. در روش خوشه‌بندي سخت، پس از خوشه‌بندي، هر داده دقيقأ به يک خوشه تعلق مي‌گيرد، مانند روش خوشه‌بندي k-means؛ ولي در خوشه‌بندي فازی پس از خوشه‌بندي، به هر نمونه يک درجه تعلق (بین صفر و یک) به ازای هر خوشه، نسبت داده مي‌شود. به‌عبارت‌دیگر، يک نمونه مي‌تواند با نسبت‌هاي متفاوتي عضو چندين خوشه باشد. مزیت‌هایی برای خوشه‌بندی فازی نسبت به خوشه‌بندی عادی k-means وجود دارد. در بسیاری از شرایط در دنیای واقعی، رفتار خوشه‌بندی فازی بسیار طبیعی‌تر است. چراکه این خوشه‌بندی به‌طور بهتری رفتار نمونه‌هایی را که به‌طور خوبی قابل جدا شدن نیستند مدل می‌کند. این دسته از خوشه‌بندی‌ها به دلیل ماهیت داده‌های جاافتاده (که به‌طور کامل معلوم نیستند)، برای هدف جایگذاری مقادیر جاافتاده، بهتر از روش‌های سخت هستند [19]. به‌علاوه، خوشه‌بندی عادی k-mean ممکن است در حالتی که نقاط اولیه به‌طور مناسب انتخاب نشوند در شرایط کمینه محلی گیر کند. اگرچه در حالت فازی، مقادیر عضویت پیوسته برای نمونه‌ها حساسیت الگوریتم‌ها را برای گیرکردن در شرایط کمینه‌ محلی کمتر می‌کند [19].

خوشه‌بندی فازی c-means معروف‌ترین روش خوشه‌بندی فازی است. این خوشه‌بندی مجموعه‌ای از نمونه‌های ورودی مثل را به c خوشه‌ی فازی تقسیم می‌کند. این کار با کمینه کردن تابع هدف () مطابق با رابطه‌ی ‏5‑1 انجام می‌شود.

|  |  |
| --- | --- |
| ‏5‑1 |  |

*در این رابطه بیانگر عملگر نُرم اقلیدسی است. ماتریس مراکز خوشه‌ها است و نشان‌دهنده مرکز خوشه‌ی k ام است. پارامتر فازی‌سازی*[[169]](#footnote-169) *است که تعیین می‌کند که خوشه‌ها چقدر با یکدیگر همپوشانی داشته باشند. مقدار آن به‌صورت تجربی معمولاً برابر با 5/1 در نظر گرفته می‌شود. بیانگر ماتریس عضویت است و مقدار نشان می‌دهد که نمونه‌ی به چه میزان به خوشه‌ی تعلق دارد. هرچه مقدار بیش‌تر باشد، با احتمال بیش‌تری به آن خوشه تعلق دارد. مجموع تعلقات یک نمونه به تمام خوشه‌ها همواره برابر با 1 است، یعنی: (اینجا هم به جای یعنی، مستقیما ارجاع بده به فرمول)*

|  |  |
| --- | --- |
| ‏5‑2 |  |

در فرآیند خوشه‌بندی فازی c-means، تابع هدف به‌طور تکراری با به‌روزرسانی ماتریس‌ عضویت و ماتریس‌ مراکز خوشه‌ها بهینه‌سازی می‌شود. شرایط لازم برای کمینه کردن ‏5‑1 با در نظر گرفتن محدودیت ‏5‑2، استفاده از معادلات ‏5‑3 و ‏5‑4 به ترتیب برای به‌روزرسانی ماتریس‌ عضویت و مراکز خوشه‌ها است.

|  |  |
| --- | --- |
| ‏5‑3 |  |
| ‏5‑4 |  |

*دقت موردنیاز، تعداد تکرارهای الگوریتم را تعیین می‌کند. این دقت، از میزان اختلاف بین دو تکرار پیاپی برای ماتریس عضویت طبق رابطه‌ی* ‏5‑6 *محاسبه می‌شود.*

|  |  |
| --- | --- |
| ‏5‑5 |  |

*که در آن و به ترتیب نشان‌دهنده ماتریس عضویت در تکرارهای r و r+1 ام هستند. در این رابطه بیانگر عملگر نُرم ماتریس است. تکرار الگوریتم هنگامی متوقف می‌شود که میزان اختلاف بین دو تکرار پیاپی برای ماتریس عضویت کمتر از یک حد آستانه تعریف‌شده مثل باشد. همان‌طور که پیداست، نسخه‌ی اصلی فازی* c-means از فاصله اقلیدسی به‌عنوان معیار سنجش فاصله استفاده می‌کند. در روش پیشنهادی در این فصل از نسخه‌ی متفاوتی از فازی c-means برای هدف جایگذاری مقادیر جاافتاده استفاده خواهد شد که در بخش‌های بعدی به آن اشاره خواهد شد. (به نظر من این بخش جاش خوبه که نزدیک کاربردشه اما امیدوارم داورها هم با من هم عقیده باشند.)

### تحلیل رابطه‌ای گری

نظریه سیستم گری (GST) برای تعامل با سیستم‌هایی با عدم قطعیت که شامل اطلاعات نامعلوم جزئی[[170]](#footnote-170) هستند، اولین بار در [96] توسط معرفی شد(ناقصه. اصلا لازمم نیست بگی توسط کی). این نظریه می‌تواند اطلاعات ارزشمندی را از داده‌های ناکامل استخراج کند؛ بنابراین به‌طور گسترده‌ برای حل مسائلی که شامل عوامل نامعلوم هستند، مورداستفاده قرارگرفته است [97]. تحلیل رابطه‌ای گری[[171]](#footnote-171) (GRA) به‌عنوان یک روش اندازه‌گیری در نظریه سیستم گری می‌تواند رابطه‌ی بین یک مشاهده‌ی (نمونه) مرجع[[172]](#footnote-172) و یک مجموعه از مشاهده‌های مقایسه شده[[173]](#footnote-173) را با استفاده از نمره‌ی رابطه‌ای گری (GRG) و ضریب رابطه‌ای گری[[174]](#footnote-174) (GRC) تعیین کند. مجموعه‌ی مشاهدات را که در آن مشاهده‌ی مرجع و مشاهدات مقایسه شده هستند، در نظر بگیرید. هر مشاهده مثل با *m* صفت به صورت نمایش داده‌ می‌شود. مقدار GRC بین دو نمونه از طریق ‏5‑6 محاسبه می‌شود.

|  |  |
| --- | --- |
| ‏5‑6 |  |

که در آن برای است. ضریب تمایز نام دارد که سطح تغییرات را با توجه به ضریب رابطه‌ای تعیین می‌کند و مقدار آن به‌صورت تجربی برابر با 5/0 قرار می‌گیرد [36].

مقدار GRC بین 0 و 1 است. هرچقدر مقدار بزرگ‌تر باشد، شباهت بین و بیش‌تر خواهد بود و بلعکس. درنهایت، مقدار GRG بین دو نمونه از طریق ‏5‑7 قابل‌محاسبه است.

|  |  |
| --- | --- |
| ‏5‑7 |  |

*که در آن*  است. مقدار GRG میزان شباهت دو نمونه بر روی مجموعه‌ای از صفات را نشان می‌دهد و مقداری بین 0 و 1 اختیار می‌کند. اگر باشد، آنگاه نمونه‌ی بیش‌تر از نمونه‌ی به نمونه‌ی مرجع () شبیه‌ است. نمونه‌ای با بیش‌ترین مقدار GRG، شبیه‌ترین نمونه به نمونه‌ی مرجع () است.

به‌طورکلی توابع شباهت ازجمله فاصله اقلیدسی برای تعیین میزان نزدیکی بین دو نمونه استفاده می‌شود. اگرچه، فواصلی مثل فاصله اقلیدسی معمولاً برای کاربردهایی در دامنه‌هایی خاص مناسب هستند. معیار GRG نسبت به معیارهایی مثل فاصله اقلیدسی مزیت‌هایی دارد. ازجمله‌ی این مزایا می‌توان به نرمال بودن و تمامیت[[175]](#footnote-175) اشاره کرد. تمامیت به این معنا است که این معیار ارتباط بین دو نمونه را با در نظر گرفتن کل فضای مجموعه‌ی مشاهدات اندازه می‌گیرد؛ درحالی‌که در معیارهایی مثل فاصله‌ی اقلیدسی این‌گونه نیست و فاصله هر دو نمونه مستقل از دیگر نمونه‌ها محاسبه می‌شود [97]. با توجه به برتری معیار GRG نسبت به معیار معمول فاصله اقلیدسی (به‌ویژه در شرایطی با مقادیر جاافتاده)، نسخه‌ی جدیدی از خوشه‌بندی فازی c-means که از این معیار استفاده می‌کند، در ادامه ارائه می‌شود. از معیار GRG برای سنجش نزدیکی نمونه‌ها نسبت به مراکز خوشه‌ها در حین فرآیند به‌روزرسانی ماتریس عضویت استفاده خواهد شد.

### معیار اطلاعات متقابل

در نظریه احتمالات و نظریه اطلاعات، اطلاعات متقابل بین دو متغیر تصادفی مثل X و Y که با I(X;Y) نمایش داده می­شود، معیاری برای نشان دادن میزان وابستگی متقابل آن دو متغیر است. به بیان ساده، این معیار میزان اطلاعات به‌دست‌آمده در مورد یک متغیر تصادفی را با دانستن مقدار متغیر تصادفی دیگر نشان می‌دهد. مزیت اصلی این معیار این است که هیچ فرضی درباره‌ی نوع ارتباط بین صفات مثل این‌که ارتباط خطی است یا خیر، نمی‌کند. بنابراین این معیار قابلیت تعمیم بیشتری نسبت معیارهایی مثل ضریب همبستگی خطی دارد [98,99].

در دامنه‌ی اطلاعات گسسته، فرمول محاسبه اطلاعات متقابل برای دو متغیر تصادفی X و Y طبق معادله‌ی ‏5‑8 است.

|  |  |
| --- | --- |
| ‏5‑8 |  |

که در آن و به ترتیب تابع جرم احتمال[[176]](#footnote-176) (pmf) متغیرهای X و Y است. نشان‌دهنده‌ی تابع جرم احتمال توأم دو متغیر X و Y است. برای دامنه‌ی پیوسته فرمول محاسبه اطلاعات متقابل برای دو متغیر تصادفی X و Y طبق معادله‌‌ی ‏5‑9 خواهد بود.

|  |  |
| --- | --- |
| ‏5‑9 |  |

که در آن و به ترتیب تابع چگالی احتمال[[177]](#footnote-177) (pdf) متغیرهای X و Y است. نشان‌دهنده‌ی تابع چگالی احتمال توأم دو متغیر X و Y است.

با محاسبه مقدار اطلاعات متقابل برای هر جفت از صفات یک مجموعه‌داده‌ مثل ، یک ماتریس اطلاعات متقابل با اندازه‌ی می‌توان ساخت که مدخل در این ماتریس، مقدار اطلاعات متقابل بین صفات و را در نشان می‌دهد. در ادامه، نماد ماتریس اطلاعات متقابل برای مجموعه‌داده را نشان می‌دهد. نماد اشاره به مدخل ردیف ام و ستونام از ماتریس اطلاعات متقابل برای مجموعه‌داده خواهد داشت. ماتریس اطلاعات متقابل به شکل ‏5‑10 خواهد بود.

|  |  |
| --- | --- |
| ‏5‑10 |  |

که در آن، میزان اطلاعات متقابل بین مقادیر دو صفت و یعنی I(;) را نشان می‌دهد.

در رویکرد پیشنهادی از ماتریس اطلاعات متقابل برای انجام عملِ انتخاب ویژگی استفاده خواهد شد.

## جزئیات مراحل پیشنهادی

در این بخش و زیربخش­های آن مراحل مختلف رویکردهای پیشنهادی برای جایگذاری مقادیر جاافتاده شرح داده می­شوند.

فرض کنید مجموعه‌داده‌ی‌‌ ناکاملِ شامل *n* نمونه است و هر نمونه مثل شامل *m* صفت عددی است. مقدار صفت ام را در رکورد ام در نشان می‌دهد. شکل ‏5‑1 رویه‌ی کلی برای روش‌ ارائه‌شده در این فصل را نمایش می‌دهد(این شکله خیییلی دوره. بهتره کلیاتشم اینجا بگی و جزئیاتش بمونه تو زیربخشها). زیربخش‌های بعدی جزئیات مراحل رویکرد پیشنهادی را شرح می‌دهند.

### محاسبه اولویت هر صفت جاافتاده برای فرآیند تخمین داده‌های جاافتاده

صفات جاافتاده در روش‌های پیشنهادی تک‌به‌تک به ترتیب اهمیتشان طبق رویکرد ارائه‌شده توسط [37] پر خواهند شد. صفتی با بیش‌ترین اهمیت ابتدا و صفتی با کمترین اهمیت در انتها پر خواهد شد. اولویت هر صفت جاافتاده برای پرشدن در سه گام محاسبه می‌شود. ابتدا مقادیر جاافتاده در مجموعه‌داده‌ی ورودی (شامل مقادیر جاافتاده) با استفاده از یک رویکرد تخمین ساده، مثل تخمین با مقدار میانگین، پر می‌شوند. سپس، یک مدل جنگل تصادفی بر روی مجموعه‌داده‌ی پرشده ساخته می‌شود. درنهایت، مدل ساخته‌شده به نمونه‌های خارج از کیسه اعمال می‌شود ("کیسه" دیگه چیه؟) تا اهمیت هر صفت اندازه‌گیری شود. جزئیات این کار در [92] شرح داده‌شده است. پس از پیدا کردن ترتیب هر صفت برای پرشدن، نمونه‌هایی که صفتی با بیش‌ترین اهمیت در آن‌ها جاافتاده است، برای ادامه فرآیند پرکردن انتخاب خواهد شد. بنابراین با فرض این­که اندیس مهم‌ترین صفت جاافتاده باشد، آنگاه نمونه‌های که در صفتام خود مقدار جاافتاده دارند تک‌به‌تک برای ادامه در مراحل بعدیِ فرآیندِ جایگذاری انتخاب خواهند شد و درنهایت پر خواهند شد. (توضیحاتت با ارجاع به خطوط کد باشه)

