تبدیل داده‌ها برای مقابله با نقاط پرت

محمدمهدی شریف بیگی

۱ مهر ۱۴۰۴

۱ مقدمه

تبدیل داده‌ها (Data Transformation) یکی از روش‌های مؤثر برای کاهش تأثیر نقاط پرت (Outliers) است. این روش با اعمال تبدیل ریاضی مناسب، مقیاس داده‌ها را تغییر داده و توزیع آن‌ها را به شکل طبیعی‌تر درآورده و تأثیر مقادیر افرطی را کاهش می‌دهد.

۲ ۱. تبدیل لگاریتمی (**Logarithmic Transformation**)

۱.۲ فرمول:

$$y = \ln(x) \quad \text{یا} \quad y = \log_{10}(x) \quad \text{یا} \quad y = \log(x)$$

۲.۲ توضیح مفصل:

تبدیل لگاریتمی یکی از پرکاربردترین روش‌های تبدیل داده است که به خصوص برای داده‌های دارای چولگی راست (right-skewed) بسیار مؤثر است.

۱.۲.۲ خصوصیات ریاضی:

برای $x > 0$، تابع لگاریتم دارای خصوصیات زیر است:

$$\frac{d}{dx}\ln(x) = \frac{1}{x} \tag{1}$$

$$\lim_{x\to\infty} \frac{\ln(x)}{x} = 0 \tag{2}$$

$$\ln(ab) = \ln(a) + \ln(b) \tag{3}$$

این خصوصیات نشان می‌دهد که:

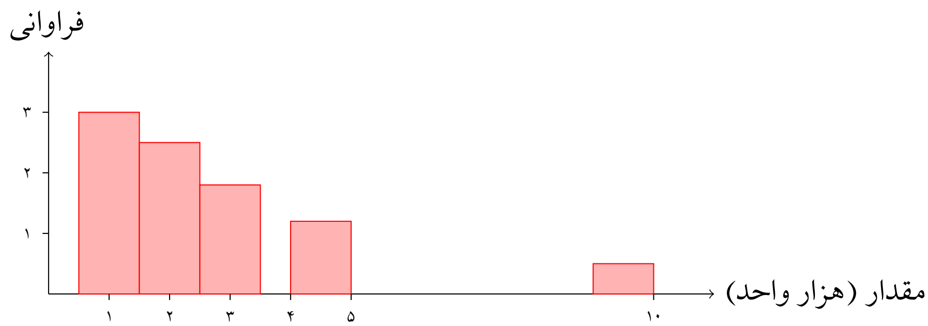
- شیب تابع با افزایش x کاهش می‌یابد
- رشد لگاریتم نسبت به x کندتر می‌شود
- ضرب در جمع تبدیل می‌شود

۲.۲.۲ مثال عملی با محاسبات کامل:

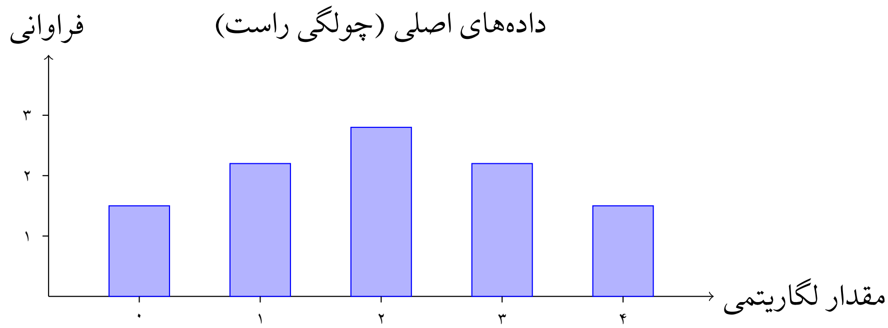
فرض کنید داده‌های اصلی درآمد (میلیون ریال): {۱, ۱۰, ۱۰۰, ۱۰۰۰, ۱۰۰۰۰} مقایسه آمارهای توصیفی

آماره	داده اصلی	داده تبدیل شده
میانگین	$\frac{۱+۱۰+۱۰۰+۱۰۰۰+۱۰۰۰۰}{۵} = ۲۲۲۲/۲$	$\frac{۰+۱+۲+۳+۴}{۵} = ۲$
میانه	۱۰۰	۲
انحراف معیار	۴۴۹۹/۴	۱/۵۸
محدوده	۱۰۰۰۰ − ۱ = ۹۹۹۹	۴ − ۰ = ۴
ضریب تغییرات	$\frac{۴۴۹۹/۴}{۲۲۲۲/۲} = ۲/۰۲۵$	$\frac{۱/۵۸}{۲} = ۰/۷۹$

