مهندسی صنایع و مدیریت شریف، (تابستان ۱۴۰۰)
دوره‌ی ۱-۳۷، شماره‌ی ۱، ص. ۷۱-۸۲، (پژوهشی)

Original Article

طراحی نمودار کنترل T^2 هتلینگ با استفاده از خوشه‌بندی

علیرضا پویا* **(استاد)**
گروه مدیریت، دانشکده‌ی علوم اداری و اقتصادی، دانشگاه فردوسی مشهد
علی یگانه (دانشجوی دکتری)
گروه مهندسی صنایع، دانشکده‌ی مهندسی، دانشگاه فردوسی مشهد
سمیه فدائی (دانشجوی دکتری)
گروه مدیریت، دانشکده‌ی علوم اداری و اقتصادی، دانشگاه فردوسی مشهد

هدف اصلی تحقیقات در زمینه‌ی کنترل فرایند آماری چندمتغیره، در نظر گرفتن همبستگی بین چندین مشخصه‌ی کیفی برای یک مرحله از فرایند است. در فاز دوم رویه‌ی کنترل فرایند چندمتغیره با استفاده ازحدود کنترلی به دست آمده از فاز اول و مشاهدات آتی، تحت کنترل بودن فرایند بررسی می‌شود. یافتن نقاط پرت فاز اول قبل از محاسبه‌ی حدود کنترلی برای حصول نتیجه‌ی مناسب اهمیت بسیار دارد. لذا در این تحقیق نمودار کنترل پیشنهادی با استفاده از روش خوشه‌بندی سلسله‌مراتبی، مشاهدات پرت را شناسایی می‌کند. اهمیت این روش در تعیین داده‌های پرت با استفاده از ضریب ناهمسانی و مجموعه‌یی از حدود کنترل متغیر است. در این روش ضریب ناهمسانی و مجموعه‌یی از حدود کنترل با استفاده از پارامترهای اندازه نمونه و تعدادمتغیرها (شاخص‌های کیفی) تعیین می‌شود، در واقع فاصله بین مشاهدات به‌صورت خوشه‌بندی مدل می‌شود و داده‌های پرت با الگوریتم بازگشتی حذف می‌شوند و سپس میانگین و ماتریس واریانس و کوواریانس T^2 بر اساس داده‌های باقیمانده تعیین می‌شود. در مرحله‌ی آخر با توجه به حد کنترل به دست آمده، آماره‌ی T^2 تعیین می‌شود. برای ارزیابی عملکرد نمودار کنترل پیشنهادی و مقایسه‌ی آن با نمودار T^2 هتلینگ معمولی در شناسایی نقاط پرت از شاخص عدم مرکزیت و روش الفارو و همکاران بر اساس تشخیص مجموعه‌ی داده‌های پرت استفاده شده است. همچنین برای بررسی و مقایسه‌ی بیشتر، دو نمودار از مجموعه داده‌های هاوکینز و فسفر نیز بررسی شده است. اهمیت و کارآیی روش پیشنهادی در شرایط وجود تعداد داده‌ی پرت زیاد در مجموعه‌ی داده‌ها، به طور محسوس مشاهده شد.

واژگان کلیدی: نمودار کنترل چندمتغیره، نمودار کنترل T_2 هتلینگ، تحلیل خوشه‌یی سلسله‌مراتبی.

alirezapooya@um.ac.ir
ali.yeganeh@mail.um.ac.ir
somayehfadaei@mail.um.ac.ir

۱. مقدمه

امروزه کیفیت به‌عنوان راهبردی تجاری شناخته می‌شود و عنوان غول صنعتی نه به کشورهایی با تولید انبوه، بلکه به کشورهایی با بالاترین درجه‌ی کیفیت در محصولات، صنایع و خدمات‌رسانی اطلاق می‌شود. حفظ دستاوردهای مرحله‌ی بهینه‌سازی قبل از ساخت در مرحله‌ی حین ساخت توسط فنون آماری به‌نام «نمودارهای کنترل» در مبحث کنترل آماری فرایند انجام می‌گیرد. کنترل فرایند آماری (SPC) به طور گسترده به‌عنوان یک روش قدرتمند برای اندازه‌گیری، طبقه‌بندی، تجزیه و تحلیل و تفسیر داده‌های فرایند برای بهبود کیفیت محصولات و خدمات با شناسایی ناپایداری و توجیه علل احتمالی به طور گسترده‌یی مورد استفاده قرار می‌گیرد.[۱] ابزار اصلی SPC نمودارهای کنترل است.[۲] هدف اولیه‌ی استفاده از نمودارهای کنترل، شناسایی هرچه سریع‌تر انحرافات بادلیل است. شوهارت[۳] استفاده از یک نمودار کنترل تک‌متغیره را برای نظارت بر ویژگی کیفیت واحد پیشنهاد کرد. با این حال، نمودارهای کنترل تک‌متغیره یک محدودیت اساسی دارند زیرا: ۱. آنها رابطه‌ی بین‌متغیرها را در فرایندهای چندمتغیره نشان نمی‌دهند؛[۴] ۲. استفاده از چندین نمودار کنترل، خطای نوع ۱ را طبق فرمول $\alpha_{overral} = ۱ - \Pi_{i=۱}^{p}(۱ - \alpha_i)$ افزایش می‌دهد. افزایش نرخ هشدار اشتباهی نمودار کنترل، باعث عملکرد ضعیف نمودار کنترل خواهد شد. همچنین سیستم‌های تولید مدرن معمولاً شامل تعدادی ویژگی‌های کیفی هستند که می‌توانند با یکدیگر همبستگی پیدا کنند. بنابراین، استفاده

* نویسنده مسئول
تاریخ: دریافت ۱۳۹۹/۳/۶، اصلاحیه ۱۴۰۰/۱/۱۷، پذیرش ۱۴۰۰/۲/۲۷.
DOI:10.24200/J65.2021.55182.2089

از نمودارهای کنترل تک‌متغیره در فرایندهای چندمتغیره ممکن است منجر به نتایج نامطلوب شود.[۵و۶]

هدف اصلی نمودارکنترل چندمتغیره، شناسایی وجود علل تغییر در ویژگی‌های کیفیت چندمتغیره است. به طور خاص، برای تجزیه و تحلیل فازیک، از یک مجموعه داده‌ی تاریخی، دو هدف دنبال می‌شود: الف) برای شناسایی تغییرات در بردار میانگین که ممکن است برآورد میانگین بردارکنترل و ماتریس کوواریانس را تحریف کند؛ ب) شناسایی و از بین بردن نقاط پرت چندمتغیره.[۷]

در داده‌های یک‌متغیره، تشخیص داده‌ی پرت ساده است. در حالت دومتغیره، هنگام نمایش پراکندگی داده‌ها، یک داده ممکن است بدون وجود مشکل غیرضروری، پرت به نظر برسد. در داده‌های چندمتغیره، هنگامی که مجموعه داده بیش از دو بعد داشته باشد، شناسایی داده‌ی پرت توسط متدهای بصری امکان‌پذیر نیست. هتلینگ روش شناخته شده‌ی چندمتغیره‌یی است که به طورگسترده در نمودارکنترل T^2 در مجموعه داده‌های چندبعدی به کار رفته است.[۸]

در نمودارکنترل T_2 فرض بر این است که داده‌ها دارای توزیع نرمال چندمتغیره با میانگین μ و واریانس σ است. اگر فرایند شامل p شاخص کیفی با تعداد m مشاهده‌ی n تایی باشد، آماره چنین تعریف می‌شود:

$$T^2 = n(x-\bar{x})' S^{-1}(x-\bar{x}) \qquad (1)$$

که در آن، $x = [x_1.x_2.\ldots.x_p]^T$ و $\bar{x}$ و S برآوردی از μ و $\sum$ است. $\bar{x}$ و S مطابق رابطه‌ی ۲ و ۳ برآوردی از میانگین کل نمونه و ماتریس کوواریانس است.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{m} x_i \qquad (2)$$

$$\sum = \frac{1}{n-1}\sum_{i=1}^{m}(X_i-\bar{X})(X_i-\bar{X})' \qquad (3)$$

در حالتی که $n = 1$ است، میسون و یانگ[۹] حدود کنترلی فاز ۱ را برمبنای توزیع β به‌صورت رابطه‌ی ۴ برای نمودارکنترل T^2 در نظرگرفته‌اند:

$$\begin{cases} UCL = \frac{p(m-1)^2}{m}\beta_\alpha, p/2, m-p-1/2 \\ LCL = 0 \end{cases} \qquad (4)$$

روش کلاسیک T^2 برای تشخیص تغییر شیفت میانگین و نقاط پرت که در رابطه‌ی ۱ توضیح داده شده است در صورتی که مجموعه داده فقط دارای یک داده پرت باشد، مناسب است.[۱۰] با این وجود، هنگامی که بیش از یک مشاهده پرت در مجموعه داده‌ها وجود داشته باشد، قدرت این تکنیک‌ها به طور قابل توجهی کاهش می‌یابد.[۱۱و۱۲] به عبارتی هنگامی که بیش از یک داده‌ی پرت در فرایند وجود داشته باشد، به دلیل اثر درون‌آوری[1] و برون‌بری،[2] تشخیص نقاط پرت مشکل می‌شود.[۱۳] درون‌آوری هنگامی اتفاق می‌افتد که تشخیص مشاهدات پرت واقعی به‌عنوان پرت اعلام نمی‌شود. در حالی که برون‌بری هنگامی اتفاق می‌افتد که مشاهدات به‌صورت نادرست به‌عنوان پرت اعلام شود.[۱۴] داده‌های پرت، در صورت عدم‌کشف، قابلیت انحراف میانگین و ماتریس کوواریانس نمودارهای کنترل چندمتغیره‌ی T^2 هتلینگ برای نظارت بر ویژگی‌های کیفیت فردی را دارند. تأثیر این تحریف این است که نمودارکنترل ساخته شده از آن غیرقابل اعتماد می‌شود، زیرا دارای اثر درون‌آوری و برون‌بری است.[۱۰] بنابراین، شناسایی نقاط پرت قبل از تجزیه و تحلیل داده‌ها بسیار مهم است به‌خصوص در نظارت برکیفیت محصول فرایند تولید.[۸]