### جایگذاری اولیه، خوشه‌بندی داده‌ها با استفاده از فازی c-means مبتنی بر گِرِی و عملیات انتخاب خوشه

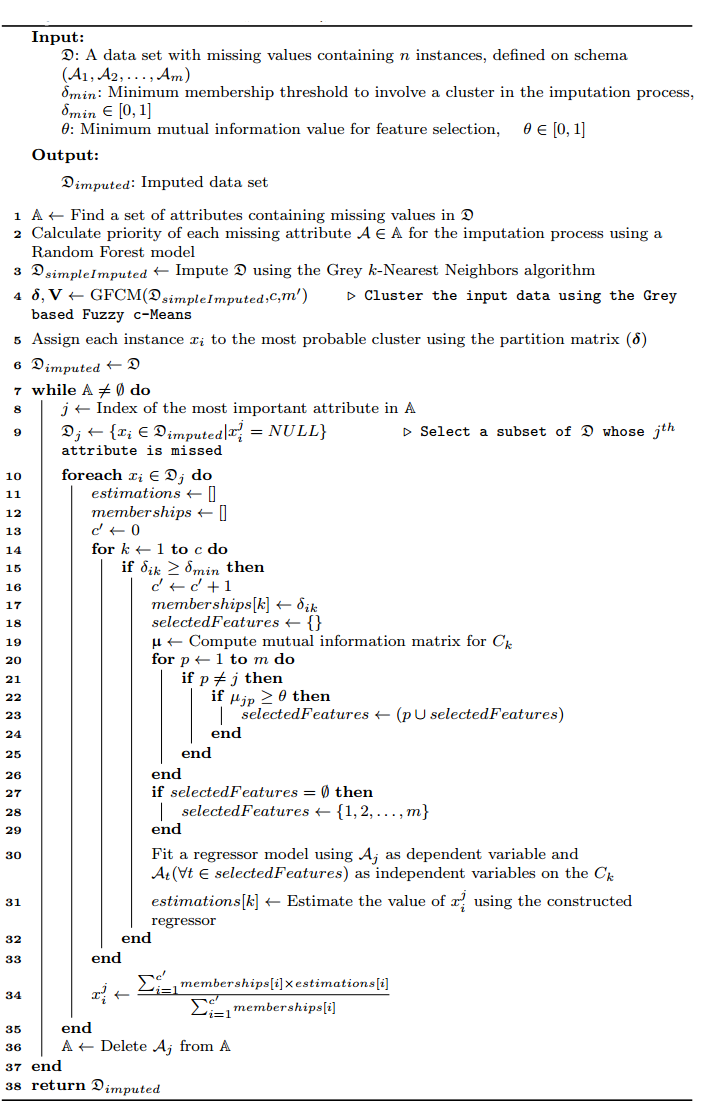
تا این مرحله، ترتیب تخمین هریک از صفات جاافتاده تعیین شد. از این مرحله به بعد، نحوه پرکردن مقادیر جاافتاده برای هر نمونه‌ی جاافتاده شرح داده خواهد شد. در رویکرد ارائه‌شده، هر مقدار جاافتاده در یک نمونه‌ی جاافتاده با استفاده از مقادیر تخمین زده‌شده توسط هریک از خوشه‌های فازیِ تشکیل‌شده، جایگذاری خواهند شد. این مقادیر از اعمال مدل‌های رگرسیون درون خوشه‌های تشکیل‌شده حاصل می‌شوند. بنابراین ابتدا داده‌ها باید خوشه‌بندی شوند. همان‌طور که گفته شد، برای این‌ منظور، از خوشه‌بندی فازی جدیدِ پیشنهادی به نام فازی c-means مبتنی بر گِرِی[[178]](#footnote-178) (GFCM) در این روش استفاده خواهد شد. همانند اکثر روش‌های یادگیری ماشین، خوشه‌بندی نیز قابلیت اعمال مستقیم به یک مجموعه‌داده‌ ناکامل را ندارد. بنابراین، ابتدا یک جایگذاری اولیه برای رفع این مشکل انجام خواهد شد؛ یعنی در ابتدا مقادیر جاافتاده در مجموعه‌داده‌ی‌ ناکاملِ ورودی توسط یک روش پایه پر می‌شوند. برای این منظور، از روش جایگذاری با مقدار میانگین استفاده‌شده است.

پس از انجام جایگذاری اولیه، داده‌ها توسط الگوریتم خوشه‌بندی GFCM به c خوشه‌ی فازی تقسیم خواهند شد. الگوریتم GFCM نسخه‌ی جدید از FCM است که از معیار GRG برای انتساب هر نمونه‌ به هر خوشه‌ی فازی بهره می‌گیرد. شکل ‏5‑2 مراحل این الگوریتم را نشان می‌دهد(حتما توضیحاتتو با ارجاع به خطوط کدها تکمیل کن. در واقع خواننده رو باید در خواندن خطوط کدها همراهی کنی). ابتدا ماتریس عضویت و مراکز خوشه‌ها به‌صورت تصادفی مقداردهی اولیه می‌شوند. سپس، الگوریتم به‌طور تکراری سعی می‌کند تا مجموع فواصل هر نمونه از مراکزی که به آن تعلق دارد را با استفاده از به‌روزرسانی ماتریس افرازبندی کمینه کند. این کار آن‌قدر ادامه می‌یابد تا تغییرات ماتریس عضویت از حد آستانه کمتر باشد و یا این‌که تعداد تکرارهای الگوریتم به حداکثر خود برسد. همان‌طور که گفته شد، به‌طور تجربی ثابت‌شده است که GRG، بهتر از فاصله‌ی اقلیدسی برای سنجش فاصله در شرایط با عدم قطعیت، عمل می‌کند. از دلایل این امر می‌توان به تمامیت و نرمال بودن اشاره کرد. بنابراین در الگوریتم ارائه‌شده از معیار GRG برای سنجش شباهت بین مراکز خوشه و نمونه‌های ورودی استفاده می‌شود. در الگوریتم FCM، از مقدار فاصله‌ی دو نمونه برای به‌روزرسانی ماتریس عضویت استفاده می‌شود، درحالی‌که معیار GRG مقدار شباهت دو نمونه را اندازه می‌گیرد. بنابراین از مقدار (1-GRG) در روابط استفاده می‌شود.

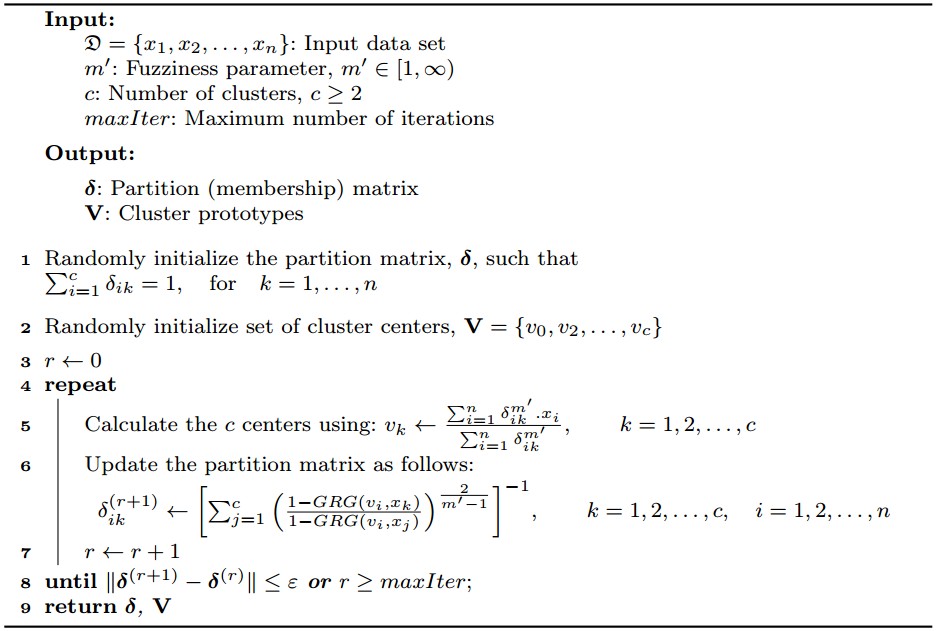
پس از پایان فرآیند خوشه‌بندی، یک ماتریس عضویت حاصل خواهد شد که میزان تعلق هر نمونه به هر خوشه را توصیف می‌کند. به‌علاوه، هر نمونه‌ می‌تواند به یک خوشه که بیش‌ترین احتمال تعلق به آن را دارد، منتسب شود. یعنی با داشتن ماتریس عضویت، هر نمونه مثل را می‌توان به خوشه‌ی ، که در آن از رابطه‌ی ‏5‑11 محاسبه می‌شود، منتسب کرد.

|  |  |
| --- | --- |
| ‏5‑11 |  |

در گام بعدی، هر نمونه‌ی جاافتاده با استفاده از اطلاعات نمونه‌های خوشه‌‌های فازی تشکیل‌شده پر خواهد شد. اگرچه در روش‌ پیشنهادی از تمام خوشه‌های تشکیل‌شده استفاده نخواهد شد؛ بلکه یک نمونه‌ی جاافتاده فقط با استفاده از اطلاعات خوشه‌هایی تخمین زده خواهد شد که با یک حداقل درجه‌ی عضویت به آن خوشه‌ها تعلق دارد. دلیل این کار این است که استفاده از داده‌های خوشه‌هایی که به نمونه‌ی جاافتاده ارتباط زیادی دارند احتمالاً منجر به بهبود تخمین یک مقدار جاافتاده می‌شود. حداقل درجه‌ی عضویت با پارامتر در الگوریتم پیشنهادی تعیین می‌شود. مقدار مناسب برای به‌طور مستقیم به تعداد خوشه‌ها یعنی c، بستگی دارد. در این پژوهش، مقدار برابر با تنظیم می‌شود. این مقداردهی تضمین می‌کند که همه‌ی خوشه‌ها شانس شرکت در تخمین یک مقدار جاافتاده را دارند(چرا؟). به‌طور خلاصه، خروجی این گام، زیرمجموعه‌ای از خوشه‌های فازی تشکیل‌شده است که شرط ذکرشده را ارضا می‌کنند.



شکل ‏5‑1 رویه کلی رویکرد پیشنهادی برای تخمین مقادیر جاافتاده



شکل ‏5‑2 مراحل الگوریتم فازی c-means مبتنی بر گری (GFCM)

### انتخاب ویژگی در هر خوشه‌ی انتخاب‌شده

پس از انتخاب خوشه‌های مطلوب، یک رویکرد انتخاب ویژگی مبتنی بر اطلاعات متقابل در هر خوشه، مستقل از دیگر خوشه‌ها انجام می‌شود. درواقع، از تمام صفاتِ نمونه‌های درون یک خوشه برای تخمین یک مقدار جاافتاده استفاده نمی‌شود؛ بلکه فقط از زیرمجموعه‌ای از صفات که ارتباط زیادی با صفت جاافتاده دارند برای تخمین مقادیر جاافتاده استفاده می‌شود. برای این منظور فقط صفاتی که یک حداقل وابستگی نسبت به صفت جاافتاده دارند، انتخاب می‌شوند. این وابستگی از طریق معیار اطلاعات متقابل سنجیده می‌شود. حداقل اطلاعات متقابل برای انتخاب شدن یک ویژگی توسط یک حد آستانه که با نمایش داده می‌شود، تعیین می‌شود. به بیان دقیق‌تر، با توجه به ماتریس اطلاعات متقابلِ درون هر خوشه، فقط صفاتی برای ادامه‌ی فرآیند جایگذاری از آن خوشه انتخاب می‌شوند که میزان اطلاعات متقابل آن‌ها نسبت به صفت جاافتاده، بیش‌تر از باشد. در شرایطی که هیچ‌یک از صفات شرط لازم را ارضا نکنند (هنگامی‌که مقدار بسیار بالا انتخاب شود)، تمام صفات انتخاب خواهند شد. یک مزیت اصلی برای انتخاب ویژگی در هر خوشه به‌طور مجزا از دیگر خوشه‌ها نسبت به حالتی که یک‌بار انتخاب ویژگی در کل داده‌ها انجام شود، وجود دارد(منظورت جمله بعده؟). پس از فرآیند خوشه‌بندی، نمونه‌های درون یک خوشه ممکن است بر روی زیرمجموعه‌ای از صفات، همبستگی بالا داشته ‌باشند، درحالی‌که در خوشه‌ی دیگر برای آن زیرمجموعه این‌چنین نباشد؛ بنابراین در این شرایط مناسب است که هر خوشه به‌طور جدا بررسی شود و عمل انتخاب ویژگی انجام شود.

### اعمال مدل رگرسیون بر هریک از خوشه‌ها و صفات انتخاب‌شده

گام بعدی در الگوریتم پیشنهادی، تخمین مقادیر جاافتاده توسط صفات انتخاب‌شده، درون هر خوشه‌ی منتخب است. الگوریتم، هریک از مقادیر جاافتاده را با استفاده از میانگین وزنی تخمین‌های حاصل از هر خوشه پر می‌کند. بنابراین اگر به تعداد خوشه‌ از مرحله‌ی قبل برای جایگذاری یک مقدار جاافتاده مثل انتخاب شوند، آنگاه تخمین برای انجام خواهد شد. هریک از این تخمین‌ها با اعمال مدل رگرسیون بر صفات انتخاب‌شده برای هریک از نمونه‌های یک خوشه به دست می‌آید. از دو مدل رگرسیون برای رویکرد پیشنهادی در این مرحله استفاده می‌شود: رگرسیون خطی چند صفتی[[179]](#footnote-179) (MLR) و رگرسیون بردار پشتیبان[[180]](#footnote-180) (SVR)(چرا از این دوتا استفاده کردی؟ چراهای انتخابهاتو پررنگ کن). معادله‌ی رگرسیون خطی چند صفتی مطابق با ‏5‑3 است. رگرسیون بردار پشتیبان، یک الگوریتم مبتنی بر هسته[[181]](#footnote-181) است که می‌توان از آن به‌عنوان رگرسیون غیرخطی استفاده کرد. درواقع، هر نمونه‌ی ورودی مثل با استفاده از یک نگاشت ویژگی غیرخطی[[182]](#footnote-182) مثل به یک فضای ویژگی مثل که از هسته‌ به‌دست‌آمده است، نگاشت می‌شود و در آن فضا رگرسیون خطی انجام می‌شود. به بیان دقیق‌تر، مدل SVR تابع رگرسیونی را مطابق با معادله‌ی ‏5‑12 با استفاده از تخمین مقادیر و از داده‌های ورودی یاد می‌گیرد [100].

|  |  |
| --- | --- |
| ‏5‑12 |  |

که در آن نشان‌دهنده‌ی ضرب داخلی در فضای است. حل معادله بالا برای داده‌ها به‌وسیله روش‌های بهینه‌سازی و به‌ویژه برنامه‌ریزی درجه دوم[[183]](#footnote-183) که روش‌های شناخته‌شده‌ای در حل مسائل محدودیت‌دار است، صورت می‌گیرد.

با توجه به مطالب ذکرشده، خروجی این گام مجموعه‌ای از مقادیر تخمین زده‌شده برای مقدار جاافتاده‌ی است که از اعمال مدل‌های رگرسیون بر داده‌های انتخاب‌شده از خوشه‌های انتخاب‌شده (در گام‌های قبل)، حاصل‌شده‌اند.

### انجام نظرخواهی وزن‌دار برای تخمین نهایی

آخرین گام در الگوریتم پیشنهادی، تخمین و جایگذاری نهایی یک مقدار جاافتاده مثل است. این کار با استفاده از رأی‌گیری وزن‌دار بر روی مقادیر حاصل‌شده از مرحله‌ی قبل حاصل می‌شود. وزن یک مقدار تخمین زده‌شده توسط یک خوشه برای اثر دهی در تخمین نهایی، برابر است با این‌که نمونه جاافتاده به چه میزانی به آن خوشه تعلق دارد؛ یعنی مقدار تخمین نهایی از میانگین وزن‌دار تخمین انجام‌شده در مرحله‌ی قبل حاصل می‌شود و وزن تخمین *k* ام () برابر با خواهد بود.

