شرکت اعتبارسنجی
حافظ سامان ایرانیان
Iranian Credit Bureau and Scoring Co.

داده‌های پرت یا Outliers

منظور از داده‌های پرت، مقادیری هستند که خارج از محدوده طبیعی یک متغیر قرار دارند. به عنوان مثال ممکن است توزیع سن محدوده‌ای بین ۱۸ تا ۹۰ سال داشته باشد، اما تعدادی داده سن با مقادیر ۷۸۹، ۱۱۱۱ و ۱۳۴ در بین داده‌ها وجود داشته باشد. وجود چنین داده‌هایی در اکثر مواقع نتیجه عملکرد نادرست کاربران است. نظر به اینکه این داده‌ها می‌توانند نتایج تحلیل را تحت تاثیر قرار دهند، معمولا ساده‌ترین و متداول‌ترین رویکرد این است که داده‌های پرت حذف شوند. البته در بعضی موارد داده‌های پرت می‌توانند به مقادیر میانگین تخصیص داده شوند، زیرا در اغلب موارد تعداد معدودی دارند و گاهی ممکن است تاثیر معنی‌داری بر نتایج نداشته باشند. با این حال به ازای هر مطالعه روی داده‌ها، می‌بایست تمامی داده‌های پرت در ابتدا بررسی شوند، زیرا ممکن است مسائلی از قبیل تقلب و دستکاری در داده‌ها نیز وجود داشته باشد.

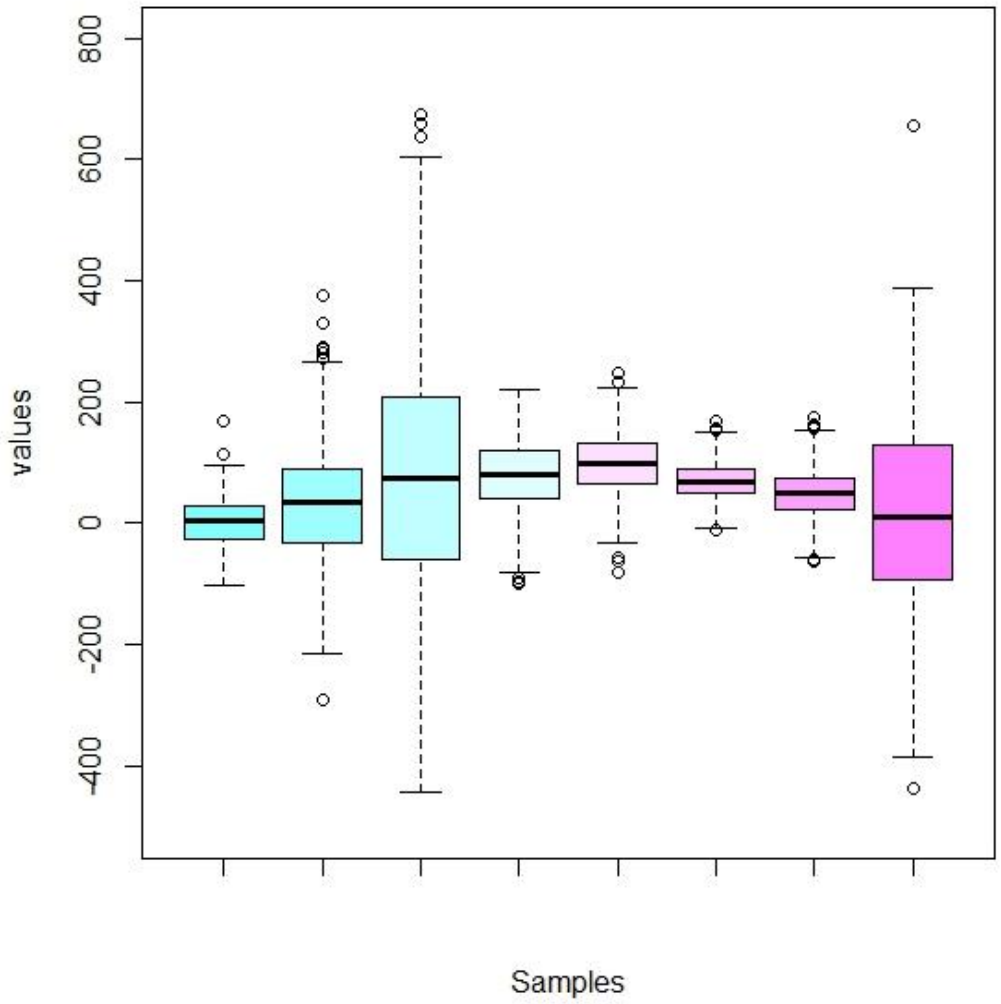
روش های مختلفی برای تشخیص داده‌های پرت وجود دارد، که می‌توان آن‌ها را به سه دسته کلی تقسیم کرد. این روش‌ها شامل تشخیص یک متغیره، دومتغیره و چندمتغیره هستند. مقاله جاری بر روی روش‌های یک متغیره متمرکز است که شامل دو گروه روش‌های دامنه توزیع و آزمون‌های آماری می‌شود. در روش دامنه توزیع، مشاهدات بررسی شده و داده‌های خارج از یک دامنه معین به عنوان داده پرت تلقی می‌شوند. مهم‌ترین موضوع در این روش تعیین دامنه یاد شده بـرای مشخص کردن داده‌های پرت است. روش سنتی در این مورد، میانگین $(\overline{X})$ به اضافه یا منهای ۳ برابر انحراف معیار (S) است که داده‌های بزرگتر از میانگین به اضافه ۳ برابر انحراف معیار و کوچکتر از میانگین منهای ۳ برابر انحراف معیار، پرت محسوب می شوند. چون ایـن روش در سایر پارامترها نیز تحت تأثیر داده‌های پرت است (در محاسبه میانگین و انحراف معیار از تمام داده‌ها از جمله داده‌های پرت استفاده می شود)، لذا روش‌های دیگری از جمله میانه به اضافه یا منهای میانه انحراف‌های