پدیده‌ی درون‌آوری و برون‌بری اتفاق می‌افتد چون میانگین و ماتریس واریانس و کوواریانس باثبات نیستند.[۱۰] یکی از راه‌های حل این مشکلات، استفاده از برآوردگرهای باثبات برای معادلات ۲ و ۳ به‌عنوان گزینه‌یی برای موقعیت و ماتریس کوواریانس است. این برآوردگرها شامل روش پیراسته[3] با بهبود T^2 هتلینگ،[۱۵و۱۶] روش اختلاف پی‌درپی سالیوان و وودال[۱۷] است که SW۱ نامیده می‌شوند و اصلاح طرح استالاکتیت[4] انکینسون و مولیرا[۱۸] به‌عنوان رویکرد دوم سالیوان و وودال، که آن را SW۲ می‌گویند. روش‌های ذکر شده در واقع می‌توانند T^2 هتلینگ را بهبود بخشند، اما هنوز در تشخیص تعداد زیادی از نقاط پرت موجود در داده‌ها مؤثر نیستند.[۸]

رسئو[۱۹] دو روش حداقل حجم بیضی‌وار (MVE) [5] و حداقل روش تعیین‌کننده‌ی کوواریانس (MCD) [6] را برای مسئله‌ی تشخیص تعداد زیادی از نقاط پرت پیشنهاد کردند. با این حال، هر دو روش نیاز به محاسبات بسیار سنگین دارد. در دو دهه‌ی گذشته بحث‌های بسیاری در مورد تقریب MVE و MCD انجام شده است. اما در دهه‌ی اخیر، در برخی از مطالعات استفاده از روش‌های تحلیل‌خوشه‌یی برای بهبود نمودارکنترل چندمتغیره آغاز شده است. برای مثال اونگ و الی[۱۰] از تحلیل خوشه‌یی رگرسیون، کانگ و بام کیم[۲۰] از تحلیل خوشه‌یی k-means، فان و همکارانش[۸] از تحلیل خوشه‌یی سلسله‌مراتبی برای نمودار کنترل چندمتغیره SW۲ استفاده کرده‌اند. رویکردهای خوشه‌بندی معمولاً از لحاظ محاسباتی کارآمدتر از روش‌های نمودار کنترل است.[۸] لذا با توجه به این که تحلیل خوشه‌یی نسبت به داده‌های پرت حساس است،[۸] در این پژوهش از تکنیک خوشه‌بندی سلسله‌مراتبی در طراحی نمودار کنترل در فاز اول، برای به دست آوردن مجموعه‌یی باثبات از مشاهدات استفاده می‌شود و تمرکز بر استفاده از رویکردهای خوشه‌یی به‌عنوان فیلتر اولیه‌ی داده‌های پرت است.

یکی از عوامل رایج خارج ازکنترل شدن نمودار، رخ دادن چندین داده‌ی پرت است. برآورد پارامترهای μ و $\sum$ مخصوصاً در نمودارکنترل T^2 هتلینگ، نسبت به یک مشاهده‌ی پرت حساس‌اند اما آماره‌های به دست آمده از نمودار کنترل مذکور، بر اساس این برآوردگرها وقتی که چندین مشاهده پرت وجود داشته باشد، بسیار ضعیف عمل می‌کند.[۲۱] همانطورکه گفته شد در پژوهش‌های زیادی برای بهبود نمودارکنترل T^2 هتلینگ، اقدام شده است، اما هنوز در تشخیص تعداد زیادی از نقاط پرت موجود در داده‌ها مؤثر نیستند.[۸] لذا دلیل استفاده از تحلیل خوشه‌یی سلسله‌مراتبی در تحقیق حاضر این است که خوشه‌بندی سلسله‌مراتبی امکان استفاده از فاصله‌های مختلف در بین داده‌ها را دارد. همچنین چون تمامی داده‌ها با یکدیگر مقایسه می‌شوند، امکان شناسایی نقاط پرت با الگوهای مختلف فراهم می‌شود.[۱۰]

برای ارزیابی عملکرد نمودارکنترل پیشنهادی ابتدا بر اساس تشخیص یک داده‌ی پرت و با استفاده از شاخص عدم مرکزیت وارگاس[۲۲] استفاده می‌شود. شاخص عدم مرکزیت (ncp) نشان‌دهنده‌ی شدت تغییر بردار میانگین خارج از کنترل (μ_1) از بردار میانگین تحت کنترل میانگین (μ_0) مطابق رابطه‌ی ۱ است. وارگاس معیار دقت روش را تشخیص دست‌کم یک سیگنال در داده‌ها دانسته است.

$$ncp = (\mu_1-\mu_0)' \sum^{-1} (\mu_1-\mu_0) \qquad (5)$$

در این تحقیق علاوه بر روش ارزیابی وارگاس، از ارزیابی عملکرد بر اساس تشخیص مجموعه‌ی داده‌های پرت الفارو و ارتگا[۱۱] نیز استفاده می‌شود. در این

روش اگر تمامی داده‌های پرت شناسایی شوند، دقت روش مورد تأیید است. سپس با استفاده از دو مجموعه داده‌ی واقعی فسفر و هاوکینز[۸] نمودار کنترل T^2 معمولی با نمودار T^2 خوشه‌بندی مقایسه می‌شود.

در ادامه‌ی این نوشتار، در بخش دوم پیشینه‌ی مطالعاتی نمودار کنترل T^2 و تلاش‌هایی که برای بهبود این نمودار انجام شده بررسی می‌شود. در بخش سوم روش پیشنهادی تحقیق حاضر معرفی خواهد شد. در بخش چهارم ارزیابی عملکرد نمودار کنترل پیشنهادی (T^2 خوشه‌بندی) با استفاده از روش وارگاس و روش الفارو و ارتگا بررسی و مقایسه می‌شود؛ سپس با استفاده از مجموعه داده‌های واقعی فسفر و هاوکینز، نسبت به ارزیابی نمودار کنترل T^2 خوشه‌بندی و مقایسه‌ی آن با نمودار T^2 هتلینگ معمول اقدام می‌شود. در بخش پایانی به بحث و نتیجه‌گیری مطالب به دست آمده خواهیم پرداخت.

۲. پیشینه‌ی تحقیق

تاکنون روش‌های برآورد باثبات متفاوتی برای شناسایی نقاط پرت و غیرمعمول چندمتغیره در فاز اول نمودارهای کنترلی مطرح شده است که مهم‌ترین و کاراترین آنها در ادامه معرفی می‌شود. رسئو[۱۹] دو روش حداقل حجم بیضی‌وار (MVE) و حداقل دترمینان کوواریانس (MCD) را برای شناسایی نقاط پرت در داده‌های چندمتغیره پیشنهاد کرد. برآوردگر حداقل حجم بیضی‌وار (MVE) اولین برآوردگر باثبات بر اساس موقعیت و پراکندگی چندمتغیره است. در حالی که ایده‌ی MVE بسیار شهودی است، یافتن برآوردگر MVE در عمل می‌تواند بسیار دشوار باشد. با افزایش اندازه‌ی نمونه و ابعاد داده‌ها محاسبات به طور نمایی افزایش می‌یابد؛ به همین دلیل رسئو و لئوری[۲۳] یک روش تقریبی برای یافتن برآوردگرهای MVE با استفاده از الگوریتم زیرنمونه‌گیری[۷] را پیشنهاد دادند. دیگر برآوردگر بردار میانگین و ماتریس کوواریانس که می‌تواند تا ۵۰٪ نقاط پرت را در مجموعه داده اصلاح کند کم‌ترین برآوردگر تعیین کوواریانس (MCD) است. روسو و ون‌درایسین[۲۴] الگوریتمی سریع برای محاسبه‌ی برآوردگر تعیین کوواریانس پیشنهاد دادند و آن را FastMCDنامیدند؛ این الگوریتم قابلیت کنترل مجموعه داده‌های بزرگ در یک زمان معقول را دارد. جنسن و همکاران[۲۵] دو روش MVE و MCD را در داده‌های چندمتغیره‌ی زیادی مقایسه کردند و نتیجه گرفتند هر دو روش در تشخیص نقاط پرت مؤثرند، اما عملکرد این دو رویکرد به اندازه‌ی نمونه و نسبت نقاط پرت موجود بستگی دارد. ویلمز و همکارانش[۲۶] الگوریتم ترکیبی حداقل دترمینان کوواریانس بازموزون (RMCD) [۸] را پیشنهاد کردند که از جایگزینی برآوردگرهای موقعیت و پراکندگی حداقل دترمینان کوواریانس بازموزون با میانگین و واریانس کلاسیک آماره‌ی T^2 هتلینگ به دست می‌آید. به طور مشابه وریات و وتاتور[۲۷] برآوردگرهای باثبات بر اساس بر اساس حداقل دترمینان کوواریانس بازموزون (RMCD) و حداقل حجم بیضی‌وار بازموزون (RMVE) برای نظارت بر مشاهدات چندمتغیره در داده‌های فاز I پیشنهاد کردند.