تکرار‌های الگوریتم ارائه‌شده آن‌قدر ادامه می‌یابد تا تمام صفات جاافتاده جایگذاری شوند. همان­طور که الگوریتم پیشنهادی نشان می‌دهد، پس از پرکردن هر صفت جاافتاده، مقادیر آن صفت برای تخمین دیگر صفات جاافتاده، در تکرارهای بعدی استفاده خواهد شد. درواقع، مجموعه‌داده‌ی ورودی پس از هر تکرار حلقه **،** به‌روزرسانی می‌شود.

## آزمایش‌های تجربی

در این قسمت، روش‌های مورد مقایسه با روش‌ پیشنهادی(منظورت مقایسه دیگر روشها با روش پیشنهادیه؟)، مجموعه‌داده‌های استفاده‌شده برای آزمایش‌ها و معیارهای ارزیابی کارایی رویکردهای متفاوت بررسی می‌شوند. علاوه بر این، آزمایش‌هایی برای بررسی تأثیر پارامتر‌های c و بر روی نتایج تخمین نهایی انجام می‌شود. عملکرد رویکرد پیشنهادی با پنج روشِ تخمین یعنی روش جایگذاری با مقدار میانگین، جایگذاری با استفاده از نزدیک­ترین همسایگان (kNN)، جایگذاری با شبکه عصبی پرسپترون چندلایه (MLP) [25]، جایگذاری با خوشه‌بندی فازی c-means (FCM) [89] و روش جایگذاری با رگرسیون افزایشی صفات (IARI) [37] مقایسه شده ­است.

### طراحی آزمایش‌ها

به‌منظور حذف عامل تصادفی بودن در آزمایش­ها، هر آزمایش 20 بار تکرار شده است. میانگین این 20 بار تکرار به‌عنوان نتیجه نهایی در این پژوهش گزارش‌شده‌اند. کدهای مورداستفاده برای رویکردهای مختلف همگی با زبان برنامه­نویسی پایتون و با کمک بعضی از بسته­های یادگیری ماشین از مخزن Scikit-Learn [94] پیاده­سازی شده­اند. همان‌طور که گفته شد، برای گام‌ تخمین در هر خوشه از دو نوع رگرسیون خطی و رگرسیون بردار پشتیبان (SVR) استفاده می‌شود. بنابراین در مقایسات انجام‌شده، دو روش پیشنهادی ارائه می‌شود. در پیاده­سازی­ها از رگرسیون خط‌الرأس که نوعی از رگرسیون خطی است، استفاده‌شده است. در این رگرسیون از تابع حداقل مربعات خطی به‌عنوان تابع هزینه­ استفاده‌شده است. همچنین از تنظیم‌کردنِ با پارامتر تنظیمِ استفاده‌شده است. در رگرسیون SVR، مقادیر پارامترهای جریمه[[184]](#footnote-184) و اپسیلون[[185]](#footnote-185) از مقادیر پیش‌فرض و به ترتیب برابر با و استفاده شد. همچنین از هسته‌ی خطی[[186]](#footnote-186) برای یادگیری در این مدل رگرسیون استفاده‌شده است.

#### مجموعه‌داده‌های مورد آزمایش

روش­های مختلف جایگذاریِ مقادیر جاافتاده بر روی پنج مجموعه‌داده‌ی از دنیای واقعی موجود در مخزن داده­ی یادگیری ماشین UCI مطابق با جدول ‏3‑1، اعمال شدند. این مجموعه‌داده‌ها بدون مقادیر جاافتاده می‌باشند. در این مجموعه‌داده­های کامل، ابتدا مقادیر جاافتاده به‌طور مصنوعی ایجاد شدند. سپس مقادیر جاافتاده با استفاده از روش­های مختلف جایگذاری پُر شدند. ازآنجایی‌که مقادیر اصلی برای مقادیر جاافتاده­ی ساخته‌شده معلوم است، بنابراین عملکرد رویکردهای مختلف با دانستن مقادیر واقعی قابل ارزیابی است.

#### پیش‌پردازش داده‌ها و نحوه ایجاد مقادیر جاافتاده

برای این­که امکان مقایسه نتایج بین رویکردهای مختلف وجود داشته باشد و مقایسه­­ها بامعنا باشند، مقادیر هر صفت از مجموعه‌داده‌های با استفاده از نرمال­سازی به روش کمینه-بیشینه، مطابق با معادله‌ی ‏3‑9 نرمال شدند.

به‌منظور ایجاد مقادیر جاافتاده در مجموعه‌داده‌ها ابتدا نیمی از صفات به‌طور تصادفی انتخاب شدند. سپس در 30%، 40%، 50% و 60% از نمونه­ها مقادیر جاافتاده به‌طور تصادفی ایجاد شد. بنابراین هر نمونه­ی جاافتاده ممکن است بین 1 تا 2/*m* صفت جاافتاده داشته باشد. همه­ی مقادیر جاافتاده مطابق با الگوی جاافتاده به‌طور تصادفی (MAR) ایجاد شدند.

#### معیارهای ارزیابی

عملکرد رویکردهای جایگذاری مقادیر جاافتاده در مقایسه با دیگر روش­ها توسط سه معیار شناخته­شده­ی ذکرشده در بخش ‏2-5 یعنی ریشه میانگین مربعات خطا (RMSE)، ضریب تعیین () و میانگین خطای مطلق (MAE) سنجیده می­شود.

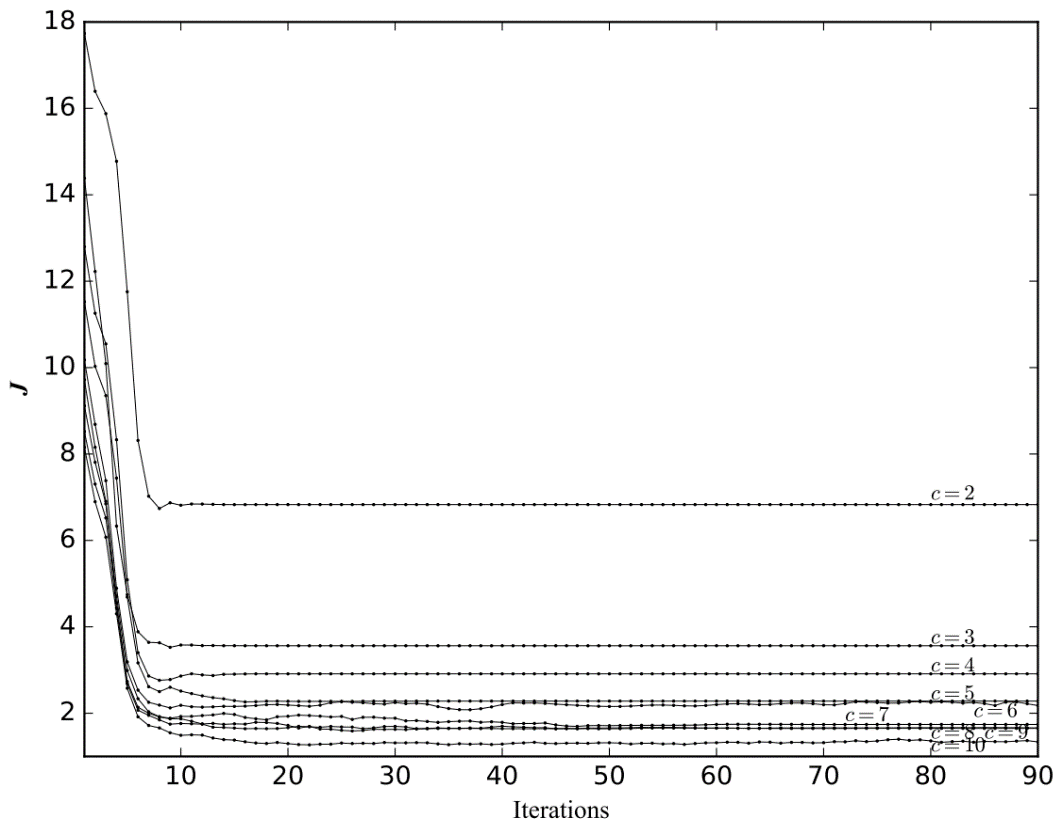
### آزمایش‌های انجام‌شده

#### بررسی عملکرد الگوریتم خوشه‌بندی فازی c-means مبتنی بر گِری

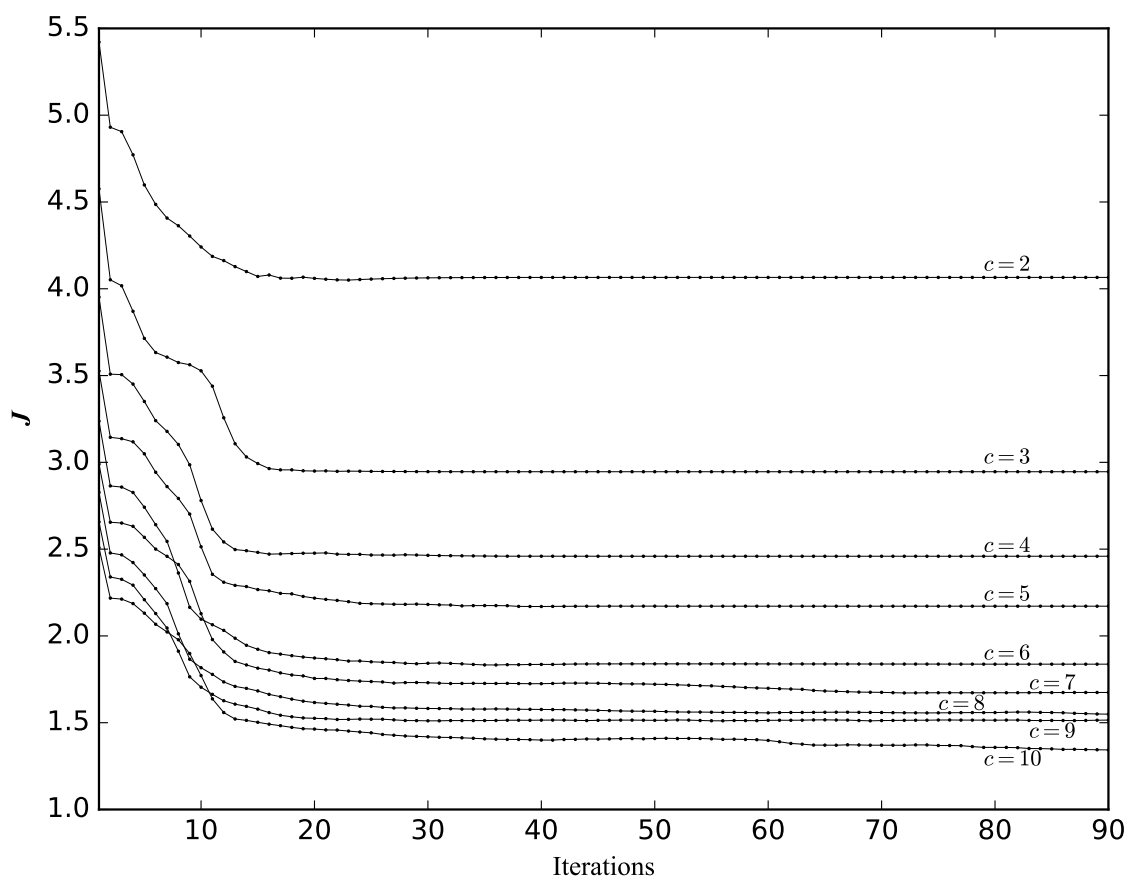
برای بررسی صحت عملکرد و همگرایی الگوریتم خوشه‌بندی پیشنهادی، مقادیر تابع هدف پس از تعداد تکرارهای مختلف، در این بخش بررسی می‌شود. برای این منظور، مقادیر تابع هدف () به ازای دفعات تکرار الگوریتم GFCM، به ازای مقادیر محاسبه و ثبت شدند. شکل ‏5‑3 تا شکل ‏5‑7 مقادیر تابع هدف () را برای دفعات تکرار 0 تا 90 و تعداد خوشه‌های تا را در مجموعه‌داده‌های مختلف نمایش می‌دهد.

همان‌طور که پیداست، برای همه مجموعه‌داده‌ها پس گذشت از انجام تعداد کمی تکرار از اجرای الگوریتم (حدود ده تکرار)(بازنویسی)، مقدار تابع هدف به میزان زیادی، تقریباً به‌صورت نزولی یکنوا کاهش‌یافته است. این تغییرات برای تمام مقادیر به همین‌گونه بوده ‌است. برای همه‌ی مجموعه‌داده‌‌ها به‌غیراز Haberman، پس تغییرات ذکرشده، (بازنویسی)مقدار تابع هدف تقریباً روند تغییرات یکسان و جزئی داشته است. اگرچه این تغییرات برای مجموعه‌داده Wholesale Customers برای مقادیر پایین از کمی صعودی بوده است، بااین‌حال، مقدار تابع هدف به سمت یک مقدار ثابت همگرا شده است. برای مجموعه‌داده‌ی Haberman اگرچه پس از تغییرات زیاد ابتدایی، تابع هدف نسبتاً نوسان داشته است، بااین‌حال بازه‌ی نوسانات برای مقدار تابع هدف برای بیشتر مقادیر نسبتاً کم بوده ‌است.

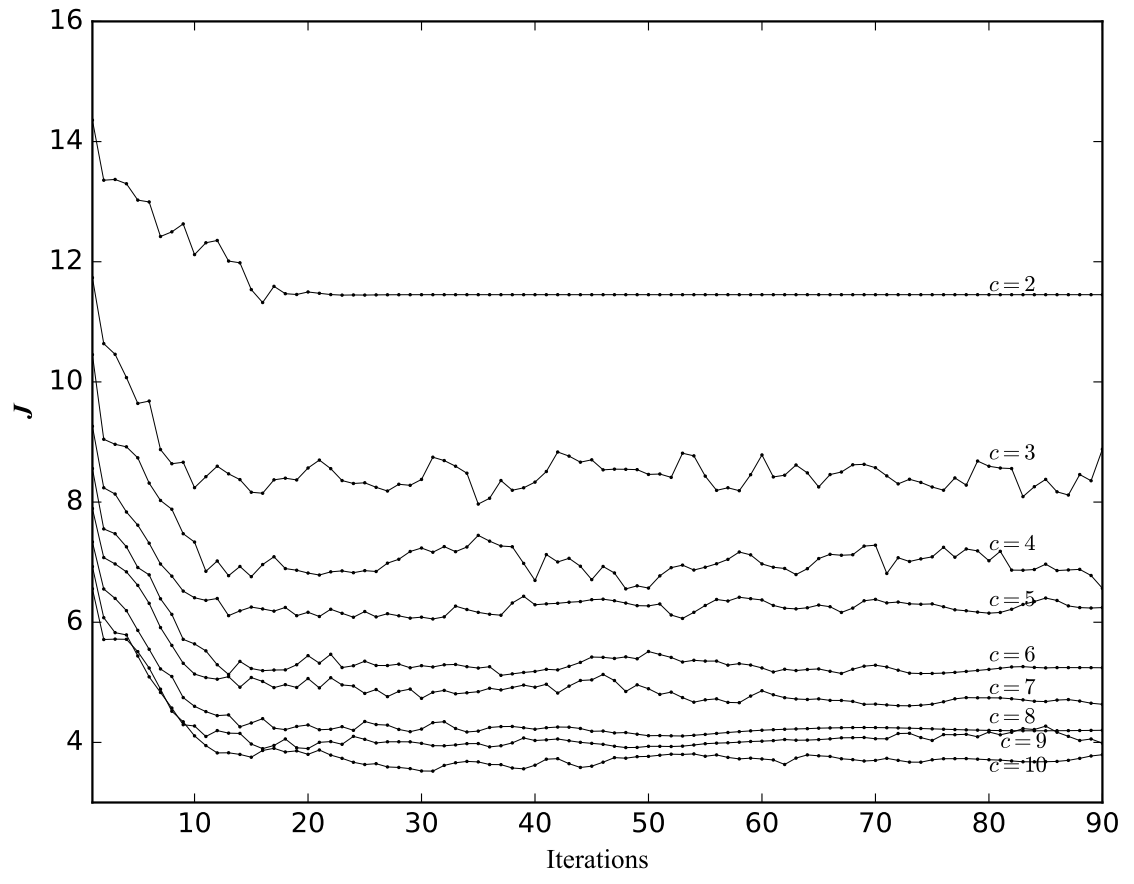
به‌طورکلی می‌توان گفت که الگوریتم استفاده‌شده به‌خوبی توانسته است که مقدار تابع هدف را حتی پس از تعداد کمی تکرار، نسبت به مقدار اولیه‌اش کم کند. این روند‌ تغییرات به‌طور تجربی، همگرایی الگوریتم ارائه‌شده و توانایی آن در کمینه کردن تابع هدف را نشان می‌دهد.