۳.۲ نمودار مقایسه‌ای بهبود یافته:



داده‌های اصلی (چولگی راست)



داده‌های تبدیل شده (نرمال)

۳ .۲ تبدیل رادیکال (Square Root Transformation)

۱.۳ فرمول:

$$y = \sqrt{x}$$

۲.۳ توضیح مفصل:

تبدیل رادیکال نسبت به تبدیل لگاریتمی ملایم‌تر است و برای داده‌هایی که دارای چولگی متوسط هستند مناسب است.

۱.۲.۳ خصوصیات ریاضی:

برای $x \geq ۰$:

$$\frac{d}{dx}\sqrt{x} = \frac{۱}{۲\sqrt{x}} \tag{۴}$$

$$\lim_{x \to ۰^+} \frac{۱}{۲\sqrt{x}} = +\infty \tag{۵}$$

$$\lim_{x \to \infty} \frac{۱}{۲\sqrt{x}} = ۰ \tag{۶}$$

۲.۲.۳ مثال عملی با محاسبات کامل:

فرض کنید داده‌های اصلی مساحت (متر مربع): {۴, ۱۶, ۶۴, ۲۵۶, ۱۰۲۴} محاسبه دقیق آمارهای توصیفی

برای داده‌های اصلی: $X = \{۴, ۱۶, ۶۴, ۲۵۶, ۱۰۲۴\}$

$$\bar{X} = \frac{۴ + ۱۶ + ۶۴ + ۲۵۶ + ۱۰۲۴}{۵} = \frac{۱۳۶۴}{۵} = ۲۷۲/۸$$

$$s_X^۲ = \frac{\sum(x_i - \bar{x})^۲}{n - ۱} = \frac{(۴ - ۲۷۲/۸)^۲ + \cdots + (۱۰۲۴ - ۲۷۲/۸)^۲}{۴}$$

$$s_X^۲ = \frac{۷۲۲۶۷/۰۴ + ۶۶۰۲۲/۲۴ + ۴۳۴۳۶/۸۴ + ۳۷۹/۲۴ + ۵۶۴۸۴۰/۶۴}{۴} = \frac{۷۴۶۹۴۶}{۴} = ۱۸۶۷۳۶/۵$$

$$s_X = \sqrt{۱۸۶۷۳۶/۵} = ۴۳۲/۰۱$$

برای داده‌های تبدیل شده: $Y = \{۲, ۴, ۸, ۱۶, ۳۲\}$

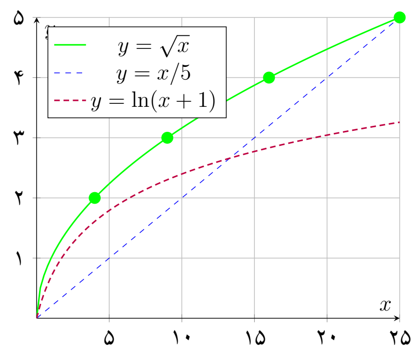
$$\bar{Y} = \frac{۲ + ۴ + ۸ + ۱۶ + ۳۲}{۵} = \frac{۶۲}{۵} = ۱۲/۴$$

$$s_Y^۲ = \frac{(۲ - ۱۲/۴)^۲ + (۴ - ۱۲/۴)^۲ + (۸ - ۱۲/۴)^۲ + (۱۶ - ۱۲/۴)^۲ + (۳۲ - ۱۲/۴)^۲}{۴}$$

$$s_Y^۲ = \frac{۱۰۸/۱۶ + ۷۰/۵۶ + ۱۹/۳۶ + ۱۲/۹۶ + ۳۸۴/۱۶}{۴} = \frac{۵۹۵/۲}{۴} = ۱۴۸/۸$$

$$s_Y = \sqrt{۱۴۸/۸} = ۱۲/۲$$

۳.۳ نمودار تابع رادیکال:



۴ ۳. تبدیل باکس‌ـ‌کاکس (Box-Cox Transformation)

۱.۴ فرمول:

$$y = \begin{cases} \frac{x^{\lambda} - ۱}{\lambda} & \text{اگر} \lambda \neq ۰ \\ \ln(x) & \text{اگر} \lambda = ۰ \end{cases}$$

۲.۴ تفسیر ریاضی مقادیر مختلف λ:

۱.۲.۴ $\lambda = ۲$ (تبدیل درجه دوم):

$$y = \frac{x^{۲} - ۱}{۲}$$

این تبدیل برای داده‌هایی که چولگی چپ دارند (left-skewed) مناسب است.