سالیوان و وودال[۱۷] دو الگوریتم با دو دیدگاه متفاوت ارائه کردند. آن‌ها نشان دادند که نمودار T^2 هتلینگ با استفاده از ماتریس کوواریانس نمونه، در شناسایی تغییرات بردار میانگین مؤثر نیست، از این رو آن‌ها بر اساس نمودار استالاکتیت اتکینسون و مولیرا، پیشنهاد استفاده از اختلاف برداری بین مشاهدات پی‌درپی برای برآورد ماتریس کوواریانس در کنترل روند را دادند.[۱۸] با این وجود وارگاس[۲۲] در یک مطالعه‌ی مقایسه‌یی T^2 با استفاده از تکنیک‌های شبیه‌سازی عملکرد نمودار کنترلی را بر اساس تشخیص یک داده‌پرت و با استفاده از شاخص عدم مرکزیت (ncp) بررسی کرد. نتایج حاصله و ارزیابی عملکرد نشان داد که نمودار کنترل T^2 با استفاده از برآوردگرهای حداقل حجم بیضی (MVE) در شناسایی تعداد نقاط پرت مؤثر است و روش SW۱ در شناسایی تعداد زیادی از نقاط انحرافی مؤثر نیست. روش SW۲ نیز در مقابل داده‌هایی با نقاط دورافتاده زیاد، آسیب‌پذیر است و همچنین به نمونه‌ی تصادفی اولیه وابسته است.

الفارو و ارتگا[۱۶] یک برآوردگر باثبات به جای T^2 با استفاده از روش پیراسته پیشنهاد کردند. این نمودار کنترل جدید از ساختار مشابه T^2 برخوردار است، اما بردار موقعیت و برآورد ماتریس کوواریانس با استفاده از پیرایش به دست می‌آید. آن‌ها رفتار و عملکرد این دو نمودار را با استفاده از یک مثال و یک مطالعه شبیه‌سازی بر اساس تشخیص مجموعه‌ی داده‌های پرت مقایسه کردند. یک سال بعد این دو پژوهش‌گر[۱۱] یک مطالعه‌ی شبیه‌سازی برای تحلیل عملکرد برآوردگرهای MVE، MCD، MCD بازموزون و برآورگر پیراسته را در موقعیت‌های مختلف انجام دادند. احسان و همکاران[۱۴] نمودار T^2 مبتنی بر ترکیب PCA با حدود کنترل چگالی کرنل برای داده‌های پیوسته و دسته‌یی مختلط ارائه کردند.

پیشینه‌ی مطالعات نشان می‌دهد که در دهه‌ی اخیر استفاده از روش‌های تحلیل‌خوشه‌یی برای بهبود نمودار کنترل چندمتغیره برای شناسایی نقاط پرت و غیرمعمول چندمتغیره، افزایش یافته است. انکارا و ویرل[۲۸] نمودار کنترل R اصلاح شده را با استفاده از روش خوشه‌یی تک‌پیوند ارائه دادند. جوب و پاکجوبی[۲۹] یک روش مبتنی بر خوشه‌ی چندمرحله‌یی رایانه‌یی را پیشنهاد کردند. آنان چند سال بعد نیز برای غلبه بر این اثر درون‌آوری، یک رویکرد مبتنی بر خوشه رایانه‌یی پیشنهاد کردند[۳۰] که آماره‌ی mcd بازموزون را با یک الگوریتم مبتنی بر خوشه‌ی چندمرحله‌یی ترکیب می‌کند. فان و همکارانش[۸] نمودار کنترل چندمتغیره‌ی SW۲ برای تشخیص پرت‌ها با استفاده از درخت خوشه‌ی سلسله‌مراتبی ارائه دادند که به طور مؤثر بتواند داده‌های پرت شناسایی کند. کانگ و بام کیم[۲۰] نمودار کنترل مبتنی بر الگوریتم خوشه‌یی k-means برای فرایندهای TFT-LCD با توزیع غیرهمگن ارائه کردند. اونگ و الی[۱۰] نمودار کنترل را بر اساس تنظیم خوشه‌یی ـ رگرسیون برای نظارت بر خصوصیات کیفیت فردی در یک محیط چندمتغیره ارائه کرده‌اند. هوانگ و همکاران[۳۱] یک الگوریتم تشخیص خوشه‌ی پرت به نام ROCF را بر اساس مفهوم گراف همسایه متقابل و با این عقیده که معمولاً اندازه‌ی خوشه‌های دورافتاده بسیار کوچک‌تر از خوشه‌های معمولی است، پیشنهاد کردند. بررسی پیشینه نشان می‌دهد که تحقیقات متعددی با روش‌های متفاوت اقدام به بهبود و توسعه‌ی نمودار کنترل T^2 هتلینگ پرداخته‌اند و در دهه‌ی اخیر ترکیب روش‌های تحلیل خوشه‌یی با نمودارهای چندمتغیره برای بهبود نمودار کنترل در شناسایی نقاط پرت، افزایش یافته است. برای مثال فان و همکارانش[۸] از خوشه‌بندی برای بهبود برآوردگر سالیوان ۲ (SW۲) استفاده کرده‌اند. جوب و پاکجوبی[۳۰] از الگوریتم خوشه‌بندی چندمرحله‌یی برای بهبود MCD رسئو استفاده کردند. همچنین اونگ و الی [۱۰] از رگرسیون خوشه‌یی برای بهبود نمودار کنترل T^2 استفاده کرده‌اند. مشاهده می‌شود در این پژوهش‌ها از خوشه‌بندی برای بهبود نوع متفاوتی از نمودار کنترل T^2 اقدام شده است؛ در حالی که در پژوهش حاضر از تحلیل خوشه‌یی سلسله‌مراتبی برای بهبود برآوردگر T^2 استفاده شده است. در این روش به‌صورت بازگشتی نقاط شناسایی شده پرت توسط الگوریتم ترکیبی پیشنهاد شده، شناسایی

می‌شوند. لذا استفاده از این ساختار ترکیبی در شناسایی داده‌های پرت و استفاده از حدود کنترل متغیر متناسب با تعداد داده‌های پرت انتخاب شده، نوآوری تحقیق حاضر است.

۳. معرفی روش پیشنهادی

۱.۳. مبانی روش پیشنهادی

اساس روش پیشنهادی در این پژوهش، خوشه‌بندی سلسله‌مراتبی مشابه روش فن و همکاران[۸] است. در خوشه‌بندی سلسله‌مراتبی داده‌ها بر اساس فاصله تفکیک می‌شوند و درنهایت خروجی خوشه‌بندی سلسله‌مراتبی، نمودار دندوگرام داده‌ها خواهد بود. بر اساس دندوگرام داده‌ها می‌توان تعداد خوشه‌ها را مشخص کرد. در خوشه‌بندی سلسله‌مراتبی نوع فاصله و نحوه‌ی محاسبه‌ی فاصله‌ی خوشه‌ها باید تعیین شود. در این مدل از فاصله‌ی اقلیدسی و فاصله‌ی بین خوشه‌ی مینیمم (single) استفاده شده است.

برای تعیین وضعیت خوشه‌ها شاخص‌های متفاوتی ارائه شده که یکی از پرکاربردترین شاخص‌ها، ضریب ناهمسانی خوشه‌هاست. این ضریب به‌وسیله‌ی ماتریس ناهمسانی که دارای $m-1$ سطر (m تعداد داده‌هاست) و ۴ ستون است بیان می‌شود.[۳۲] ستون اول این ماتریس میانگین فاصله‌ی داده‌های بین دو خوشه در بالاترین سطح است. ستون دوم انحراف معیار فاصله‌ها و ستون سوم تعداد فاصله‌ی موجود بین دوخوشه و ستون چهارم ضریب ناهمسانی است. هر چه این ضریب مقدار کم‌تری باشد یعنی شباهت دو خوشه کم‌تر است و خوشه‌بندی مناسبی صورت پذیرفته‌است. برای توضیحات بیش‌تر می‌توان به مرجع مربوطه مراجعه کرد.[۳۲]

در این پژوهش پس از انجام خوشه‌بندی سلسله‌مراتبی داده‌ها به دو خوشه تقسیم می‌شوند. سپس ضریب ناهمسانی برای دوخوشه‌ی نهایی (سطر آخر ماتریس ناهمسانی) محاسبه و با حد مجاز آن که بر اساس سیگنال خطا تعیین می‌شود مقایسه می‌شود. اگر از حد مجاز کم‌تر باشد خوشه‌بندی تأیید می‌شود و در غیر این صورت داده‌ی پرت وجود ندارد. در صورت تأیید خوشه‌بندی، خوشه با تعداد داده‌ی کم‌تر، معادل داده‌های پرت در نظرگرفته شده و خوشه‌ی داده‌های اولیه‌ی پرت نامیده می‌شود. درگام دوم برای تمامی داده‌ها آماره‌ی T^2 محاسبه می‌شود و داده‌ی دارای بیشینه مقدار T^2 به‌عنوان داده‌ی انتخاب شده‌ی پرت در نظر گرفته می‌شود. اگر این داده در خوشه‌ی داده‌های اولیه‌ی پرت وجود داشت داده‌ی پرت ثانویه در نظر گرفته می‌شود و از مجموعه‌ی داده‌ها کنارگذاشته می‌شود. الگوریتم فوق دوباره تکرار می‌شود تا داده‌ی پرت ثانویه‌ی جدید ایجاد نشود. در انتهای مرحله‌ی فوق، داده‌ها به دو دسته‌ی داده‌های پرت ثانویه و داده‌های اصلی محاسباتی تقسیم می‌شوند. داده‌های پرت ثانویه از مجموعه‌ی داده‌ها کنارگذاشته می‌شود و با توجه به داده‌های اصلی محاسباتی میانگین و ماتریس واریانس کوواریانس مطابق رابطه‌ی ۲ و ۳ ایجاد می‌شود.

در مرحله‌ی بعدی با توجه به میانگین و ماتریس کوواریانس ایجاد شده بر اساس داده‌های اصلی محاسباتی، آماره‌ی T^2 برای تمامی داده‌ها محاسبه و با حد مجاز آن مقایسه می‌شود. اگر از حد مجاز آن بیش‌تر شود، آن داده پرت قطعی شناخته می‌شود. در قسمت طراحی پارامترهای مدل نحوه‌ی تعیین حدود در روش تشریح شده است.