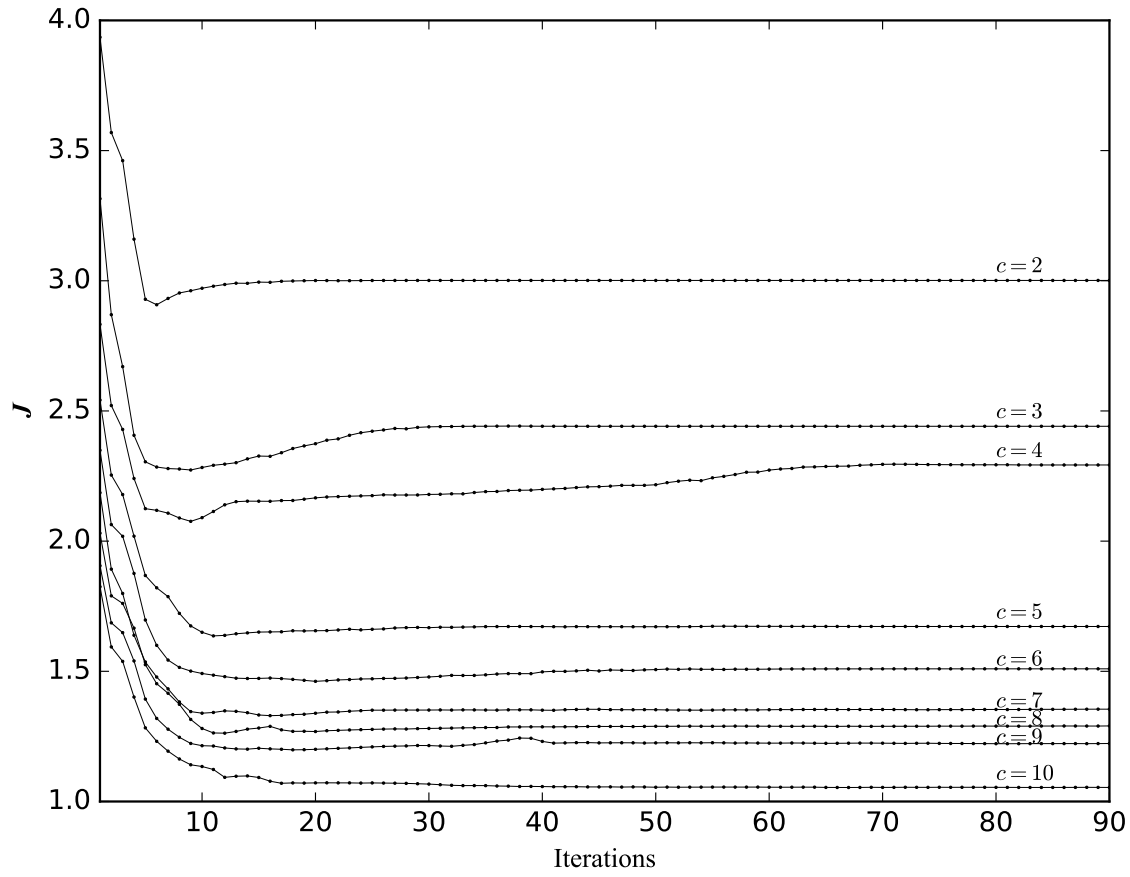
شکل ‏5‑3 مقادیر تابع هدف برحسب تعداد تکرار الگوریتم GFCM در مجموعه‌داده Iris



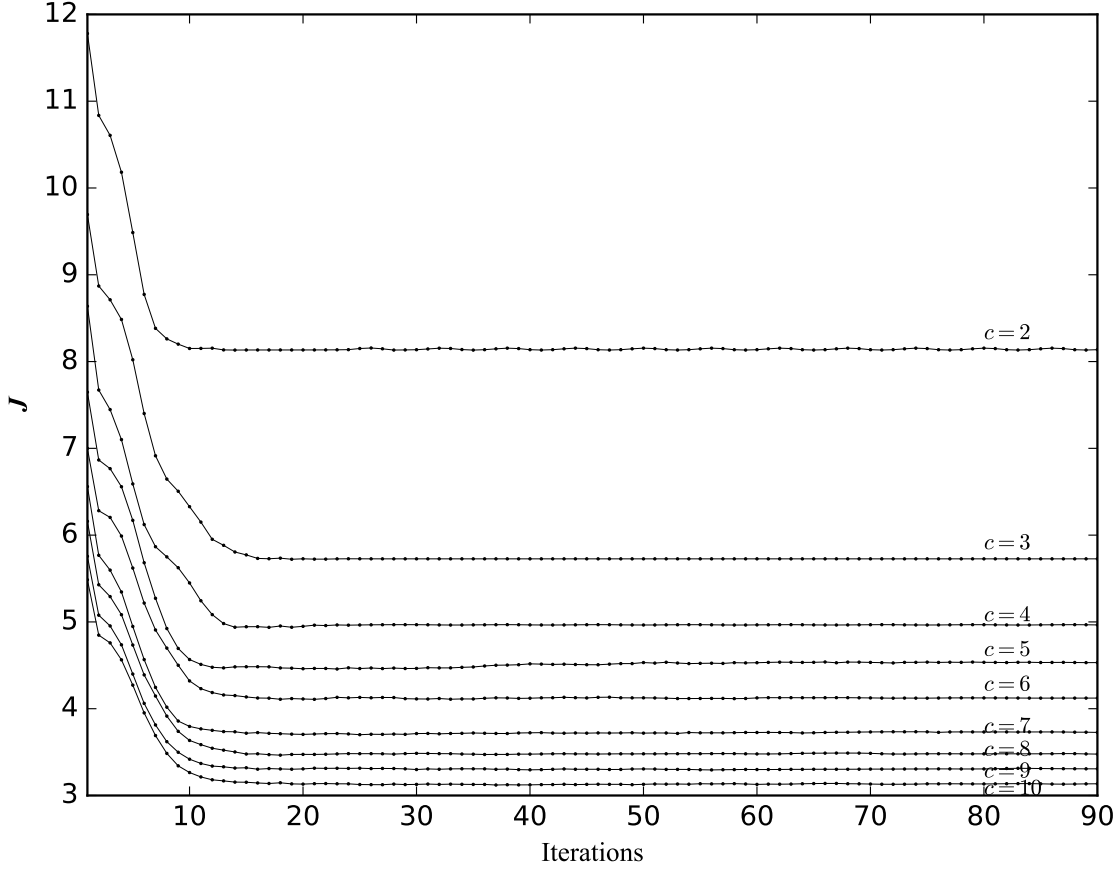
شکل ‏5‑4 مقادیر تابع هدف برحسب تعداد تکرار الگوریتم GFCM در مجموعه‌داده Glass



شکل ‏5‑5 مقادیر تابع هدف برحسب تعداد تکرار الگوریتم GFCM در مجموعه‌داده Haberman



شکل ‏5‑6 مقادیر تابع هدف برحسب تعداد تکرار الگوریتم GFCM در مجموعه‌داده Wholesale Customers



شکل ‏5‑7 مقادیر تابع هدف برحسب تعداد تکرار الگوریتم GFCM در مجموعه‌داده Wine

#### مقایسه عملکرد رویکردهای ارائه‌شده با دیگر روش‌ها توسط معیار میانگین مربعات خطا

در این بخش، عملکرد رویکرد پیشنهادی (جایگذاری با استفاده از خوشه‌بندی فازی c-means مبتنی بر گِرِی (GFCMI[[187]](#footnote-187))) با پنج روش دیگر که در بخش‌های پیشین ذکر شد، با استفاده از معیار میانگین مربعات خطا (RMSE) مقایسه می‌شود. همان‌طور که گفته شد، با توجه به اینکه برای گام‌ تخمین در هر خوشه از دو نوع رگرسیون چندگانه خطی (MLR) و رگرسیون بردار پشتیبان (SVR) استفاده‌شده است، بنابراین در مقایسات انجام‌شده، دو روش پیشنهادی ارائه می‌شود.

جدول ‏5‑1 تا

جدول ‏5‑5 (چرا این مدلی نوشته شده؟)مقدار RMSE را برای رویکردهای مختلف و درصدهای مختلف جاافتادگی در مجموعه‌داده‌های مختلف نمایش می‌دهد. مقادیر کمتر برای RMSE، بیانگر عملکرد بهتر است. بهترین عملکرد برای هر درصد جاافتادگی به‌صورت پررنگ و کج نوشته‌شده است. برای مجموعه‌داده Wine، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. فقط در حالت 40%، نوع MLR کمی بهتر از نوع SVR عمل کرده ‌است. برای مجموعه‌داده Wholesale Customers، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. همچنین نوع SVR در تمامی حالات از نوع MLR بهتر عمل کرده است. برای مجموعه‌داده Haberman، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. همچنین فقط در حالت 60%، نوع SVR کمی بهتر از نوع MLR عمل کرده است. برای مجموعه‌داده Iris، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. فقط در حالت 60%، نوع MLR کمی بهتر از نوع SVR عمل کرده است. برای مجموعه‌داده Glass، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. همچنین نوع MLR در تمامی حالات از نوع SVR بهتر عمل کرده است. به‌طورکلی می‌توان گفت که میانگین مربعات خطا برای رویکردهای پیشنهادی، کمتر از دیگر روش‌ها برای مجموعه‌داده‌های آزمایش‌شده است. اگرچه در برخی موارد عملکرد نوع SVR و MLR با یکدیگر متفاوت است، بااین‌حال هر دو نوع بهتر از دیگر روش‌ها عمل می‌کنند.

جدول ‏5‑1 مقایسه عملکرد رویکردهای جایگذاری توسط معیار RMSE در مجموعه‌داده Wine

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RMSE** | | | | | | | **درصد**  **جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/036360 | 0/023976 | 0/033730 | 0/030167 | 0/022538 | 0/021777 | ***0/021563*** | **30%** |
| 0/041661 | 0/028334 | 0/039631 | 0/034952 | 0/026348 | ***0/025403*** | 0/025503 | **40%** |
| 0/046409 | 0/031679 | 0/042908 | 0/039003 | 0/029118 | 0/028236 | ***0/028088*** | **50%** |
| 0/050661 | 0/035406 | 0/046779 | 0/042911 | 0/032450 | 0/031470 | ***0/031352*** | **60%** |

جدول ‏5‑2 مقایسه عملکرد رویکردهای جایگذاری توسط معیار RMSE در مجموعه‌داده Wholesale Customers

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RMSE** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/020691 | 0/015604 | 0/020098 | 0/018431 | 0/013842 | 0/013638 | ***0/012927*** | **30%** |
| 0/024958 | 0/017603 | 0/025692 | 0/022408 | 0/016467 | 0/016308 | ***0/015172*** | **40%** |
| 0/027446 | 0/020071 | 0/029130 | 0/024785 | 0/019145 | 0/018172 | ***0/017070*** | **50%** |
| 0/030793 | 0/022856 | 0/033716 | 0/028014 | 0/021546 | 0/021062 | ***0/020014*** | **60%** |

جدول ‏5‑3 مقایسه عملکرد رویکردهای جایگذاری توسط معیار RMSE در مجموعه‌داده Haberman

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RMSE** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/072809 | 0/090718 | 0/076416 | 0/072974 | 0/082576 | ***0/062842*** | 0/063411 | **30%** |
| 0/087862 | 0/110032 | 0/093408 | 0/088205 | 0/098874 | ***0/077887*** | 0/078400 | **40%** |
| 0/099338 | 0/123904 | 0/105968 | 0/099741 | 0/113329 | ***0/089366*** | 0/089492 | **50%** |
| 0/110476 | 0/135649 | 0/117827 | 0/111404 | 0/126812 | 0/100697 | ***0/100591*** | **60%** |

جدول ‏5‑4 مقایسه عملکرد رویکردهای جایگذاری توسط معیار RMSE در مجموعه‌داده Iris

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RMSE** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/091236 | 0/03086334 | 0/053599 | 0/064343 | 0/029823 | 0/026096 | ***0/025713*** | **30%** |
| 0/104716 | 0/03443357 | 0/059999 | 0/073202 | 0/033643 | 0/031683 | ***0/031501*** | **40%** |
| 0/117367 | 0/04063529 | 0/067326 | 0/083416 | 0/039206 | 0/038280 | ***0/038023*** | **50%** |
| 0/128874 | 0/04524099 | 0/073726 | 0/093255 | 0/042593 | ***0/041197*** | 0/041815 | **60%** |

جدول ‏5‑5 مقایسه عملکرد رویکردهای جایگذاری توسط معیار RMSE در مجموعه‌داده Glass

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RMSE** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/031926 | 0/028065 | 0/036946 | 0/030478 | 0/024559 | ***0/023559*** | 0/024403 | **30%** |
| 0/037667 | 0/033988 | 0/043886 | 0/035865 | 0/029146 | ***0/028146*** | 0/029077 | **40%** |
| 0/041271 | 0/038739 | 0/047221 | 0/039348 | 0/033161 | ***0/032161*** | 0/032211 | **50%** |
| 0/044692 | 0/043048 | 0/052514 | 0/042803 | 0/036080 | ***0/03508*** | 0/035226 | **60%** |

#### مقایسه عملکرد رویکردهای ارائه شده با دیگر روش‌ها توسط معیار ضریب تعیین

در این بخش، عملکرد رویکرد پیشنهادی (GFCMI) با پنج روش دیگر که در بخش‌های پیشین ذکر شد، با استفاده از معیار ضریب تعیین (CoD) مقایسه می‌شود.