تمام داده‌ها از میانه (Median ± ۳ MAD) و نمودار جعبه‌ای (Box Plot) ارایه شده که تحت تأثیر داده‌های پرت قرار نمی‌گیرند. میانه انحراف‌های تمام داده‌ها از میانه (MAD) از رابطه زیر محاسبه می‌شود:

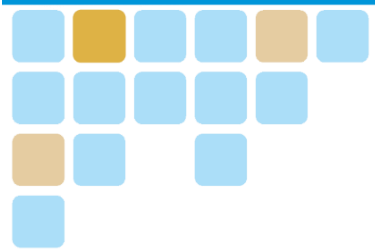
$$MAD = ۱.۴۸۲\ Median(|x_i - x_{median}|)$$

مقدار ثابت ۱/۴۸۲ برای تبدیل MAD به برآورد ناااریبی از انحراف معیار (امید ریاضی انحراف معیار نمونه برابر با انحراف معیار جامعه) داده‌های گوسی (نرمال) است.

علاوه بر رویکرد فوق، نمودار جعبه‌ای نیز از روش‌های دامنه محسوب می‌گردد. این روش نموداری برای نشان دادن موقعیت، پراکندگی و چولگی داده‌ها به‌کار می‌رود و از فراوانی برای تشخیص داده‌های پرت استفاده می‌کند. این نمودار با استفاده از یک مستطیل (Box) و دو خط یا میله در دو طرف مستطیل و به‌وسیله میانه، چارک‌های اول ($Q_۱$) و سوم ($Q_۳$) و کمترین و بیشترین مقادیر رسم می‌شود. طول مستطیل برابر با فاصله چارکی (Interquartile Range (IQR)) یعنی تفاوت بین چارک سوم و چارک اول یا $IQR = Q_۳ - Q_۱$ است. در یک نوع نمودار جعبه‌ای که از آن برای تشخیص داده‌های پرت استفاده می‌شود، داده‌هایی که کوچکتر از $Q_۱ - ۱.۵IQR$ یا بزرگتر از $Q_۳ + ۱.۵IQR$ باشند جزء داده‌های پرت خفیف و داده‌هایی که کوچکتر از $Q_۱ - ۳IQR$ یا بزرگتر از $Q_۳ + ۳IQR$ باشند جزء داده‌های پرت قوی محسوب می‌شوند. شکل زیر نمودار جعبه‌ای برای نمونه‌های مختلف را نشان می‌دهد؛ داده‌های پرت در این شکل با دایره مشخص شده‌اند.



شکل. نمودار جعبه‌ای جهت مشخص کردن داده‌های پرت (دایره‌ها)



منابع

۱. حکیم‌خانی، شاهرخ. علیجان‌پور، احمد. (۱۳۸۹). تشخیص داده‌های پرت در روش منشایابی رسوب. *مجله پژوهش‌های حفاظت آب و خاک*، ۱۷(۱)، ۲۳-۴۳.

۲. Siddiqi, N. (2017). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring (2nd ed)*. John Wiley & Sons, Inc.