در نهایت ملاک نهایی تصمیم‌گیری برای داده‌ی پرت در این روش، داده‌ی پرت قطعی خواهد بود. بر این اساس اگر داده‌یی در رده پرت ثانویه تعیین شد ولی پس از محاسبه‌ی آماره‌ی مربوط به آن، پرت قطعی نبود (از حد ایجاد شده‌ی نمودار کم‌تر بود) آن داده پرت محسوب نمی‌شود.

۲.۳. طراحی پارامترهای مدل

دو پارامتر ضریب ناهمسانی و حدود کنترل باید در این روش طراحی شود. ابتدا باید پارامتر ضریب ناهمسانی مجاز را برای تأیید خوشه‌بندی تعیین کرد. برای تعیین این پارامتر برای هر حالت m و p مشخص، داده‌های نرمال با میانگین صفر و ماتریس واریانس ـ کوواریانس همانی شبیه‌سازی می‌شود. میانگین ضریب ناهمسانی برای خوشه‌های نهایی (سطر آخر ماتریس ناهمسانی) به‌عنوان این حد تعیین می‌شود.

در مرحله‌ی دوم باید حدود کنترل تنظیم شود. در این روش از حدود مجاز متغیر برای داده‌ها استفاده شده است. این حدود بر اساس تعداد داده‌ی موجود در خوشه‌ی داده‌های پرت ثانویه تغییر می‌کند. به‌طورکلی می‌توان به‌صورت نظری برمبنای توزیع احتمال یا با شبیه‌سازی، حدود کنترل (LCL و UCL) را تعیین کرد.[۲۷] حد پایین در روش T^2 صفر در نظرگرفته می‌شود، اما برای حد بالا در این روش به‌ازای هر حالت تعداد داده‌ی موجود در خوشه‌ی داده‌های اولیه یک مقدار در نظر گرفته می‌شود به‌طوری که نسبت تعداد سیگنال در هر حالت به تعداد هر حالت از داده‌ی موجود در خوشه‌ی اولیه مساوی شود. به‌عنوان مثال ورگاس برای آماره‌ی T^2 با پارامترهای $m=۳۰$ و $p=۲$ حد بالای ۱۰/۵ را با شبیه‌سازی برای نسبت سیگنال خطای ۰/۰۵ محاسبه کرده است.[۲۷] این بدان معناست که در صورت شبیه‌سازی در حالت میانگین ۰ و ماتریس واریانس همانی در ۵ درصد مواقع نقاط بالای حد کنترل قرار می‌گیرد و سیگنال خطای اشتباه صادر می‌شود.

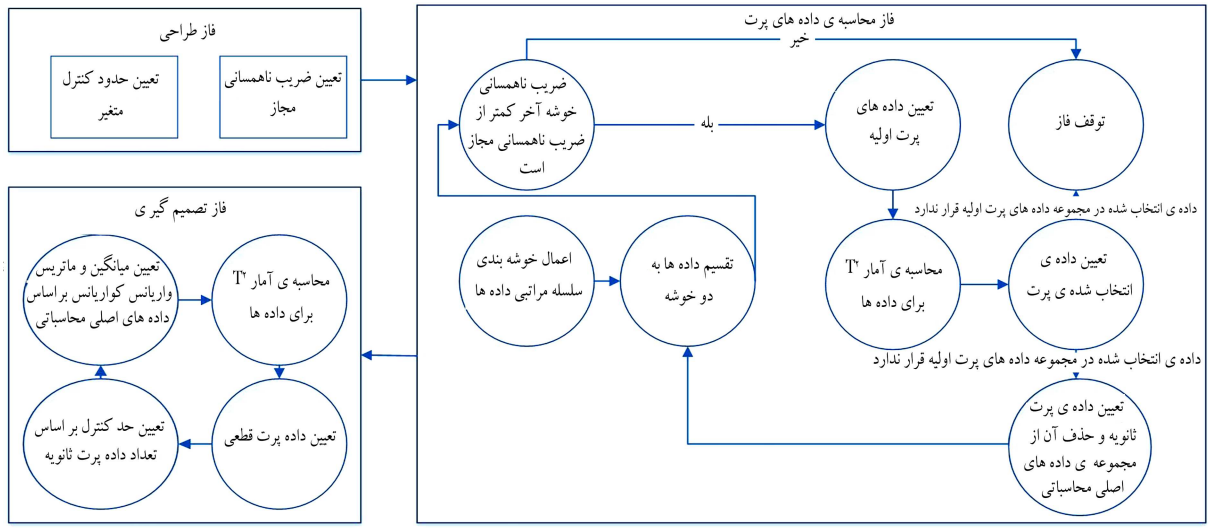
در حقیقت در این روش بر اساس تعداد داده‌ی ثانویه‌ی پرتِ شناسایی شده (به‌وسیله‌ی خوشه‌بندی سلسله‌مراتبی) حد کنترل بالای نمودار T^2 تغییر می‌کند. برای حالت $m=۳۰$ و $p=۲$ در جدول ۱ محاسبات تعیین حدود کنترل متغیر بر اساس ضریب ناهمسانی ۱/۲ براساس ۵۰۰۰ شبیه‌سازی ارائه شده است. در سطر اول نشان داده شده که در هر مرحله از شبیه‌سازی چند داده‌ی پرت ثانویه شناسایی شده و بر اساس آن آماره‌ی T^2 ساخته شده است. در سطر دوم در صورتی که بخواهیم ۰/۰۵ خطا داشته باشیم، تعداد سیگنال مجاز بیان شده است (از ضرب سطر اول در ۰/۰۵ سطر دوم حاصل می‌شود). در حقیقت این نسبت بیان می‌کند که از هر حالت شبیه‌سازی چندبار باید سیگنال اشتباه صادر شود.

به‌عنوان مثال در ۵۰۰۰ شبیه‌سازی، در ۱۷۶۸ حالت روش هیچ داده‌ی پرتی را در حالت اولیه شناسایی نمی‌کند و برای این حالت حد کنترل ۸/۸۷، در ۸۰ حالت نمودار سیگنال خارج کنترل برای داده‌های با $m=۳۰$ و $p=۲$ صادر می‌کند. در حالت ایده‌آل باید این اعداد بر روی ۸۸/۴ تنظیم شود که امکان‌پذیر نیست. دقت شود که سایر حالت‌های بزرگ‌تر از ۸ داده‌ی پرت اولیه به‌دلیل کم بودن احتمال رخ دادن در نظر گرفته نمی‌شود و حد کنترل آن‌ها نیز معادل حد کنترل ۲۸/۵ در نظر گرفته می‌شود. با این رویه به ازای حالت‌های مختلف از داده اولیه پرت، حدود کنترل متغیر تعریف می‌شود. توجه شود که تعداد حالات شبیه‌سازی در سطر اول باید دقیقاً برابر ۵۰۰۰ شود که سایر حالات که بیش از ۸ داده پرت اولیه در خوشه‌بندی شناخته شده است. برای اختصار جدول ارائه نشده است رویه‌ی فوق برای هر یک از حالات m و p باید انجام گیرد و حدکنترل مورد نظر تعیین شود.

این ساختار ترکیبی و استفاده از حدود کنترل متغیر متناسب با تعداد داده‌های پرت انتخاب شده نوآوری و مزیت اصلی این پژوهش است. بر این اساس، در این روش هدف اصلی این است که با توجه به خوشه‌بندی صورت گرفته، به تعدادی

جدول ۱. مراحل طراحی حد کنترل به ازای نسبت سیگنال اشتباهی ۰٫۰۵ برای m و p برابر ۳۰ و ۲ و ضریب ناهمسانی ۱٫۲.

	تعداد داده‌های پرت ثانویه									
	۰	۱	۲	۳	۴	۵	۶	۷	۸	جمع کل
تعداد حالت ایجاد شده در شبیه‌سازی	۱۷۶۸	۱۳۲۵	۸۳۸	۴۸۹	۲۶۳	۱۲۸	۷۸	۴۷	۲۵	۴۹۶۱
تعداد سیگنال خطای مجاز	۸۸٫۴	۶۶٫۲۵	۴۱٫۹	۲۴٫۴۵	۱۳٫۱۵	۶٫۴	۳٫۹	۲٫۳۵	۱٫۲۵	۲۴۸٫۰۵
تعداد سیگنال خطا براساس حدود تعیین شده	۸۰	۶۱	۲۸	۲۵	۱۰	۱۹	۸	۱۶	۷	۲۵۴
حد بالای کنترل	۸٫۸۷	۱۸٫۲	۲۴٫۳۹	۲۷	۳۰٫۵	۲۸٫۵	۳۳٫۵	۲۸٫۵	۲۸٫۵	



شکل ۱. فلوچارت روش پیشنهادی برای تعیین داده پرت.

داده‌ی پرت ابتدایی یا انتخاب شده دست یافت و سپس براساس آن‌ها، آماره‌ی T^2 محاسبه می‌شود که این آماره با حدکنترل مخصوص به خود که با توجه به تعداد داده‌ی پرت انتخاب شده مشخص شده، مقایسه می‌شود. روش پیشنهادی این پژوهش را می‌توان در ۳ فاز مطابق فلوچارت شکل ۱ طبقه‌بندی کرد:

۱. طراحی؛

۲. محاسبه‌ی داده‌ی پرت؛

۳. تصمیم‌گیری.

در فاز اول با استفاده از پارامترهای اندازه نمونه (m) و تعداد متغیرها (تعداد شاخص‌های کیفی) (p) ضریب ناهمسانی تعیین می‌شود. سپس با استفاده از پارامتر تعیین شده (ضریب ناهمسانی مجاز) و پارامتر نسبت سیگنال خطا، حدود کنترل متغیر مطابق جدول ۱ تعیین می‌شود.