جدول ‏5‑6 تا جدول ‏5‑10 مقدار CoD را برای رویکردهای مختلف و درصدهای مختلف جاافتادگی در مجموعه‌داده‌های مختلف نمایش می‌دهد. مقادیر بیش‌تر برای CoD، بیانگر عملکرد بهتر است. بهترین عملکرد برای هر درصد جاافتادگی به‌صورت پررنگ و کج نوشته‌شده است. برای مجموعه‌داده Wine، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. فقط در حالت 40%، نوع MLR کمی بهتر از نوع SVR عمل کرده است. برای مجموعه‌داده Wholesale Customers، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. همچنین نوع SVR در تمامی حالات از نوع MLR بهتر عمل کرده است. برای مجموعه‌داده Haberman، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. همچنین فقط در حالت 60%، نوع SVR کمی بهتر از نوع MLR عمل کرده است. متأسفانه برای مجموعه‌داده Iris، رویکردهای پیشنهادی کمی بدتر از روش تخمین با kNN عمل کرده‌اند. بااین‌حال، رویکردهای پیشنهادی در همه‌ی موارد، بهتر از باقی روش‌ها (به‌جز kNN) عمل کرده‌اند. برای مجموعه‌داده Glass، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. همچنین نوع SVR در تمامی حالات از نوع MLR بهتر عمل کرده است. به‌طورکلی می‌توان گفت که تقریباً مقدار ضریب تعیین برای رویکردهای پیشنهادی، بیشتر از دیگر روش‌ها برای مجموعه‌داده‌های آزمایش‌شده (به‌غیراز Iris) است. اگرچه در مجموعه‌داده Iris عملکرد رویکردهای پیشنهادی ضعیف‌تر است، بااین‌حال بسیار نزدیک به بهترین روش عمل کرده است.

جدول ‏5‑6 مقایسه عملکرد رویکردهای جایگذاری توسط معیار CoD در مجموعه‌داده Wine

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CoD** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/967889 | 0/985899 | 0/972275 | 0/977855 | 0/987501 | 0/988349 | ***0/988561*** | **30%** |
| 0/957826 | 0/980367 | 0/961761 | 0/970248 | 0/983043 | ***0/984233*** | 0/984032 | **40%** |
| 0/947717 | 0/975561 | 0/955174 | 0/963016 | 0/979283 | 0/980570 | ***0/980725*** | **50%** |
| 0/937771 | 0/969484 | 0/946659 | 0/955308 | 0/974294 | 0/975870 | ***0/975990*** | **60%** |

جدول ‏5‑7 مقایسه عملکرد رویکردهای جایگذاری توسط معیار CoD در مجموعه‌داده Wholesale Customers

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CoD** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/951196 | 0/972532 | 0/955519 | 0/960172 | 0/978164 | 0/978153 | ***0/980073*** | **30%** |
| 0/930522 | 0/966029 | 0/927108 | 0/943055 | 0/970223 | 0/970019 | ***0/973992*** | **40%** |
| 0/917500 | 0/955881 | 0/907306 | 0/931986 | 0/958613 | 0/963018 | ***0/967131*** | **50%** |
| 0/897446 | 0/943630 | 0/877256 | 0/914579 | 0/948900 | 0/951047 | ***0/955537*** | **60%** |

جدول ‏5‑8 مقایسه عملکرد رویکردهای جایگذاری توسط معیار CoD در مجموعه‌داده Haberman

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CoD** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/891742 | 0/831435 | 0/880756 | 0/891247 | 0/860367 | ***0/901628*** | 0/899919 | **30%** |
| 0/842639 | 0/752718 | 0/821772 | 0/841378 | 0/800664 | ***0/852525*** | 0/850664 | **40%** |
| 0/798956 | 0/686665 | 0/771137 | 0/797291 | 0/738527 | ***0/808816*** | 0/808269 | **50%** |
| 0/751297 | 0/624871 | 0/717118 | 0/747035 | 0/672536 | 0/760239 | ***0/760720*** | **60%** |

جدول ‏5‑9 مقایسه عملکرد رویکردهای جایگذاری توسط معیار CoD در مجموعه‌داده Iris

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CoD** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/878237 | ***0/990787*** | 0/957512 | 0/939097 | 0/986698 | 0/989888 | 0/990196 | **30%** |
| 0/839869 | ***0/988030*** | 0/947157 | 0/921499 | 0/983200 | 0/985101 | 0/985251 | **40%** |
| 0/798915 | ***0/982349*** | 0/933204 | 0/898267 | 0/977338 | 0/979304 | 0/979532 | **50%** |
| 0/757660 | ***0/977398*** | 0/920310 | 0/872956 | 0/973208 | 0/975822 | 0/976248 | **60%** |

جدول ‏5‑10 مقایسه عملکرد رویکردهای جایگذاری توسط معیار CoD در مجموعه‌داده Glass

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CoD** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/964872 | 0/972769 | 0/953004 | 0/967879 | 0/979049 | 0/977604 | ***0/980819*** | **30%** |
| 0/951631 | 0/962560 | 0/934284 | 0/956060 | 0/970817 | 0/970971 | ***0/973106*** | **40%** |
| 0/942019 | 0/953744 | 0/924029 | 0/947205 | 0/962392 | 0/962215 | ***0/965073*** | **50%** |
| 0/932190 | 0/944492 | 0/906177 | 0/937727 | 0/955709 | 0/955289 | ***0/957461*** | **60%** |

#### مقایسه عملکرد رویکردهای ارائه شده با دیگر روش‌ها توسط معیار میانگین مطلق خطا

در این بخش، عملکرد رویکرد پیشنهادی (GFCMI) با پنج روش دیگر که در بخش‌های پیشین ذکر شد، با استفاده از معیار میانگین مطلق خطا (MAE) مقایسه می‌شود.

جدول ‏5‑11 تا

جدول ‏5‑15 مقدار MAE را برای رویکردهای مختلف و درصدهای مختلف جاافتادگی در مجموعه‌داده‌های مختلف نمایش می‌دهد. مقادیر کمتر برای MAE، بیانگر عملکرد بهتر است. بهترین عملکرد برای هر درصد جاافتادگی به‌صورت پررنگ و کج نوشته‌شده است. برای مجموعه‌داده‌های Wholesale Customers، Haberman و Wine، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. همچنین در همه‌ موارد نوع SVR کمی بهتر از نوع MLR عمل کرده است. متأسفانه برای مجموعه‌داده Iris، رویکردهای پیشنهادی کمی بدتر از روش تخمین با kNN عمل کرده‌اند. بااین‌حال، رویکردهای پیشنهادی در همه‌ی موارد، بهتر از باقی روش‌ها (به‌جز kNN) عمل کرده‌اند. برای مجموعه‌داده Glass، در همه‌ی موارد، هر دو نوع از رویکرد پیشنهادی بهتر از باقی روش‌ها عمل کرده‌اند. همچنین نوع SVR در تمامی حالات غیر از 40%، از نوع MLR بهتر عمل کرده است. به‌طورکلی می‌توان گفت که تقریباً مقدار ضریب تعیین برای رویکردهای پیشنهادی، بیشتر از دیگر روش‌ها برای مجموعه‌داده‌های آزمایش‌شده (به‌غیراز Iris) است.

جدول ‏5‑11 مقایسه عملکرد رویکردهای جایگذاری توسط معیار MAE در مجموعه‌داده Wine

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MAE** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/004955 | 0/003013 | 0/004448 | 0/004028 | 0/002849 | 0/002760 | ***0/002721*** | **30%** |
| 0/006549 | 0/004120 | 0/006009 | 0/005356 | 0/003862 | 0/003719 | ***0/003695*** | **40%** |
| 0/008121 | 0/005093 | 0/007198 | 0/006666 | 0/004717 | 0/004622 | ***0/004542*** | **50%** |
| 0/009680 | 0/006194 | 0/008570 | 0/007991 | 0/005733 | 0/005598 | ***0/005512*** | **60%** |

جدول ‏5‑12 مقایسه عملکرد رویکردهای جایگذاری توسط معیار MAE در مجموعه‌داده Wholesale Customers

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MAE** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/002872 | 0/001965 | 0/003246 | 0/002400 | 0/001790 | 0/001777 | ***0/001566*** | **30%** |
| 0/004001 | 0/002639 | 0/004585 | 0/003364 | 0/002487 | 0/002503 | ***0/002169*** | **40%** |
| 0/004881 | 0/003294 | 0/005720 | 0/004130 | 0/003137 | 0/003094 | ***0/002733*** | **50%** |
| 0/005942 | 0/004028 | 0/007242 | 0/005070 | 0/003778 | 0/003800 | ***0/003411*** | **60%** |

جدول ‏5‑13 مقایسه عملکرد رویکردهای جایگذاری توسط معیار MAE در مجموعه‌داده Haberman

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MAE** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/018002 | 0/021412 | 0/019068 | 0/017740 | 0/019884 | 0/017030 | ***0/016193*** | **30%** |
| 0/025404 | 0/030644 | 0/027170 | 0/025140 | 0/027844 | 0/024455 | ***0/023712*** | **40%** |
| 0/032405 | 0/038785 | 0/034811 | 0/032105 | 0/035945 | 0/031497 | ***0/030888*** | **50%** |
| 0/039709 | 0/046719 | 0/042956 | 0/039483 | 0/044271 | 0/038865 | ***0/038396*** | **60%** |

جدول ‏5‑14 مقایسه عملکرد رویکردهای جایگذاری توسط معیار MAE در مجموعه‌داده Iris

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MAE** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/023348 | ***0/005184*** | 0/012332 | 0/015609 | 0/006201 | 0/005565 | 0/005499 | **30%** |
| 0/030674 | ***0/006778*** | 0/015974 | 0/020358 | 0/008134 | 0/007709 | 0/007640 | **40%** |
| 0/038663 | ***0/009194*** | 0/019834 | 0/026163 | 0/010600 | 0/009201 | 0/009820 | **50%** |
| 0/046295 | ***0/011513*** | 0/023732 | 0/032100 | 0/012683 | 0/011503 | 0/011878 | **60%** |

جدول ‏5‑15 مقایسه عملکرد رویکردهای جایگذاری توسط معیار MAE در مجموعه‌داده Glass

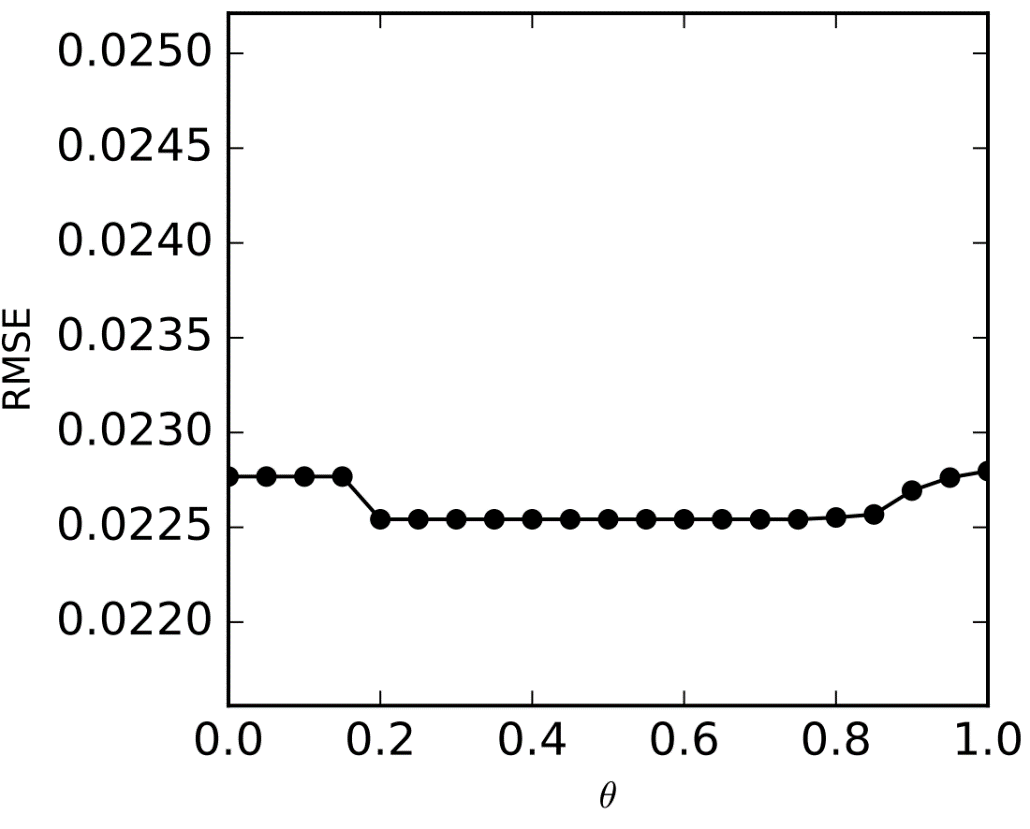
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MAE** | | | | | | | **درصد جاافتادگی** |
| **Mean** | **kNN** | **MLP** | **FCM** | **IARI** | **GFCMI (MLR)** | **GFCMI (SVR)** |
| 0/004039 | 0/003002 | 0/004790 | 0/003713 | 0/002660 | ***0/002647*** | 0/002682 | **30%** |
| 0/005585 | 0/004196 | 0/006621 | 0/005108 | 0/003728 | 0/003915 | ***0/003706*** | **40%** |
| 0/006855 | 0/052267 | 0/008014 | 0/006270 | 0/004578 | 0/004502 | ***0/004222*** | **50%** |
| 0/008282 | 0/006525 | 0/009776 | 0/007636 | 0/005591 | 0/005626 | ***0/005250*** | **60%** |

#### تأثیر پارامترهای c و

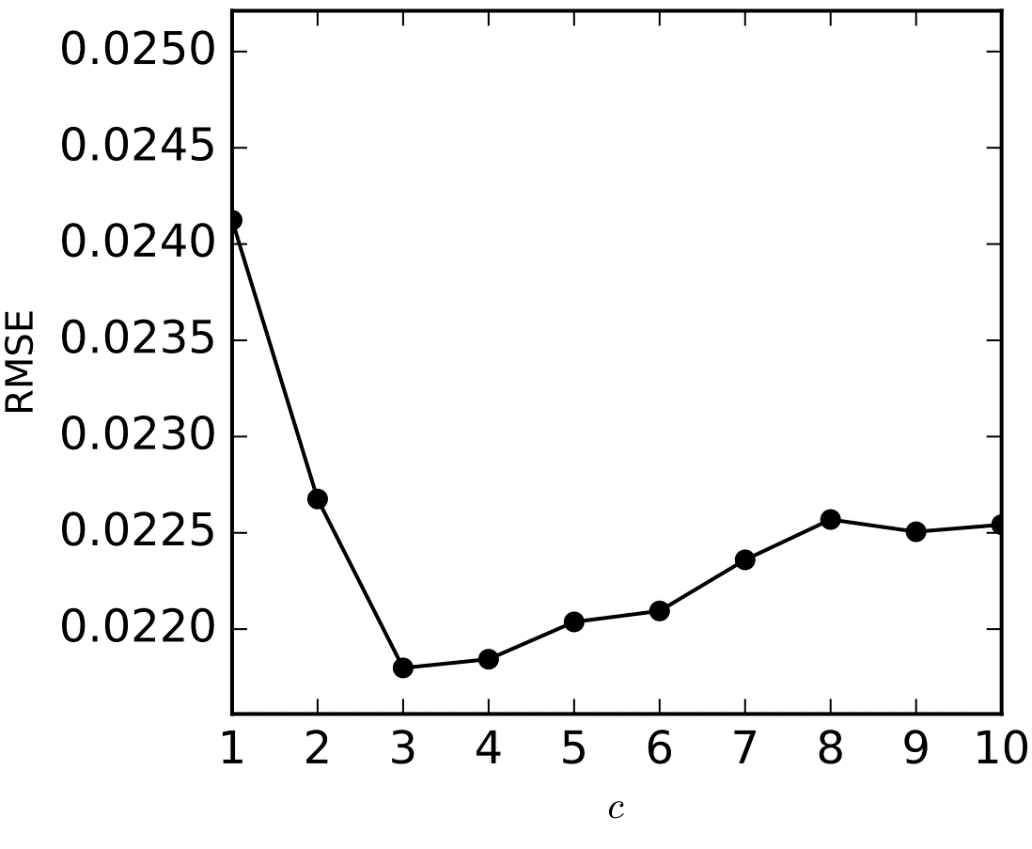
در این بخش تأثیر پارامترهای ورودی بر روی عملکرد رویکردهای ارائه‌شده بررسی می­شود. همان­طور که در بخش­های پیش توضیح داده ­شد، در روش­های ارائه‌شده دو پارامتر اصلی c و θ نیاز به تنظیم دارند. پارامتر θ حداقل میزان همبستگی برای انتخاب در عمل انتخاب ویژگی را تعیین می‌کند. پارامتر c، تعداد خوشه‌ها را برای مرحله خوشه‌بندی تعیین می‌‌‌کند.