۲.۲.۴ $\lambda = ۱$ (تبدیل خطی):

$$y = x - ۱$$

این تبدیل فقط یک انتقال ساده است و شکل توزیع را تغییر نمی‌دهد.

۳.۲.۴ $\lambda = ۰/۵$ (تبدیل رادیکال):

$$y = \frac{\sqrt{x} - ۱}{۰/۵} = ۲(\sqrt{x} - ۱)$$

۴.۲.۴ $\lambda = ۰$ (تبدیل لگاریتمی):

$$y = \ln(x)$$

۵.۲.۴ $\lambda = -0.5$ (تبدیل معکوس رادیکال):

$$y = \frac{x^{-0.5} - 1}{-0.5} = 2\left(1 - \frac{1}{\sqrt{x}}\right)$$

۶.۲.۴ $\lambda = -1$ (تبدیل معکوس):

$$y = 1 - \frac{1}{x}$$

۳.۴ روش یافتن λ بهینه:

پارامتر λ از طریق بیشینه‌سازی تابع درستنمایی محاسبه می‌شود:

$$L(\lambda) = -\frac{n}{2}\ln\left(\frac{\mathrm{RSS}(\lambda)}{n}\right) + (\lambda - 1)\sum_{i=1}^{n}\ln(x_i)$$

که در آن:

$$\mathrm{RSS}(\lambda) = \sum_{i=1}^{n}\left[y_i(\lambda) - \bar{y}(\lambda)\right]^2 \tag{7}$$

$$y_i(\lambda) = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(x_i) & \lambda = 0 \end{cases} \tag{8}$$

الگوریتم محاسبه:

۱. مقادیر مختلف λ را در بازه $[-2, 2]$ با گام ۰/۱ آزمایش کنید

۲. برای هر λ، داده‌ها را تبدیل کنید

۳. $L(\lambda)$ را محاسبه کنید

۴. λ که بیشترین $L(\lambda)$ را دارد، انتخاب کنید

۴.۴ مثال عملی کامل:

داده‌های اصلی: $\{9, 25, 100, 400\}$، فرض کنید $\lambda = 0.5$ بهینه محاسبه شده است.
گام ۱: تبدیل داده‌ها

$$y_1 = \frac{9^{0.5} - 1}{0.5} = \frac{3 - 1}{0.5} = \frac{2}{0.5} = 4 \tag{9}$$

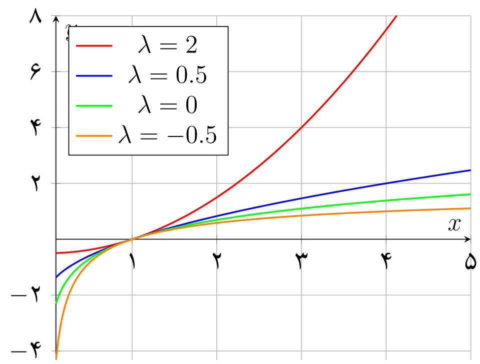
$$y_2 = \frac{25^{0.5} - 1}{0.5} = \frac{5 - 1}{0.5} = \frac{4}{0.5} = 8 \tag{10}$$

$$y_3 = \frac{100^{0.5} - 1}{0.5} = \frac{10 - 1}{0.5} = \frac{9}{0.5} = 18 \tag{11}$$

$$y_4 = \frac{400^{0.5} - 1}{0.5} = \frac{20 - 1}{0.5} = \frac{19}{0.5} = 38 \tag{12}$$

گام ۲: محاسبه آمارهای توصیفی
داده‌های اصلی: میانگین = $۱۳۳/۵ = \frac{۹+۲۵+۱۰۰+۴۰۰}{۴}$، انحراف معیار = ۱۸۱/۹
داده‌های تبدیل شده: میانگین = $۱۷ = \frac{۴+۸+۱۸+۳۸}{۴}$، انحراف معیار = ۱۵/۶
ضریب تغییرات کاهش یافته: از $۱/۳۶ = \frac{۱۸۱/۹}{۱۳۳/۵}$ به $۰/۹۲ = \frac{۱۵/۶}{۱۷}$

۵.۴ نمودار اثر مقادیر مختلف λ:



۵ مقایسه کمی سه روش

۱.۵ مثال کاربردی: داده‌های درآمد

داده‌های درآمد (میلیون تومان): $\{۲, ۵, ۸, ۱۲, ۱۵, ۲۵, ۴۵, ۱۲۰\}$

۱.۱.۵ محاسبات دقیق:

۱. آمارهای اصلی:

$$\bar{x} = \frac{۲ + ۵ + ۸ + ۱۲ + ۱۵ + ۲۵ + ۴۵ + ۱۲۰}{۸} = \frac{۲۳۲}{۸} = ۲۹ \tag{۱۳}$$

$$s^۲ = \frac{\sum(x_i - ۲۹)^۲}{۷} = \frac{۱۱۹۱۴}{۷} = ۱۷۰۲ \tag{۱۴}$$

$$s = \sqrt{۱۷۰۲} = ۴۱/۲۶ \tag{۱۵}$$

$$\text{چولگی} = \frac{\sum(x_i - ۲۹)^۳/۸}{s^۳} = ۱/۸۹ \tag{۱۶}$$

۲. تبدیل لگاریتمی:

$$y_{\text{log}} = \{\ln(۲), \ln(۵), \ln(۸), \ln(۱۲), \ln(۱۵), \ln(۲۵), \ln(۴۵), \ln(۱۲۰)\}$$

$$= \{۰/۶۹, ۱/۶۱, ۲/۰۸, ۲/۴۸, ۲/۷۱, ۳/۲۲, ۳/۸۱, ۴/۷۹\}$$

میانگین: $\bar{y}_{\text{log}} = ۲/۶۷$، انحراف معیار: $s_{\text{log}} = ۱/۳۵$، چولگی: ۰/۲۳
۳. تبدیل رادیکال:

$$y_{\text{sqrt}} = \{\sqrt{۲}, \sqrt{۵}, \sqrt{۸}, \sqrt{۱۲}, \sqrt{۱۵}, \sqrt{۲۵}, \sqrt{۴۵}, \sqrt{۱۲۰}\}$$

$$= \{۱/۴۱, ۲/۲۴, ۲/۸۳, ۳/۴۶, ۳/۸۷, ۵/۰۰, ۶/۷۱, ۱۰/۹۵\}$$

میانگین: $\bar{y}_{\text{sqrt}} = ۴/۵۶$، انحراف معیار: $s_{\text{sqrt}} = ۳/۲۴$، چولگی: ۰/۶۷

آماره	داده اصلی	لگاریتمی	رادیکال
میانگین	۰.۲۹	۶۷.۲	۵۶.۴
انحراف معیار	۲۶.۴۱	۳۵.۱	۲۴.۳
ضریب تغییرات	۴۲.۱	۵۱.۰	۷۱.۰
چولگی	۸۹.۱	۲۳.۰	۶۷.۰

۶ راهنمای عملی انتخاب روش

۱.۶ الگوریتم تصمیم‌گیری علمی:

۱. محاسبه ضریب چولگی ($\gamma_۱$):

$$\gamma_۱ = \frac{E[(X-\mu)^۳]}{\sigma^۳} = \frac{\sum_{i=۱}^{n}(x_i-\bar{x})^۳/n}{s^۳}$$

۲. قانون تصمیم‌گیری:

- اگر $|\gamma_1| < 0.5$: نیازی به تبدیل نیست
- اگر $0.5 \leq |\gamma_1| < 1$: تبدیل رادیکال
- اگر $|\gamma_1| \geq 1$: تبدیل لگاریتمی یا باکس–کاکس

۳. آزمون نرمالیتی شاپیرو–ویلک: $W = \frac{\left(\sum_{i=۱}^{n} a_i x_{(i)}\right)^۲}{\sum_{i=۱}^{n}(x_i-\bar{x})^۲}$
اگر $p\text{-value} > 0.05$، داده‌ها نرمال هستند.

۷ نتیجه‌گیری

تبدیل داده‌ها ابزار قدرتمندی برای مقابله با نقاط پرت و بهبود نرمالیتی است. انتخاب روش مناسب باید بر اساس:

- تحلیل کمی چولگی داده‌ها
- آزمون‌های آماری نرمالیتی
- ماهیت علمی متغیرها
- هدف نهایی تحلیل

همیشه قبل و بعد از تبدیل، آزمون‌های آماری مناسب را انجام دهید تا از مؤثر بودن تبدیل اطمینان حاصل کنید.