در فاز دوم محاسبه‌ی داده پرت با توجه به ضریب ناهمسانی تعیین شده در فاز اول به دست می‌آید. محاسبه‌ی داده پرت در هر مرحله از تکرار الگوریتم خوشه‌بندی سلسله‌مراتبی صورت می‌گیرد و داده‌ها به دو خوشه تحت کنترل و خارج از کنترل تقسیم می‌شود (با توجه به ضریب ناهمسانی). خوشه‌یی که دارای تعداد اعضای کم‌تری است به‌عنوان داده‌های پرت اولیه در نظر گرفته می‌شود. سپس داده‌های پرت اولیه بر این اساس مشخص می‌شود و آماره‌ی T^2 برای آن‌ها محاسبه می‌شود. داده‌ی پرت انتخاب شده با توجه به این آماره و حدکنترل متغیر به دست می‌آید. در صورتی که این داده در مجموعه‌ی پرت اولیه باشد، به‌عنوان داده‌ی پرت قطعی شناخته می‌شود. پس از مشخص شدن داده‌های پرت از مجموعه داده‌ها حذف می‌شوند.

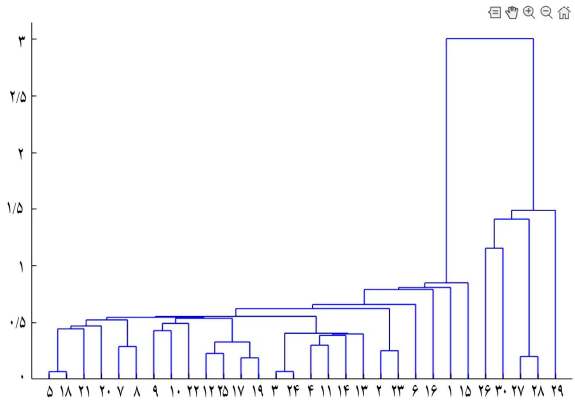
در فاز تصمیم‌گیری، میانگین و ماتریس واریانس کوواریانس T^2 بر اساس داده‌های باقیمانده در فاز دوم محاسبه خواهد شد، سپس مقادیر آماره‌ی T^2 با توجه به حد کنترل به دست آمده در فاز اول برای تمامی داده‌ها محاسبه می‌شود.

۳.۳. مثال مفهومی

برای تشریح بهتر الگوریتم فوق از یک مثال عددی استفاده می‌شود. داده‌های جدول ۳ شامل ۲۵ داده ابتدایی ایجاد شده با توزیع نرمال دومتغیره با میانگین صفر و واریانس همانی ۱ است و ۵ داده‌ی آخر بر اساس ncp برابر ۲۰ (یا میانگین ۳٫۱۶) ایجاد شده است (در حقیقت این ۵ داده پرت محسوب می‌شوند).

جدول ۲. مراحل خوشه‌بندی و آماره‌های محاسبه شده بر روی داده‌ها.

مرحله	داده‌های پرت اولیه	مقدار ضریب ناهمسانی	داده با بیشترین آماره‌ی T^2	مقدار آماره‌ی T^2	داده پرت انتخاب شده	داده پرت قطعی
۱	۳۰ – ۲۹ – ۲۸ – ۲۷ – ۲۶	۱٫۱	۲۶	۹٫۶۴۶۳۴۹۰۱۵	۲۶	۲۶
۲	۲۹ – ۲۸ – ۲۷ – ۲۶	۱٫۱	۲۹	۹٫۰۱۲۶۹۶۵۱۹	۲۹	۲۹(۳۰)
۳	۲۸ – ۲۷ – ۲۶	۱٫۱	۲۸	۸٫۹۷۲۸۱۱۷۶۵	۲۸	۲۸(۲۹)
۴	۲۷ – ۲۶	۱٫۱۲	۲۷	۹٫۰۱۰۰۲۰۹۱۷	۲۷	۲۷(۲۸)
۵	۲۶	۰٫۷۰۷	۲۶	۱۳٫۸۳۱۳۲۴۰۲	۲۶	۲۶(۲۷)
۶	۱۵	۰٫۷۰۷	۱۶	۵٫۴۵۰۰۳۴۷۳۱	-	-
بردار میانگین داده‌ها پس از حذف داده‌های پرت			ماتریس واریانس کوواریانس داده‌ها پس از حذف داده‌های پرت			
۰٫۰۴۱۵	۰٫۰۰۷۵		−۰٫۰۱۸۷	۰٫۶۴		
			۰٫۵۸۷۶	−۰٫۰۱۸۷		



شکل ۲. دندوگرام داده‌ها در اولین مرحله.

با توجه به مقادیر آماره‌ی T^2 در سومین ستون جدول ۳، روش T^2 معمولی با توجه به حدود کنترل شبیه‌سازی شده ۱۰٫۵ برای $m = ۳۰$ و $p = ۲$ هیچ داده‌ی پرتی را در این مجموعه نشان نمی‌دهد.

در نمودار کنترل پیشنهادی، اولین مرحله با خوشه‌بندی سلسله‌مراتبی با نرم‌افزار متلب دندوگرام شکل ۲ به دست می‌آید. در تقسیم داده‌ها به دو خوشه، مطابق شکل ۲ داده‌های ۲۶، ۲۷، ۲۸، ۲۹ و ۳۰ پرت اولیه هستند. به‌علت این که ضریب ناهمسانی خوشه‌ی آخر بیش‌تر از ۱٫۲ است این خوشه‌بندی تأیید می‌شود. سپس آماره‌ی T^2 معمولی برای کل داده‌ها محاسبه شده است (نتایج مربوط به این مرحله، در ستون سوم جدول ۳ گزارش شده است.) مشاهده می‌شود، داده‌ی ۲۶ دارای بیشترین آماره‌ی T^2 است و در مجموعه‌ی داده‌های پرت اولیه قرار دارد و به‌عنوان داده‌ی پرت ثانویه از داده‌ها حذف می‌شود. در جدول ۳ این داده با رنگ زرد مشخص شده است (مقدار آماره ۹٫۶۴۶ است).

در مرحله‌ی دوم برای ۲۹ داده‌ی بعدی دندوگرام رسم می‌شود (به‌منظور اختصار سایر دندوگرام‌ها نشان داده نشده است). در این مرحله داده‌های ۲۶، ۲۷، ۲۸ و ۲۹ در خوشه‌ی داده‌های پرت اولیه قرار دارند (دقت شود که در داده‌های اصلی این داده‌ها ۲۷، ۲۸، ۲۹ و ۳۰ هستند و در جدول ۲ در پرانتز نشان داده شده‌اند). در این مرحله آماره‌ی T^2 برای ۲۹ داده در ستون چهارم جدول ۳ محاسبه شده است و داده‌ی ۲۹ فعلی و ۳۰ اصلی دارای بیشترین مقدار آماره است که در خوشه‌ی داده‌های پرت نیز قرار دارد و از داده‌ها حذف می‌شود (داده‌ی پرت ثانویه). مرحله‌ی سوم و چهارم و پنجم نیز به‌همین شکل انجام می‌گیرد و داده‌های ۲۸ فعلی (۲۹ اصلی) و ۲۷ فعلی (۲۸ اصلی) و ۲۶ فعلی (۲۷ اصلی) از مجموعه‌ی داده‌ها حذف می‌شوند. در مرحله‌ی ششم دندوگرام داده‌ی ۱۵ فعلی را داده‌ی پرت اولیه تشخیص می‌دهد ولی آماره‌ی T^2 برای ۲۵ داده‌ی باقیمانده در داده‌ی ۱۶ دارای بیشترین مقدار است (ستون هشتم جدول ۳). در این قسمت فاز محاسبه‌ی داده‌ی پرت به پایان می‌رسد.

در فاز تصمیم‌گیری باید ابتدا بر مبنای ۲۵ داده‌ی باقیمانده میانگین و ماتریس واریانس ـ کوواریانس را محاسبه کرد. این مقادیر در جدول ۲ گزارش شده است (سطر انتهایی). در مرحله‌ی نهایی بر مبنای این مقادیر میانگین و ماتریس واریانس ـ کوواریانس آماره‌ی T^2 برای کل داده‌ها محاسبه می‌شود که در ستون نهم جدول ۳ گزارش شده است. به‌دلیل این که در این الگوریتم ۵ داده‌ی پرت اولیه شناسایی شد، در این حالت حد کنترل باید (۲۸٫۵ حد کنترل به‌ازای ۵ داده‌ی ثانویه پرت در جدول ۱) در نظر گرفته شود که با توجه به این حدکنترل هر ۵ داده‌ی انتهایی پرت محسوب و با رنگ قرمز مشخص شده است.

این ۵ داده به‌عنوان داده‌ی پرت قطعی در نظر گرفته می‌شود که به‌طور کاملاً صحیح توسط روش شناسایی شده‌اند.

۴. ارزیابی عملکرد

۱.۴. عملکرد روش بر اساس تشخیص یک داده‌ی پرت

ورگاس برای تشخیص داده‌ی پرت از شاخص ncp بر اساس رابطه‌ی ۵ برای ایجاد k داده‌ی پرت استفاده کرد.[۲۶] او بر اساس شبیه‌سازی، سیگنال خطا را در حد ۰٫۰۵ تنظیم کرد و ملاک دقت روش را تشخیص حداقل یک سیگنال در داده‌ها قرار داد. در نمودار کنترل پیشنهادی این پژوهش، به‌ازای هر حالت m و p حدود کنترل (حد کنترل بالا) تعیین می‌شود و سپس در ۵۰۰۰ شبیه‌سازی حالت‌هایی که سیگنال داده شود

جدول ۳. داده‌های مثال تشریحی به همراه مقادیر آماره‌های محاسبه‌شده برای هر داده در هر مرحله.