شکل ‏5‑8 تأثیر پارامتر θ بر مقدار RMSE را در مجموعه‌داده Wine برای روش پیشنهادی از نوع MLR نشان می‌دهد. همان‌طور که پیداست، پس انجام عمل انتخاب ویژگی مقدار خطای RMSE کاهش داشته است. به‌طورکلی هنگامی‌که تعداد ویژگی‌ها در یک مجموعه‌داده بالا باشد، عمل انتخاب ویژگی تأثیر مثبتی بر عملکرد رویکرد پیشنهادی خواهد داشت. اگر تعداد ویژگی‌ها بسیار کم باشد، مثل مجموعه‌داده Haberman، عمل انتخاب ویژگی در مقدار RMSE کم تأثیر و یا بی‌تأثیر خواهد بود. با توجه به این‌ نکته که اگر مقدار θ بسیار بالا انتخاب شود و بنابراین صفتی شرایط لازم را ارضا نکند، تمام صفات استفاده خواهند شد، مشاهده می‌شود که در مقادیر بالا برای θ، میزان خطا مانند حالتی که انتخاب ویژگی انجام نمی‌شود، خواهد بود

شکل ‏5‑9 تأثیر پارامتر c بر مقدار RMSE را در مجموعه‌داده Wine برای روش پیشنهادی از نوع MLR نشان می‌دهد. همان‌طور که پیداست، خوشه‌بندی داده‌ها مقدار خطای تخمین را کاهش داده است. درواقع مقدار خطا در یک مقدار بهینه‌شده است. مقدار c بهینه برای مجموعه‌داده‌های مختلف، متفاوت است. به‌طورکلی، آزمایش‌ها نشان ‌دادند، افزایش مقدار c از یک حد بهینه به بعد، به‌تدریج باعث افزایش مقدار خطای تخمین می‌شود. به دلیل محدودیت فضا، نتایج برای باقی مجموعه‌داده‌ها و دیگر معیارها نشان داده نشده­اند.



شکل ‏5‑8 تأثیر پارامتر θ بر مقدار RMSE در مجموعه‌داده Wine



شکل ‏5‑9 تأثیر پارامتر c بر مقدار RMSE در مجموعه‌داده Wine

## خلاصه و نتیجه‌گیری

در این فصل رویکردی جدید با استفاده از خوشه‌بندی c-means فازی مبتنی بر گِرِی و انتخاب ویژگی مبتنی بر معیار اطلاعات متقابل برای تخمین مقادیر جاافتاده عددی ارائه شد. در این روش، از یک نسخه‌ی جدید برای الگوریتم خوشه‌بندی فازی c-means برای هدف افراز بندی فازی داده‌ها استفاده شد.

روش پیشنهادی با پنج روش مختلف جایگذاری مقادیر جاافتاده مقایسه شد. در آزمایش­های انجام‌شده از پنج مجموعه‌داده‌ی مختلف و سه معیار ارزیابی مختلف استفاده شد. نتایج نشان داد که در اکثر موارد، عملکرد رویکرد پیشنهادی بهتر از روش­های مقایسه شده بود­ه است. درنهایت، تأثیر پارامترهای ورودی بر عملکردهای پیشنهادی بررسی شدند. مطالعات آینده می‌تواند شامل ارزیابی عملکرد رویکردهای پیشنهادی بر دیگر مجموعه‌داده‌ها و دیگر مکانیزم­های تولید مقادیر جاافتاده یعنی MCAR و MNAR خواهد بود.

# نتیجه‌گیری، بحث و پیشنهاد‌ها

## مقدمه

این فصل به بررسی نتایج تحقیق و همچنین کارهای آتی برای این پژوهش، می‌پردازد. ابتدا روش‌های پیشنهادی به‌طور خلاصه مرور می‌شوند، سپس برای هر یک از روش‌های پیشنهادی، چالش‌های اصلی بیان می‌شوند. درنهایت، کارهای پیش‌ رو که می‌تواند در راستای این پژوهش باشد، مشخص می‌شوند.

## نتایج تحقیق

در این پژوهش به بررسی مشکل مقادیر جاافتاده در مجموعه‌داده‌ها پرداخته‌ شد. برای عملیات جایگذاری مقادیر جاافتاده سه دسته رویکرد متفاوت ارائه شد. تأکید اصلی روش‌های پیشنهادی، توجه به افراز‌هایی از مجموعه‌داده‌ها به‌جای کل مجموعه‌داده است. این سه دسته عبارتنداز:

1. روشی دومرحله‌ای مبتنی بر خوشه‌بندی داده‌ها و استفاده از ‌ترکیب روش k نزدیک‌ترین همسایگان وزن‌دار و رگرسیون خطی برای پرکردن مقادیر جاافتاده عددی
2. ده روش برای جایگذاری مقادیر جاافتاده با استفاده از روش­های مبتنی بر همبستگی
3. جایگذاری مقادیر جاافتاده با استفاده از خوشه‌بندی فازی c-means مبتنی بر گِرِی و انتخاب ویژگی مبتنی بر معیار اطلاعات متقابل

در دسته‌ی اول سعی شد تا با سنجشِ همبستگی در افرازهای مختلف داده‌ها، در صورت وجود روابط قوی از مدل رگرسیون و در غیر این صورت از روش نزدیک‌ترین همسایگان برای تخمین مقادیر جاافتاده درون هر خوشه استفاده کرد. در دسته‌ی دوم، روش­هایی ارائه شد تا با اعمال مدل‌های رگرسیون خطی بر زیرمجموعه‌هایی از داده‌ها با همبستگی بالا نسبت به صفت جاافتاده، دقت جایگذاری افزایش یابد. انتخاب زیرمجموعه­های مطلوب با استفاده رویکردهایی مبتنی بر انتخاب روبه‌جلو صورت گرفت. در دسته‌‌ی سوم، پس از خوشه‌بندی فازی c-means مبتنی بر گِرِی، با انتخاب صفات مرتبط به صفت جاافتاده توسط معیار اطلاعات متقابل درون هر خوشه، مقادیر جاافتاده جایگذاری شدند.

به‌طورکلی، نتایج نشان می‌دهند که افرازبندی مناسب داده‌ها می‌تواند منجر به افزایش دقت در تخمین مقادیر جاافتاده شود. نکته‌ی مهم این است که نحوه‌ی افرازبندیِ مناسب به‌درستی انتخاب شود. این افرازبندی‌ها می‌توانند مبتنی بر معیارهای متفاوتی باشند. ازجمله این معیارها می‌توان به فاصله و همبستگی بین نمونه‌ها اشاره کرد. افرازبندی می‌تواند فازی و غیرفازی (سخت) باشد. به نظر می‌رسد که افرازبندی فازی به‌طور بهتری می‌تواند رفتار مجموعه‌داده‌های ناکامل را مدل‌سازی کند. نکته‌ی مهم دیگر این است که برای پر کردن هر ویژگی جاافتاده از یک نمونه‌ی جاافتاده، لازم است که از اطلاعات صفات مناسب برای پرکردن صفت جاافتاده استفاده کرد. این صفات مناسب می‌تواند با استفاده از بررسی میزان همبستگی صفات، انتخاب شوند.

چالش اصلی در روش‌های پیشنهادی انتخاب بهینه‌ی پارامترهای ورودی برای هریک از روش‌ها است. اگرچه این چالش نیز با عملیاتِ زمان­برِ جستجوی کامل برای انتخاب بهترین ترکیب پارامترها قابل برطرف شدن است. راه‌حل دیگر برای این مسئله استفاده از روش‌ها بهینه‌سازی مثل الگوریتم ژنتیک است.

## پیشنهادهایی برای پژوهش‌های آینده

رویکردهای پیشنهادشده شامل مراحل مختلفی هستند. برای هریک از مراحل می‌توان تغییراتی را برای بررسی و حصول نتایج بهتر احتمالی اعمال کرد. برای دسته‌ی اول از رویکردهای پیشنهادی، می‌توان تأثیر دیگر روش‌های خوشه‌بندی مثل خوشه‌بندی سلسله مراتبی، خوشه‌بندی مبتنی بر چگالی، خوشه‌بندی با شبکه عصبی SOM و دیگر روش‌های افرازی بندی را در مرحله‌ اول الگوریتم بررسی کرد. به‌علاوه، می‌توان مدل‌های رگرسیون دیگری را به‌جای رگرسیون خطی جایگزین کرد. برای دسته‌ی دوم از رویکردهای پیشنهادی، می‌توان از معیارهای دیگری برای سنجش میزان همبستگی استفاده نمود؛ معیارهایی که قابلیت سنجش همبستگی غیرخطی را نیز دارند (مثل معیار اطلاعات متقابل). به‌علاوه، می‌توان مدل‌های رگرسیون دیگری را به‌جای رگرسیون خطی در مرحله تخمین نهایی جایگزین کرد. برای دسته‌ی سوم از رویکردهای پیشنهادی، می‌توان تأثیر مدل‌های رگرسیون دیگر را به‌جای رگرسیون خطی و رگرسیون بردار پشتیبان در مرحله تخمین، بررسی کرد.

برای بهبود پیچیدگی زمانی تمام رویکردهای پیشنهادی، می‌توان پس از فرآیند افرازبندی، پردازش درون هریک از افرازها برای تخمین مقادیر جاافتاده را به‌صورت موازی انجام داد و درنهایت نتایج را ترکیب کرد.

علاوه بر موارد ذکرشده، به‌طورکلی مطالعات آینده می‌تواند شامل ارزیابی عملکرد رویکردهای پیشنهادی بر دیگر مجموعه‌داده‌ها (مجموعه‌داده‌های حجیم‌تر) و دیگر مکانیزم­های تولید مقادیر جاافتاده غیر از MAR یعنی MCAR و MNAR باشد.

# منابع و مراجع

**مراجع:**

[1] Han, J., Kamber, M., & Pei, J. (2011a). *Data Mining: Concepts and Techniques* (3rd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[2] Hosseininezhad, F., & Salajegheh, A. (2013b). Study and Comparison of Partitioning Clustering Algorithms. *Iranian Journal of Medical Informatics*, *2*(1).

[3] Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002c). Models and issues in data stream systems. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 1–16).

[4] Bai, L., Liang, J., & Dang, C. (2011d). An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. *Knowledge-Based Systems*, *24*(6), 785–795. https://doi.org/10.1016/j.knosys.2011.02.015

[5] Berry, M., & Linoff, G. (1999e). *Mastering data mining: The art and science of customer relationship management*. John Wiley & Sons, Inc.

[6] Pendharkar, P. C., & Subramanian, G. (2003f). *Managing Data Mining Technologies in Organizations*. IGI Global. Retrieved from http://dl.acm.org/citation.cfm?id=778595.778626

[7] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996g). Advances in Knowledge Discovery and Data Mining. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 1–34). Menlo Park, CA, USA: American Association for Artificial Intelligence. Retrieved from http://dl.acm.org/citation.cfm?id=257938.257942

[8] Tsai, C. F., & Chang, F. Y. (2016h). Combining instance selection for better missing value imputation. *Journal of Systems and Software*, *122*, 63–71. https://doi.org/10.1016/j.jss.2016.08.093

[9] Aydilek, I. B., & Arslan, A. (2013i). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, *233*, 25–35. https://doi.org/10.1016/j.ins.2013.01.021

[10] Ishay, R. Ben, & Herman, M. (2015j). A novel algorithm for the integration of the imputation of missing values and clustering. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9166, pp. 115–129). https://doi.org/10.1007/978-3-319-21024-7\_8

[11] Tang, N. (2014k). Big data cleaning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8709 LNCS, pp. 13–24). https://doi.org/10.1007/978-3-319-11116-2\_2

[12] Salim, N., Ibrahim, R., & others. (2011l). Towards data quality into the data warehouse development. In *Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on* (pp. 1199–1206).

[13] Vagin, V., & Fomina, M. (2011m). Problem of knowledge discovery in noisy databases. *International Journal of Machine Learning and Cybernetics*, *2*(3), 135–145. https://doi.org/10.1007/s13042-011-0028-x

[14] Qin, Y., Zhang, S., Zhu, X., Zhang, J., & Zhang, C. (2009n). POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. *Expert Systems with Applications*, *36*(2 PART 2), 2794–2804. https://doi.org/10.1016/j.eswa.2008.01.059

[15] Rahm, E., & Do, H. H. (2000o). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3–13.

[16] Minton, S. N., Nanjo, C., Knoblock, C. A., Michalowski, M., & Michelson, M. (2005p). A heterogeneous field matching method for record linkage. In *Proceedings - IEEE International Conference on Data Mining, ICDM* (pp. 314–321). https://doi.org/10.1109/ICDM.2005.7

[17] Eckerson, W. (2011q). Data Quality and the BottomLine: Achieving Business Success through a Commitment to High Quality Data. *Tech Target*, (June), 1–39. Retrieved from www.tdwi.org/research

[18] Rahman, M. G., & Islam, M. Z. (2010r). A decision tree-based missing value imputation technique for data pre-processing. In *Conferences in Research and Practice in Information Technology Series* (Vol. 121, pp. 41–50). Darlinghurst, Australia, Australia: Australian Computer Society, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=2483628.2483635

[19] Garc’\ia, S., Luengo, J., & Herrera, F. (2015s). *Data preprocessing in data mining*. Springer.