	X_1	X_2	آماره T^2 معمولی	آماره T^2 در مرحله ۱	آماره T^2 در مرحله ۲	آماره T^2 در مرحله ۳	آماره T^2 در مرحله ۴	آماره T^2 در مرحله ۵	آماره T^2 نهایی
۱	۱.۸۱۲	۰.۱۵۷	۱.۲۲۸	۱.۱۶۸	۱.۱۱۰	۱.۸۵۷	۲.۴۵۷	۵.۱۳۰	۵.۱۳۰
۲	۰.۲۵۴	۱.۱۶۱	۰.۷۰۸	۱.۰۷۸	۱.۴۶۱	۱.۳۷۴	۱.۵۴۴	۲.۲۵۶	۲.۲۵۶
۳	-۰.۸۲۵	-۰.۴۲۲	۰.۸۰۰	۰.۷۶۶	۰.۷۷۹	۰.۸۰۸	۰.۹۲۱	۱.۴۸۶	۱.۴۸۶
۴	-۰.۳۰۵	-۰.۷۴۳	۰.۸۳۱	۰.۹۴۰	۱.۰۷۲	۰.۹۹۸	۱.۰۱۹	۱.۲۲۷	۱.۲۲۷
۵	۰.۸۹۶	-۰.۴۶۴	۱.۰۶۳	۱.۱۴۳	۱.۱۳۳	۱.۳۶۷	۱.۳۵۳	۱.۶۲۳	۱.۶۲۳
۶	-۱.۳۶۳	۰.۹۸۱	۳.۴۲۹	۳.۸۱۶	۳.۹۲۹	۴.۳۹۶	۴.۲۳۷	۴.۳۰۹	۴.۳۰۹
۷	-۰.۰۴۴	-۱.۲۳۳	۱.۸۴۲	۲.۲۸۴	۲.۶۹۳	۲.۶۰۷	۲.۵۹۵	۲.۷۷۸	۲.۷۷۸
۸	۰.۲۲۷	-۱.۱۳۲	۱.۸۲۷	۲.۲۲۸	۲.۵۵۰	۲.۵۳۵	۲.۴۴۶	۲.۳۹۷	۲.۳۹۷
۹	۰.۲۹۵	۰.۴۰۵	۰.۰۵۶	۰.۰۶۱	۰.۰۸۲	۰.۰۵۷	۰.۱۰۵	۰.۳۶۵	۰.۳۶۵
۱۰	۰.۵۱۵	۰.۰۳۹	۰.۱۴۵	۰.۱۱۰	۰.۰۷۴	۰.۱۱۷	۰.۱۵۰	۰.۴۰۳	۰.۴۰۳
۱۱	-۰.۳۰۱	-۰.۴۴۴	۰.۵۱۲	۰.۵۱۷	۰.۵۳۹	۰.۴۸۶	۰.۴۷۸	۰.۵۶۵	۰.۵۶۵
۱۲	-۰.۱۶۰	۰.۶۹۳	۰.۵۶۱	۰.۶۸۸	۰.۷۸۵	۰.۷۴۷	۰.۷۲۰	۰.۷۵۷	۰.۷۵۷
۱۳	-۰.۵۴۵	-۰.۱۲۸	۰.۵۱۳	۰.۴۶۳	۰.۴۲۲	۰.۴۱۸	۰.۴۱۳	۰.۵۳۴	۰.۵۳۴
۱۴	-۰.۱۱۱	-۰.۱۰۶	۰.۲۵۴	۰.۲۱۶	۰.۱۸۰	۰.۱۳۹	۰.۰۹۹	۰.۰۶۱	۰.۰۶۱
۱۵	-۱.۶۴۴	-۰.۱۹۲	۲.۰۵۹	۱.۹۶۰	۱.۸۸۲	۲.۳۳۶	۲.۶۸۷	۴.۳۸۹	۴.۳۸۹
۱۶	۰.۹۵۳	۱.۵۵۷	۰.۷۳۲	۱.۳۲۷	۲.۰۹۶	۲.۰۴۴	۲.۷۷۴	۵.۴۵۰	۵.۴۵۰
۱۷	-۰.۷۵۴	۰.۷۳۰	۱.۴۸۸	۱.۶۶۹	۱.۷۳۶	۱.۸۵۰	۱.۷۴۴	۱.۶۶۱	۱.۶۶۱
۱۸	۰.۸۲۹	-۰.۴۶۲	۰.۹۸۶	۱.۰۶۲	۱.۰۵۶	۱.۲۵۶	۱.۲۳۲	۱.۴۴۵	۱.۴۴۵
۱۹	-۰.۶۸۰	۰.۵۵۷	۱.۱۰۶	۱.۱۹۸	۱.۲۱۳	۱.۲۸۹	۱.۲۰۸	۱.۱۵۵	۱.۱۵۵
۲۰	۰.۷۰۳	-۰.۹۱۵	۱.۸۲۹	۲.۱۳۱	۲.۲۹۴	۲.۴۶۳	۲.۳۳۶	۲.۲۵۰	۲.۲۵۰
۲۱	۰.۳۹۵	-۰.۵۶۰	۰.۷۸۷	۰.۸۷۶	۰.۹۱۵	۰.۹۵۳	۰.۸۸۲	۰.۸۲۸	۰.۸۲۸
۲۲	۱.۰۰۷	۰.۰۷۱	۰.۳۷۰	۰.۳۳۱	۰.۲۹۰	۰.۵۱۷	۰.۶۹۵	۱.۵۶۳	۱.۵۶۳
۲۳	۰.۱۷۶	۱.۴۰۰	۱.۲۲۲	۱.۷۹۴	۲.۳۵۶	۲.۲۴۷	۲.۴۲۷	۳.۲۱۲	۳.۲۱۲
۲۴	-۰.۷۹۰	-۰.۴۸۰	۰.۷۹۲	۰.۷۷۲	۰.۸۰۴	۰.۸۱۶	۰.۹۳۱	۱.۴۹۸	۱.۴۹۸
۲۵	-۰.۳۵۱	۰.۵۷۰	۰.۶۵۷	۰.۷۴۰	۰.۷۸۱	۰.۷۷۳	۰.۷۱۳	۰.۶۵۷	۰.۶۵۷
۲۶	۳.۰۶۳	۴.۷۰۵	۹.۶۴۶	-	-	-	-	-	۵۳.۰۴۷
۲۷	۳.۹۹۶	۲.۲۲۴	۳.۸۲۰	۴.۳۸۳	۶.۱۰۴	۸.۵۴۰	۱۳.۸۳۱	-	۳۳.۸۲۲
۲۸	۴.۰۰۳	۲.۴۲۴	۳.۸۷۱	۴.۶۴۷	۶.۶۹۳	۹.۰۱۰	-	-	۳۵.۵۴۸
۲۹	۵.۱۱۱	۱.۲۳۵	۸.۹۴۱	۸.۶۱۹	۸.۹۷۳	-	-	-	۴۳.۷۰۸
۳۰	۳.۸۱۳	۳.۸۲۴	۵.۹۲۶	۹.۰۱۳	-	-	-	-	۴۸.۴۱۸

به‌عنوان حالت صحیح شناخته می‌شود و نسبت تعداد حالت‌هایی که روش داده پرت تشخیص می‌دهد به تعداد کل شبیه‌سازی ثبت می‌شود. در شکل ۳ تا ۵ مقایسه‌ی ارزیابی عملکرد نمودار T^2 خوشه‌بندی در شناسایی نقاط پرت با MVE، MCD، SW۱، SW۲ و T^2 معمولی بر اساس $m = ۳۰$ و $p = ۲$ ارائه شده است. در شکل‌های ۳ تا ۵ ارزیابی عملکرد به ترتیب با ۳، ۵ و ۷ داده‌ی پرت آورده شده است.

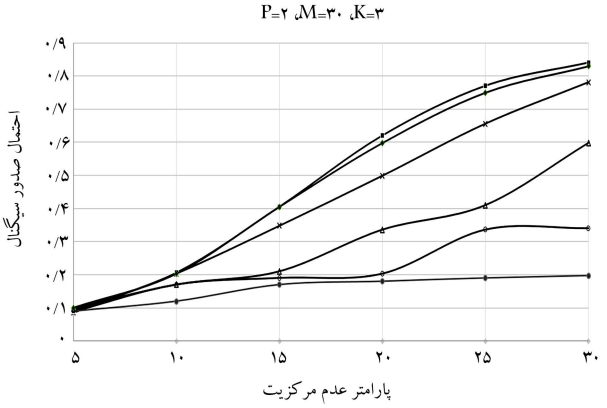
مقایسه‌ی عملکرد نمودار T^2 خوشه‌بندی با نمودار T^2 معمولی در جداول ۴ تا ۷ آمده است. در جدول ۴ مقایسه‌ی عملکرد نمودار T^2 با نمودار کنترل T^2 خوشه‌بندی از نظر درصد تشخیص داده‌ی پرت به‌ازای $P = ۳$ آورده شده است. هر دو نمودار در شاخص عدم مرکزیت برابر با ۵ ($\text{ncp} = ۵$) در تعداد نقاط پرت مختلف و با تعداد داده‌ها متفاوت، عملکرد تقریباً مشابه دارند اما با افزایش شاخص عدم مرکزیت عملکرد نمودار T^2 خوشه‌بندی بهتر می‌شود.

در جدول ۵ نیز درصد تشخیص دو نمودار با ۵ متغیر آورده شده است. در این جدول نیز با افزایش مقدار ncp عملکرد نمودار خوشه‌بندی به مراتب بهتر بوده است.

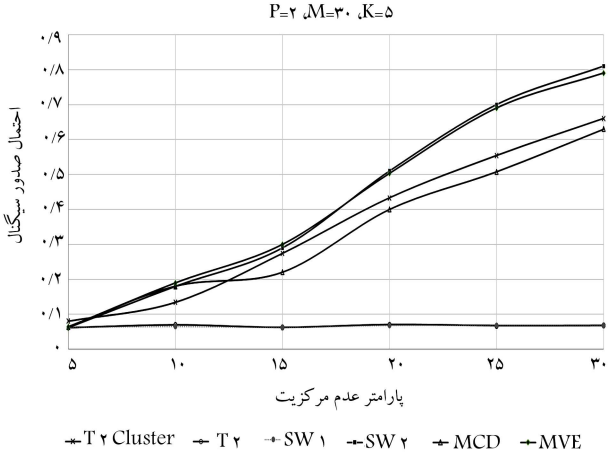
در جدول ۶ درصد تشخیص داده پرت، به‌ازای $p = ۱۰$ نشان داده شده است. عملکرد هر دو نمودار مانند جداول ۴ و ۵ در ($\text{ncp} = ۵$) تقریبا مشابه است و با افزایش مقدار شاخص عدم مرکزیت عملکرد نمودار خوشه‌بندی افزایش یافته‌است. نکته قابل مشاهده دیگر اینست که به طور کلی با افزایش تعدادمتغیرها عملکرد هر دو نمودار کاهش می‌یابد.

جدول ۴. درصد تشخیص داده پرت به‌ازای $p = ۳$ در نمودار T^2 و نمودار T^2 خوشه‌بندی به‌ازای مقادیر مختلف شاخص عدم مرکزیت.

ncp	$p = ۳$									Methods
	$m = ۳۰$			$m = ۵۰$			$m = ۱۰۰$			
	$K = ۲$	۴	۶	۲	۵	۱۰	۵	۱۰	۲۰	
۵	۰٫۰۷۶	۰٫۰۷۶	۰٫۰۵	۰٫۰۹۳	۰٫۰۸۲	۰٫۰۵۹	۰٫۱۲۱	۰٫۰۸۴	۰٫۰۴۵	T^2
	۰٫۰۸۹	۰٫۰۷۶	۰٫۰۵۹	۰٫۰۹۲	۰٫۰۹	۰٫۰۶۳	۰٫۱۳۳	۰٫۱۰۷	۰٫۰۶۴	$T^2 Cluster$
۱۵	۰٫۲۶۳	۰٫۰۷۷	۰٫۰۴۴	۰٫۴۴۷	۰٫۱۳۷	۰٫۰۵۷	۰٫۴۴۱	۰٫۱۳۷	۰٫۰۳۹	T^2
	۰٫۴۲	۰٫۲۵۷	۰٫۱۱۸	۰٫۴۰۶	۰٫۳۵۶	۰٫۱۰۴	۰٫۵۸۶	۰٫۲۸۴	۰٫۰۵۸۸	$T^2 Cluster$
۲۵	۰٫۴۴۹	۰٫۰۸	۰٫۰۵۱	۰٫۷۶۱	۰٫۱۸۲	۰٫۰۳۸	۰٫۶۸۸	۰٫۱۷۱	۰٫۰۴۵	T^2
	۰٫۷۴۱	۰٫۴۹۷	۰٫۲۴۶	۰٫۷۷۲	۰٫۷۰۱	۰٫۲۱۲	۰٫۸۹۶	۰٫۵۶۶	۰٫۰۹۸	$T^2 Cluster$



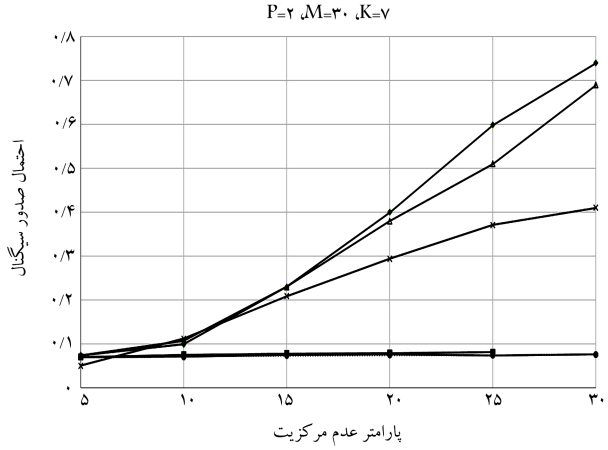
شکل ۳. احتمال شناسایی سیگنال با سه داده پرت.



شکل ۴. احتمال شناسایی سیگنال با پنج داده پرت.

۲.۴. عملکرد روش بر اساس تشخیص مجموعه‌ی داده‌های پرت

در این قسمت بر اساس روش آلفارو و همکاران عمل می‌شود.[۲] در این بخش داده‌های تحت کنترل با میانگین ۰ و ماتریس واریانس کوواریانس ۱ بر روی قطر



شکل ۵. احتمال شناسایی سیگنال با هفت داده پرت.

اصلی و ۰٫۹ بر روی سایر قسمت‌ها ایجاد می‌شوند. داده‌های پرت با میانگین ۵ ایجاد می‌شوند. حدود سیگنال خطا در دو روش بر روی حد مشخصی با شبیه‌سازی تنظیم می‌شوند. در این روش اگر تمامی داده‌های پرت شناسایی شوند، دقت روش مورد تأیید است. در جدول ۷ دو عدد برای ستون‌های T^2 و T^2 خوشه‌بندی ذکر شده است. عدد اول نسبت تعداد دفعاتی است که روش توانسته است کلیه‌ی داده‌های پرت را شناسایی نماید و عدد دوم نرخ سیگنال خطا در آن حالت است. سیگنال خطا هنگامی حساب می‌شود که روش در داده‌های غیرپرت (تفاوتی ندارد که در یک داده یا چند داده از یک مجموعه‌ی داده سیگنال صادر شود) سیگنال دهد.

۳.۴. عملکرد روش بر اساس مجموعه داده فسفر

در این مجموعه داده $m = ۱۸$ و $p = ۲$ است.[۳۳و۳۴] مطابق جدول ۱ ابتدا باید حدود کنترل برای سیگنال ۰٫۰۵ خطا تعیین شود.

در این مجموعه داده حدود کنترل در سطح خطای ۰٫۰۵ به‌صورت [۱۰٫۰۷, ۴۷٫۲, ۵۷٫۳۹, ۴۹, ۴۴٫۵, ۵۳٫۵, ۳۳٫۵, ۲۰٫۵] است. در اجرای الگوریتم، مجموعه داده‌های پرت ثانویه به‌ترتیب ۱۷، ۶، ۱، ۸ و ۴ است که باید حد

جدول ۵. درصد تشخیص داده پرت به‌ازای $p = 5$ در نمودار T^2 و نمودار T^2 خوشه‌بندی به‌ازای مقادیر مختلف شاخص عدم مرکزیت.

روش	$p = 5$									پارامتر عدم مرکزیت
	$m = 100$			$m = 50$			$m = 30$			
	۲۰	۱۰	۵	۱۰	۵	۲	۶	۴	۲	
T^2	۰/۰۴۵	۰/۰۷۴	۰/۰۹۴	۰/۰۶۲	۰/۰۷۱	۰/۰۷۶	۰/۰۶۲	۰/۰۷۵	۰/۰۶۸	۵
$T^2 Cluster$	۰/۰۶۵	۰/۰۷۹۲	۰/۰۹	۰/۰۴۲	۰/۰۶۰۸	۰/۰۶۳۶	۰/۰۶۹	۰/۰۶۳	۰/۰۸۴۴	
T^2	۰/۰۵	۰/۰۹۳	۰/۲۶۷	۰/۰۵	۰/۱۱	۰/۲۷	۰/۰۴۹	۰/۰۷۶	۰/۱۶۹	۱۵
$T^2 Cluster$	۰/۰۵۹	۰/۱۴۸	۰/۳۵۵	۰/۰۵۳	۰/۱۵۸	۰/۲۲۶	۰/۰۸۷	۰/۱۴۹	۰/۲۳۳	
T^2	۰/۰۶	۰/۱۵۵	۰/۴۵۶	۰/۰۴۶	۰/۱۱۸	۰/۵۳	۰/۰۵۳	۰/۰۸۴	۰/۲۸	۲۵
$T^2 Cluster$	۰/۰۶۷	۰/۲۷۲	۰/۷۲۷	۰/۰۷۶	۰/۳۷۴	۰/۵۰۹	۰/۱۲۴	۰/۳۰۹	۰/۴۹۹	

جدول ۶. درصد تشخیص داده پرت به‌ازای $p = 10$ در نمودار T^2 و نمودار T^2 خوشه‌بندی به‌ازای مقادیر مختلف شاخص عدم مرکزیت.

روش	$p = 10$									پارامتر عدم مرکزیت
	$m = 100$			$m = 50$			$m = 30$			
	۲۰	۱۰	۵	۱۰	۵	۲	۶	۴	۲	
T^2	۰/۰۵۸	۰/۰۷۶	۰/۰۶۸	۰/۰۶۳	۰/۰۵۸	۰/۰۵۶	۰/۰۴۸	۰/۰۴۹	۰/۰۶۱	۵
$T^2 Cluster$	۰/۰۵۴	۰/۰۶۱	۰/۰۶۵	۰/۰۵۹	۰/۰۵۸	۰/۰۶۶	۰/۰۵۲	۰/۰۵۹	۰/۰۵۹	
T^2	۰/۰۵۲	۰/۰۷	۰/۱۳۸	۰/۰۵۳	۰/۰۷۱	۰/۱۳۳	۰/۰۵۵	۰/۰۵۶	۰/۰۶۲	۱۵
$T^2 Cluster$	۰/۰۶	۰/۰۹۵	۰/۱۵۹	۰/۰۵۹	۰/۰۹۸	۰/۱۷۳	۰/۰۵۴	۰/۰۶۹	۰/۱۲۹	
T^2	۰/۰۴۸	۰/۰۷۲	۰/۲۲۹	۰/۰۵۹	۰/۰۸۲	۰/۲۷۵	۰/۰۵۷	۰/۰۵۷	۰/۱	۲۵
$T^2 Cluster$	۰/۰۵۹	۰/۱۱۷	۰/۳۵۷	۰/۰۶۱	۰/۱۴۶	۰/۴۱	۰/۰۶۴	۰/۰۸۹	۰/۲۶	

جدول ۷. درصد تشخیص صحیح تمامی داده‌های پرت و سیگنال خطا در $m = 25$ برای دو نمودار T^2 و T^2 خوشه‌بندی.