[20] Rahman, M. G., & Islam, M. Z. (2016t). Missing value imputation using a fuzzy clustering-based EM approach. *Knowledge and Information Systems*, *46*(2), 389–422. https://doi.org/10.1007/s10115-015-0822-y

[21] Sun, G., Shao, J., Han, H., & Ding, X. (2016u). Missing value imputation for wireless sensory soil data: A comparative study. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9784, pp. 172–184). https://doi.org/10.1007/978-3-319-42553-5\_15

[22] Bache, K., & Lichman, M. (2013v). UCI Machine Learning Repository. *University of California Irvine School of Information*. https://doi.org/University of California, Irvine, School of Information and Computer Sciences

[23] García-Laencina, P. J., Sancho-Gómez, J.-L., & Figueiras-Vidal, A. R. (2010w). Pattern classification with missing data: a review. *Neural Computing and Applications*, *19*(2), 263–282. https://doi.org/10.1007/s00521-009-0295-6

[24] Ye, C., & Wang, H. (2014x). Capture Missing Values Based on Crowdsourcing. In *Wireless Algorithms, Systems, and Applications* (pp. 783–792). New York, NY, USA: Springer-Verlag New York, Inc. https://doi.org/10.1007/978-3-319-07782-6\_70

[25] Silva-Ram??rez, E. L., Pino-Mej??as, R., & L??pez-Coello, M. (2015y). Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing Journal*, *29*, 65–74. https://doi.org/10.1016/j.asoc.2014.09.052

[26] Amiri, M., & Jensen, R. (2016z). Missing data imputation using fuzzy-rough methods. *Neurocomputing*, *205*, 152–164. https://doi.org/10.1016/j.neucom.2016.04.015

[27] Farhangfar, A., Kurgan, L., & Dy, J. (2008aa). Impact of imputation of missing values on classification error for discrete data. *International Journal of Communication Networks and Distributed Systems*, *41*(12), 3692–3705.

[28] Maletic, J., & Marcus, A. (2000ab). Data Cleansing: Beyond Integrity Analysis. In *Iq* (pp. 1–10). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.5212&rep=rep1&type=pdf

[29] Wang, H., & Wang, S. (2010ac). Mining incomplete survey data through classification. *Knowledge and Information Systems*, *24*(2), 221–233. https://doi.org/10.1007/s10115-009-0245-8

[30] Ye, C., Wang, H., Li, J., Gao, H., & Cheng, S. (2016ad). Crowdsourcing-enhanced missing values imputation based on Bayesian network. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9642, pp. 67–81). https://doi.org/10.1007/978-3-319-32025-0\_5

[31] Tran, C. T., Zhang, M., Andreae, P., & Xue, B. (2016ae). Improving performance for classification with incomplete data using wrapper-based feature selection. *Evolutionary Intelligence*, *9*(3), 81–94. https://doi.org/10.1007/s12065-016-0141-6

[32] Rahman, M. G., & Islam, M. Z. (2013af). Data quality improvement by imputation of missing values. In *International Conference on Computer Science and Information Technology* (pp. 82–88).

[33] Barnard, J., & Meng, X.-L. (1999ag). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, *8*(1), 17–36. https://doi.org/10.1177/096228029900800103

[34] Tran, C. T., Zhang, M., & Andreae, P. (2016ah). A genetic programming-based imputation method for classification with missing data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9594, pp. 149–163). https://doi.org/10.1007/978-3-319-30668-1\_10

[35] Little, R. J., & Rubin, D. B. (2002ai). *Statistical analysis with missing data.* (Vol. Second Edi). John Wiley & Sons.

[36] Tian, J., Yu, B., Yu, D., & Ma, S. (2014aj). Missing data analyses: A hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering. *Applied Intelligence*, *40*(2), 376–388. https://doi.org/10.1007/s10489-013-0469-x

[37] van Stein, B., & Kowalczyk, W. (2016ak). An incremental algorithm for repairing training sets with missing values. In J. P. Carvalho, M.-J. Lesot, U. Kaymak, S. Vieira, B. Bouchon-Meunier, & R. R. Yager (Eds.), *Communications in Computer and Information Science* (Vol. 611, pp. 175–186). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40581-0\_15

[38] Rubin, D. B. (1987al). *Multiple Imputation for Nonresponse in Surveys*. *John Wiley & Sons* (Vol. 81). John Wiley & Sons. https://doi.org/10.1002/9780470316696

[39] Little, R. J. A., & Rubin, D. B. (2002am). Statistical Analysis with Missing Data, Second Edition. In *Statistical Analysis with Missing Data* (pp. 1–23). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119013563.ch1

[40] Samad, T., & Harp, S. (1992an). Self–organization with partial data. *Network: Computation in Neural Systems*, *3*(2), 205–212. https://doi.org/10.1088/0954-898X/3/2/008

[41] Singh, N., Javeed, A., Chhabra, S., & Kumar, P. (2015ao). Missing Value Imputation with Unsupervised Kohonen Self Organizing Map. In N. R. Shetty, N. H. Prasad, & N. Nalini (Eds.), *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2015, Volume 1* (pp. 61–76). New Delhi: Springer India. https://doi.org/10.1007/978-81-322-2550-8\_7

[42] García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2013ap). Classifying patterns with missing values using Multi-Task Learning perceptrons. *Expert Systems with Applications*, *40*(4), 1333–1341. https://doi.org/10.1016/j.eswa.2012.08.057

[43] Gupta, A., & Lam, M. S. (1996aq). Estimating Missing Values Using Neural Networks. *The Journal of the Operational Research Society*, *47*(2), 229. https://doi.org/10.2307/2584344

[44] Nkuna, T. R., & Odiyo, J. O. (2011ar). Filling of missing rainfall data in Luvuvhu River Catchment using artificial neural networks. *Physics and Chemistry of the Earth, Parts A/B/C*, *36*(14–15), 830–835. https://doi.org/10.1016/j.pce.2011.07.041

[45] Nordbotten, S. (1996as). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data. *Journal of Official Statistics*, *12*(4), 385–401.

[46] Sharpe, P., & Solly, R. (1995at). Dealing with missing values in neural network based diagnostic systems. *Neural Computing and Applications*, *3*(xx), 73–77.

[47] Silva-Ramírez, E.-L., Pino-Mejías, R., López-Coello, M., & Cubiles-de-la-Vega, M.-D. (2011au). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, *24*(1), 121–129. https://doi.org/10.1016/j.neunet.2010.09.008

[48] Yoon, S.-Y., & Lee, S.-Y. (1999av). Training Algorithm with Incomplete Data for Feed-Forward Neural Networks. *Neural Processing Letters*, *10*(3), 171–179. https://doi.org/10.1023/A:1018772122605

[49] Marseguerra, M., & Zoia, A. (2005aw). The AutoAssociative Neural Network in signal analysis: II. Application to on-line monitoring of a simulated BWR component. *Annals of Nuclear Energy*, *32*(11), 1207–1223.

[50] Marwala, T., & Chakraverty, S. (2006ax). Fault classification in structures with incomplete measured data using autoassociative neural networks and genetic algorithm. *Current Science-Bangalore-*, *90*(4), 542. Retrieved from http://www.iisc.ernet.in/currsci/feb252006/542.pdf

[51] Batista, G. E. A. P. A., & Monard, M. C. (2002ay). A study of k-nearest neighbour as an imputation method. *Frontiers in Artificial Intelligence and Applications*, *87*(251–260), 251–260.

[52] Batista, G., & Monard, M. C. (2003az). Experimental comparison of K-nearest neighbor and mean or mode imputation methods with the internal strategies used by C4. 5 and CN2 to treat missing data. *University of Sao Paulo*, *34*.

[53] Bose, S., Das, C., Dutta, S., & Chattopadhyay, S. (2012ba). A novel interpolation based missing value estimation method to predict missing values in microarray gene expression data. In *2012 International Conference on Communications, Devices and Intelligent Systems (CODIS)* (pp. 318–321). https://doi.org/10.1109/CODIS.2012.6422202

[54] Dixon, J. K. (1979bb). Pattern recognition with partly missing data. *IEEE Transactions on Systems, Man and Cybernetics*, *9*(10), 617–621. https://doi.org/10.1109/TSMC.1979.4310090

[55] García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R., & Verleysen, M. (2009bc). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, *72*(7–9), 1483–1493. https://doi.org/10.1016/j.neucom.2008.11.026

[56] Jiang, C., & Yang, Z. (2015bd). CKNNI: An improved KNN-based missing value handling technique. In D.-S. Huang & K. Han (Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9227, pp. 441–452). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-22053-6\_47

[57] Krishnamoorthy, R., Kumar, S. S., & Neelagund, B. (2014be, May). A new approach for data cleaning process. *Recent Advances and Innovations in Engineering (ICRAIE), 2014*. https://doi.org/10.1109/ICRAIE.2014.6909249

[58] Zhang, C., Zhu, X., Zhang, J., Qin, Y., & Zhang, S. (2007bf). GBKII: an imputation method for missing values. In Z.-H. Zhou, H. Li, & Q. Yang (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 1080–1087). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71701-0\_122

[59] Huang, C. C., & Lee, H. M. (2004bg). A grey-based nearest neighbor approach for missing attribute value prediction. *Applied Intelligence*, *20*(3), 239–252. https://doi.org/10.1023/B:APIN.0000021416.41043.0f

[60] Hruschka, E. R., Hruschka, E. R., & Ebecken, N. F. F. (2003bh). Evaluating a Nearest-Neighbor Method to Substitute Continuous Missing Values. In T. (Tom) D. Gedeon & L. C. C. Fung (Eds.), *AI 2003: Advances in Artificial Intelligence: 16th Australian Conference on AI, Perth, Australia, December 3-5, 2003. Proceedings* (pp. 723–734). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24581-0\_62

[61] Franco, L., & Escuela, T. (2006bi). Missing data imputation in breast cancer prognosis. *BioMed*, *6*, 323–328.

[62] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., … Altman, R. B. (2001bj). Missing value estimation methods for DNA microarrays. *Bioinformatics. 2001 Jun;*, *17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

[63] Batista, G. E. A. P. A., & Monard, M. C. (2003bk). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, *17*(5–6), 519–533. https://doi.org/10.1080/713827181

[64] Schneider, T. (2001bl). Analysis of incomplete climate data: Estimation of Mean Values and covariance matrices and imputation of Missing values. *Journal of Climate*, *14*(5), 853–871. https://doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2

[65] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977bm). Maximum likelihood from incomplete data via the {EM} algorithm. *Journal Royal Statistical Society*, *39*(1), 1–38.

[66] Deb, R., & Liew, A. W.-C. (2016bn). Missing value imputation for the analysis of incomplete traffic accident data. *Information Sciences*, *339*, 274–289. https://doi.org/10.1016/j.ins.2016.01.018

[67] Gelman, A., & Hill, J. (2007bo). *Data Analysis Using Regression and Multilevel/Hierarchical\nModels*. Cambridge university press.

[68] Zhang, S. (2011bp). Shell-neighbor method and its application in missing data imputation. *Applied Intelligence*, *35*(1), 123–133. https://doi.org/10.1007/s10489-009-0207-6

[69] Wu, C.-H. W. C.-H., Wun, C.-H. W. C.-H., & Chou, H.-J. C. H.-J. (2004bq). Using association rules for completing missing data. In *Fourth International Conference on Hybrid Intelligent Systems (HIS’04)* (pp. 236–241). https://doi.org/10.1109/ICHIS.2004.91

[70] Wu, J., Song, Q., & Shen, J. (2007br). An novel association rule mining based missing nominal data imputation method. In *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference on* (Vol. 3, pp. 244–249).

[71] Ragel, A., & Cr??milleux, B. (1999bs). MVC - a preprocessing method to deal with missing values. *Knowledge-Based Systems*, *12*(5–6), 285–291. https://doi.org/10.1016/S0950-7051(99)00022-2

[72] Li, D., Deogun, J., Spaulding, W., & Shuart, B. (2004bt). Towards Missing Data Imputation : A Study of Fuzzy K-means Clustering Method. *Springer-Verlag Berlin Heidelberg*, *40*(c), 573–579. https://doi.org/10.1007/978-3-540-25929-9\_70

[73] Ayuyev, V. V., Jupin, J., Harris, P. W., & Obradovic, Z. (2009bu). Dynamic clustering-based estimation of missing values in mixed type data. In T. B. Pedersen, M. K. Mohania, & A. M. Tjoa (Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5691 LNCS, pp. 366–377). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03730-6\_29

[74] Ankaiah, N., & Ravi, V. (2011bv). A Novel Soft Computing Hybrid for Data Imputation. In *7th international conference on data mining (DMIN)* (Vol. 57).

[75] Azim, S., & Aggarwal, S. (2014bw). Hybrid Model for Data Imputation: Using Fuzzy c-means and Multi Layer Perceptron. In *I2014 IEEE nternational Advance Computing Conference (IACC)* (pp. 1281–1285). https://doi.org/10.1109/IAdCC.2014.6779512

[76] Bashir, S., Razzaq, S., Maqbool, U., Tahir, S., & Baig, A. R. (2009bx). Using Association Rules for Better Treatment of Missing Values. *arXiv Preprint arXiv:0904.3320*. Retrieved from http://arxiv.org/abs/0904.3320

[77] Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010by). Missing value imputation based on k-mean clustering with weighted distance. In S. Ranka, A. Banerjee, K. K. Biswas, S. Dua, P. Mishra, R. Moona, … C.-L. Wang (Eds.), *Communications in Computer and Information Science* (Vol. 94 CCIS, pp. 600–609). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-14834-7\_56

[78] Nelwamondo, F. V., Golding, D., & Marwala, T. (2013bz). A dynamic programming approach to missing data estimation using neural networks. *Information Sciences*, *237*, 49–58. https://doi.org/10.1016/j.ins.2009.10.008

[79] Lv, Z., Zhao, J., Liu, Y., Wang, W., Tran, C. T., Zhang, M., … Lichman, M. (2016ca). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, *40*(2), 257–286. https://doi.org/10.1002/9781119013563.ch1

[80] Di Zio, M., Scanu, M., Coppola, L., Luzi, O., & Ponti, A. (2004cb). Bayesian networks for imputation. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *167*(2), 309–322. https://doi.org/10.1046/j.1467-985X.2003.00736.x

[81] Duan, L., Yue, K., Qian, W., & Liu, W. (2013cc). Cleaning Missing Data Based on the Bayesian Network. In *International Conference on Web-Age Information Management* (pp. 348–359).

[82] Hruschka, E. R., Hruschka, E. R., & Ebecken, N. F. F. (2007cd). Bayesian networks for imputation in classification problems. *Journal of Intelligent Information Systems*, *29*(3), 231–252. https://doi.org/10.1007/s10844-006-0016-x

[83] Rahman, M. G., & Islam, M. Z. (2013ce). kDMI: A novel method for missing values imputation using two levels of horizontal partitioning in a data set. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8347 LNAI, pp. 250–263). https://doi.org/10.1007/978-3-642-53917-6\_23

[84] Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014cf). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis*, *72*, 92–104. https://doi.org/10.1016/j.csda.2013.10.025

[85] Carpenter, J., & Kenward, M. (2013cg). *Multiple imputation and its application - Errata May 2013* (Vol. 1). John Wiley & Sons. Retrieved from http://books.google.com/books?hl=en&lr=&id=rwR548UmRaIC&oi=fnd&pg=PT6&dq=Multiple+Imputation+and+its+application&ots=lWyA-vbS\_U&sig=X65aGeSS9uDm288vZB0\_KkL-duU

[86] Rubin, D. B., & Schenker, N. (1991ch). Multiple imputation in health???are databases: An overview and some applications. *Statistics in Medicine*, *10*(4), 585–598. https://doi.org/10.1002/sim.4780100410

[87] Kwon, O., & Sim, J. M. (2013ci). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, *40*(5), 1847–1857. https://doi.org/10.1016/j.eswa.2012.09.017

[88] Sim, J., Kwon, O., & Lee, K. C. (2016cj). Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets. *Expert Systems with Applications*, *46*, 485–493. https://doi.org/10.1016/j.eswa.2015.11.004

[89] Raja, P. S., Thangavel, K., Wilson, D. R., Martinez, T. R., Pawlak, Z., Tsai, C.-F., … Jensen, R. (2016ck). Soft Clustering Based Missing Value Imputation. *Machine Learning*, *11*(5), 257–286.

[90] Jain, A. K. (2010cl). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[91] Breiman, L. (2001cm). Random forests. *Machine Learning*, *45*(1), 5–32.

[92] Hastie, T., Tibshirani, R., & Friedman, J. (2009cn). The Elements of Statistical Learning. *Springer Series in Statistics*, *20*(2), 158–61. https://doi.org/10.1007/978-0-387-84858-7

[93] Day, A. (2010co). *Mastering Financial Mathematics in Microsoft Excel: A Practical Guide for Business Calculations (The Mastering Series)*. Pearson UK.

[94] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011cp). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825–2830.

[95] Hoerl, A. E., & Kennard, R. W. (1970cq). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[96] Ju-Long, D., & Julong, D. (1989cr). Introduction to grey system theory. *Systems & Control Letters*, *1*(5), 1–24.

[97] Pan, R., Yang, T., Cao, J., Lu, K., & Zhang, Z. (2015cs). Missing data imputation by K nearest neighbours based on grey relational structure and mutual information. *Applied Intelligence*, *43*(3), 614–632.

[98] Lv, Z., Zhao, J., Liu, Y., Wang, W., Folch-Fortuny, A., Villaverde, A. F., … Banga, J. R. (2015ct). Enabling network inference methods to handle missing data and outliers. *BMC Bioinformatics*, *367*(1), 311–323.

[99] Lv, Z., Zhao, J., Liu, Y., & Wang, W. (2016cu). Data imputation for gas flow data in steel industry based on non-equal-length granules correlation coefficient. *Information Sciences*, *367*–*368*, 311–323. https://doi.org/10.1016/j.ins.2016.05.046

[100] Wang, G., Yeung, D.-Y., & Lochovsky, F. H. (2008cv). A new solution path algorithm in support vector regression. *IEEE Transactions on Neural Networks*, *19*(10), 1753–1767.

**Abstract:**

Preprocessing is one of the most important steps in the knowledge discovery process. Data cleaning task in preprocessing plays a crucial role to ensure high-quality data. Completeness is an essential aspect of high-quality data. Missing data should be imputed before they are used for data mining and machine learning algorithms. In this research, several novel imputation methods are proposed to fill the missing values. The main point in these methods is that considering partitions of data instead of the whole data set can improve the imputation results. Three different groups of approaches are proposed in this research for the missing value imputation goal. In the first group, a two-step approach based on the clustering and using a combination of weighted k nearest neighbors and linear regression methods are presented. The second group consists of ten correlation based approaches for the missing values imputation. In the third group, missing data are imputed using a Grey based Fuzzy c-Means and a Mutual Information based feature selection.

Three well-known evaluation criteria are used to assess the performance of the proposed imputation methods. The results of proposed methods are compared with those of other existing methods, considering various well-known real data sets in the machine learning and data mining fields. The experimental results show that the proposed methods perform better than other compared methods, in most of the cases.

**Keywords**: Data cleaning, Data quality, Missing values imputation, Partitioning



**Shahid Rajaee Teacher Training University**

**Faculty of Computer Engineering**

**Missing value estimation and inconsistencies detection in data improvement using data partitioning**

**By: Amir Masoud Sefidian**

**Under Supervision of Dr. Negin Daneshpour**

**A thesis submitted to the Graduate Studies Office in partial fulfillment of**  
**the requirements for the degree of M.Sc. in**

**Software Engineering**

**September 2016**

1. Transient [↑](#footnote-ref-1)
2. Raw Data [↑](#footnote-ref-2)
3. Data Mining [↑](#footnote-ref-3)
4. Knowledge Discovery from Data [↑](#footnote-ref-4)
5. Web [↑](#footnote-ref-5)
6. Pre-processing [↑](#footnote-ref-6)
7. Data Cleaning [↑](#footnote-ref-7)
8. Completeness [↑](#footnote-ref-8)
9. Noise Data [↑](#footnote-ref-9)
10. Data Integration [↑](#footnote-ref-10)
11. Data selection [↑](#footnote-ref-11)
12. Retrieved [↑](#footnote-ref-12)
13. Data Transformation [↑](#footnote-ref-13)
14. Pattern Evaluation [↑](#footnote-ref-14)
15. Knowledge presentation [↑](#footnote-ref-15)
16. Visualization [↑](#footnote-ref-16)
17. Classification [↑](#footnote-ref-17)
18. Clustering [↑](#footnote-ref-18)
19. Regression [↑](#footnote-ref-19)
20. Facts [↑](#footnote-ref-20)
21. Single Source [↑](#footnote-ref-21)
22. Multiple Sources [↑](#footnote-ref-22)
23. Imperfections [↑](#footnote-ref-23)
24. Inconsistencies [↑](#footnote-ref-24)
25. Errors [↑](#footnote-ref-25)
26. Outliers [↑](#footnote-ref-26)
27. Missing Values [↑](#footnote-ref-27)
28. Heterogeneous [↑](#footnote-ref-28)
29. Pattern Recognition [↑](#footnote-ref-29)
30. Correctness [↑](#footnote-ref-30)
31. Accuracy [↑](#footnote-ref-31)
32. Integrity [↑](#footnote-ref-32)
33. Consistency [↑](#footnote-ref-33)
34. Timeliness [↑](#footnote-ref-34)
35. Interpretability [↑](#footnote-ref-35)
36. Data Engineering [↑](#footnote-ref-36)
37. Data Preparation [↑](#footnote-ref-37)
38. Data Type Conversion [↑](#footnote-ref-38)
39. Values Normalization [↑](#footnote-ref-39)
40. Schema Evolution [↑](#footnote-ref-40)
41. Schema Constraints [↑](#footnote-ref-41)
42. Reference Constraints [↑](#footnote-ref-42)
43. Attribute [↑](#footnote-ref-43)
44. Record Type [↑](#footnote-ref-44)
45. Duplicate Records [↑](#footnote-ref-45)
46. Record Linkage [↑](#footnote-ref-46)
47. Redundancy [↑](#footnote-ref-47)
48. Naming Conflicts [↑](#footnote-ref-48)
49. Structural Conflicts [↑](#footnote-ref-49)
50. Data Scrubbling [↑](#footnote-ref-50)
51. Data Warehouse [↑](#footnote-ref-51)
52. Federated Database [↑](#footnote-ref-52)
53. Unknown [↑](#footnote-ref-53)
54. Null [↑](#footnote-ref-54)
55. Missing Value Imputation [↑](#footnote-ref-55)
56. Missing Value Estimation [↑](#footnote-ref-56)
57. Missing Value Filling [↑](#footnote-ref-57)
58. Query [↑](#footnote-ref-58)
59. Applicability [↑](#footnote-ref-59)
60. Power System Failure [↑](#footnote-ref-60)
61. Manual Data Entry Procedures [↑](#footnote-ref-61)
62. Missing Rate [↑](#footnote-ref-62)
63. Non-applicability [↑](#footnote-ref-63)
64. Generalizability [↑](#footnote-ref-64)
65. Biased Results [↑](#footnote-ref-65)
66. Correlation-based Missing value Imputation Methods [↑](#footnote-ref-66)
67. Grey based Fuzzy c-Means [↑](#footnote-ref-67)
68. Mutual Information [↑](#footnote-ref-68)
69. Missing Data Mechanism [↑](#footnote-ref-69)
70. Missing Completely At Random [↑](#footnote-ref-70)
71. Missing At Random [↑](#footnote-ref-71)
72. Not Missing At Random [↑](#footnote-ref-72)
73. Plausible [↑](#footnote-ref-73)
74. Bias [↑](#footnote-ref-74)
75. Quantitative [↑](#footnote-ref-75)
76. Mode [↑](#footnote-ref-76)
77. Qualitative [↑](#footnote-ref-77)
78. Distribution [↑](#footnote-ref-78)
79. Artificial Neural Networks [↑](#footnote-ref-79)
80. Self Organizing Map [↑](#footnote-ref-80)
81. Multi Layer Perceptron [↑](#footnote-ref-81)
82. Target Attribute [↑](#footnote-ref-82)
83. Auto Associative Neural Networks [↑](#footnote-ref-83)
84. k-Nearest Neighbors Imputation [↑](#footnote-ref-84)
85. Nominal Attributes [↑](#footnote-ref-85)
86. Grey [↑](#footnote-ref-86)
87. Grey Relational Grade [↑](#footnote-ref-87)
88. Grey-Based K-NN Iteration Imputation [↑](#footnote-ref-88)
89. Label [↑](#footnote-ref-89)
90. Grey Relational Coefficient [↑](#footnote-ref-90)
91. Manhattan [↑](#footnote-ref-91)
92. Euclidean [↑](#footnote-ref-92)
93. Class-based K-clusters Nearest Neighbor Imputation [↑](#footnote-ref-93)
94. Weighted k-Nearest Neighbors [↑](#footnote-ref-94)
95. Predictive Model [↑](#footnote-ref-95)
96. Expectation Maximization Imputation [↑](#footnote-ref-96)
97. Covariance [↑](#footnote-ref-97)
98. Decision Tree-based Missing Value Imputation [↑](#footnote-ref-98)
99. Fit [↑](#footnote-ref-99)
100. Response Variable [↑](#footnote-ref-100)
101. Dependent Variable [↑](#footnote-ref-101)
102. Target Variable [↑](#footnote-ref-102)
103. Predictors [↑](#footnote-ref-103)
104. Independent Variables [↑](#footnote-ref-104)
105. Multiple Linear Regression [↑](#footnote-ref-105)
106. Logistic Regression [↑](#footnote-ref-106)
107. Dichotomous [↑](#footnote-ref-107)
108. Multinomial Logistic Regression [↑](#footnote-ref-108)
109. Categorical [↑](#footnote-ref-109)
110. Iteratively [↑](#footnote-ref-110)
111. Association Rules [↑](#footnote-ref-111)
112. Transactional [↑](#footnote-ref-112)
113. Lift [↑](#footnote-ref-113)
114. Confidence [↑](#footnote-ref-114)
115. Support [↑](#footnote-ref-115)
116. Extended Robust Association Rules [↑](#footnote-ref-116)
117. Fuzzy [↑](#footnote-ref-117)
118. Membership Function [↑](#footnote-ref-118)
119. Sparse [↑](#footnote-ref-119)
120. Refine [↑](#footnote-ref-120)
121. Topology [↑](#footnote-ref-121)
122. Dynamic Programming [↑](#footnote-ref-122)
123. Genetic Algorithm [↑](#footnote-ref-123)
124. Support Vector Regression [↑](#footnote-ref-124)
125. Bayesian Networks [↑](#footnote-ref-125)
126. Recursive Partitioning [↑](#footnote-ref-126)
127. Crowdsourcing [↑](#footnote-ref-127)
128. Imputation Times [↑](#footnote-ref-128)
129. Single Imputation [↑](#footnote-ref-129)
130. Multiple Imputation [↑](#footnote-ref-130)
131. Classifier [↑](#footnote-ref-131)
132. Root Mean Squared Error [↑](#footnote-ref-132)
133. Coefficient of Determination [↑](#footnote-ref-133)
134. Mean Absolute Error [↑](#footnote-ref-134)
135. Weighted k-Nearest Neighbor [↑](#footnote-ref-135)
136. Linear Regression [↑](#footnote-ref-136)
137. Correlation [↑](#footnote-ref-137)
138. Pearson Correlation Coefficient [↑](#footnote-ref-138)
139. Most Correlated [↑](#footnote-ref-139)
140. Fitting [↑](#footnote-ref-140)
141. Least Square Method [↑](#footnote-ref-141)
142. Python [↑](#footnote-ref-142)
143. Min-Max [↑](#footnote-ref-143)
144. Correlation-based Missing value Imputation Methods [↑](#footnote-ref-144)
145. Forward Selection [↑](#footnote-ref-145)
146. Random Forest [↑](#footnote-ref-146)
147. Base Set [↑](#footnote-ref-147)
148. Incremental Attribute Regression Imputation [↑](#footnote-ref-148)
149. Redundant [↑](#footnote-ref-149)
150. Incremental [↑](#footnote-ref-150)
151. Maximization [↑](#footnote-ref-151)
152. Estimation [↑](#footnote-ref-152)
153. Out of bag samples [↑](#footnote-ref-153)
154. k-combination [↑](#footnote-ref-154)
155. Exhaustive Search [↑](#footnote-ref-155)
156. Linear Ridge Regression [↑](#footnote-ref-156)
157. Linear Least Squares [↑](#footnote-ref-157)
158. Cost Function [↑](#footnote-ref-158)
159. Regularization [↑](#footnote-ref-159)
160. Regularization Parameter [↑](#footnote-ref-160)
161. Overfitting [↑](#footnote-ref-161)
162. z-score [↑](#footnote-ref-162)
163. Mutual Information [↑](#footnote-ref-163)
164. Grey Systems Theory [↑](#footnote-ref-164)
165. Grey Relational Grade [↑](#footnote-ref-165)
166. Midowski [↑](#footnote-ref-166)
167. Hard Clustering [↑](#footnote-ref-167)
168. Fuzzy (Soft) Clustering [↑](#footnote-ref-168)
169. Fuzzification [↑](#footnote-ref-169)
170. Partially Unkonwn [↑](#footnote-ref-170)
171. Grey Relational Analysis [↑](#footnote-ref-171)
172. Referential Observation [↑](#footnote-ref-172)
173. Compared Observations [↑](#footnote-ref-173)
174. Grey Relational Coefficient [↑](#footnote-ref-174)
175. Wholeness [↑](#footnote-ref-175)
176. Probability Mass Function [↑](#footnote-ref-176)
177. Probability Density Function [↑](#footnote-ref-177)
178. Grey based Fuzzy c-Means [↑](#footnote-ref-178)
179. Multiple Linear Regression [↑](#footnote-ref-179)
180. Support Vector Regression [↑](#footnote-ref-180)
181. Kernel [↑](#footnote-ref-181)
182. Nonlinear Feature Map [↑](#footnote-ref-182)
183. Quadratic Programming [↑](#footnote-ref-183)
184. Penalty Parameter [↑](#footnote-ref-184)
185. Epsilon Parameter [↑](#footnote-ref-185)
186. Linear Kernel [↑](#footnote-ref-186)
187. Grey based Fuzzy c-Means Imputation [↑](#footnote-ref-187)