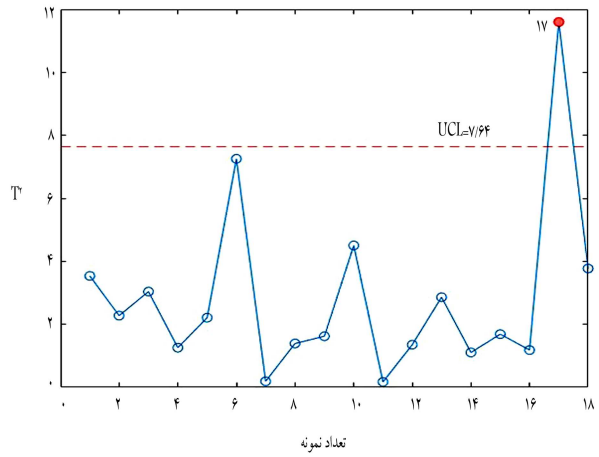
p	k	T^2		$T^2 Cluster$	
		درصد تشخیص	سیگنال خطا	درصد تشخیص	سیگنال خطا
۲	۰	-	۱/۱	-	۱/۵۶
۲	۱	۶۳/۱۲	۰/۳۴	۵۱/۹۴	۱/۷۶
۲	۲	۰/۱۸	۰/۲۴	۳۱/۶۶	۱/۵۶
۲	۳	۰	۰/۰۲	۲۳/۱۶	۱/۶۴
۳	۰	-	۰/۲۲	-	۰/۳
۳	۱	۳۱/۸	۰/۱	۲۰/۸۲	۰/۳۲
۳	۲	۰/۰۴	۰/۰۸	۷	۰/۳۸
۳	۳	۰	۰/۲	۳/۹۸	۰/۶

کنترل آن را ۵۳/۵ در نظرگرفت. نمودار آماره‌های T^2 معمولی و خوشه‌بندی مطابق شکل ۶ و ۷ است.

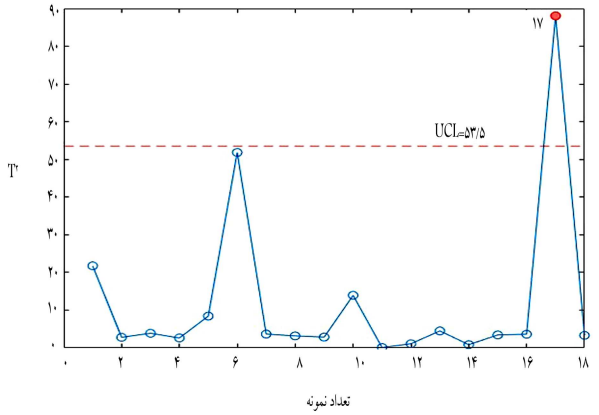
در این مجموعه داده یک داده‌ی پرت وجود دارد. نمودار T^2 معمولی در صورتی که مجموعه داده فقط دارای یک داده پرت باشد، عالی عمل می‌کند.[۱۰] با توجه به شکل ۶ و ۷ نیز مشاهده می‌شود عملکرد هر نمودار مشابه هم است و هر دو نمودار داده‌ی پرت ۱۷ را شناسایی کرده‌اند.

۴.۴. عملکرد روش بر اساس مجموعه داده هاوکینز

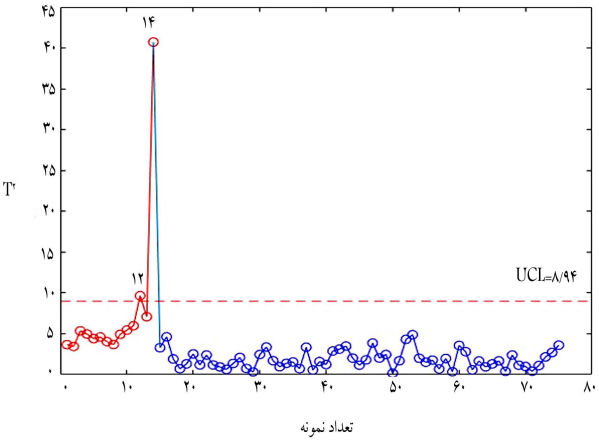
این مجموعه داده شامل ۷۵ مشاهده با چهار متغیر است،[۳۵] در این مجموعه داده ۱۴ داده‌ی ابتدایی پرت هستند. در شکل ۸ مشاهده می‌شود نمودار T^2 معمولی از ۱۴ داده پرت داده‌ی ۱۲ و ۱۴ را شناسایی کرده است در صورتی که نمودار T^2 خوشه‌بندی در شکل ۹ داده‌های ۱، ۲ تا ۱۴ را پرت ثانویه تشخیص می‌دهد و



شکل ۶. نمودار T^2 معمولی در مجموعه داده فسفر.



شکل ۷. نمودار T^2 خوشه‌بندی در مجموعه داده فسفر.

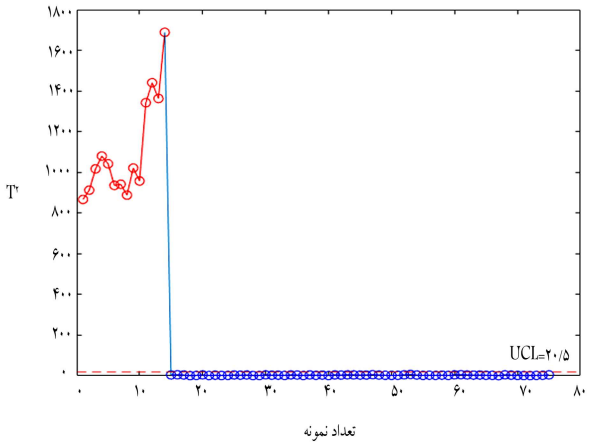


شکل ۸. نمودار T^2 معمولی در مجموعه داده‌ی هاوکینز.

به‌ازای ۱۴ مقدار داده در مجموعه داده‌های پرت ثانویه این روش مقدار UCL را ۲۰/۵ گزارش می‌دهد.

۵. بحث

با توجه به حساسیت پایین نمودارکنترل T^2، برای بهبود آن، نمودارکنترل پیشنهادی برمبنای روش خوشه‌بندی سلسه‌مراتبی ارائه شده است که تحت تأثیر داده‌های



شکل ۹. نمودار T^2 خوشه‌بندی در مجموعه داده‌ی هاوکینز.

انحرافی در نمونه یا داده‌های نامتعارف قرار نمی‌گیرد و نقاط پرت موجود در فاز اول نمودارهای کنترلی چندمتغیره را شناسایی و حذف می‌کند. روش پیشنهادی برای بهبود عملکرد نمودار کنترل T^2 در مثال مفهومی تشریح شد. برای تأیید عملکرد روش پیشنهادی نسبت به سایر روش‌های موجود در ادبیات که برای بهبود نمودارکنترل T^2 معرفی شده است، شبیه‌سازی و مقایسه انجام شد. بنابراین ابتدا عملکرد نمودارکنترل پیشنهادی در شناسایی نقاط پرت با روش حجم بیضی‌وار (MVE) و حداقل دترمینان کوواریانس (MCD) رسئو،[۱۹] برآوردگرهای SW۱، SW۲ سالیوان و وودال[۱۷] و نمودار T^2 معمولی بر اساس $m = ۳۰$ و $p = ۲$ مقایسه شده است. در حالتی که سه نقطه‌ی پرت وجود داشته باشد، مطابق شکل ۳ مشاهده می‌شود، نمودارکنترل پیشنهادی (T^2 خوشه‌بندی) از نمودار T^2 معمولی، SW۱ سالیوان و وودال[۱۷] و MCD رسئو[۱۹] به مراتب عملکرد بهتری دارد و عملکرد تقریباً مشابهی با SW۲ سالیوان و وودال[۱۷] و MVE رسئو[۱۹] دارد. با این حال این روش‌ها در مقایسه با روش خوشه‌بندی، به لحاظ عملی بسیار وقت‌گیر است.

چنان‌که گفته شد هدف پژوهش حاضر استفاده از خوشه‌بندی سلسه‌مراتبی برای بهبود نمودارکنترل T^2 معمولی است، لذا نمودارکنترل پیشنهادی (T^2 خوشه‌بندی) با نمودار T^2 معمولی در حالت‌های مختلف شبیه‌سازی و مقایسه شد. در جدول ۴ نشان می‌دهد به ازای $\text{ncp} = ۵$ با دو نقطه‌ی پرت، درصد تشخیص داده‌ی پرت در نمودارکنترل T^2 خوشه‌بندی نزدیک به نمودار T^2 معمولی است که با افزایش اندازه‌ی (m) این اختلاف کم‌تر نیز شده است، اما با افزایش نقاط پرت درصد تشخیص داده‌ی پرت در نمودار T^2 معمولی به‌شدت کاهش داشته است. همچنین مشاهده می‌شود با افزایش شاخص عدم مرکزیت (ncp) قدرت نمودارکنترل T^2 خوشه‌بندی نسبت به نمودار T^2 معمولی به‌مراتب بهتر است. در جدول ۵ تعداد متغیرها از ۳ به ۵ افزایش داده شد؛ مشاهده می‌شود با افزایش تعداد متغیرها عملکرد هر دو نمودار افت دارد اما در نمودارکنترل T^2 خوشه‌بندی کاهش شدیدی حاصل نشده است.

برای تحلیل بیشتر، عملکرد نمودارکنترل T^2 خوشه‌بندی با عملکرد نمودار T^2 معمولی بر اساس تشخیص مجموعه داده‌های پرت، شبیه‌سازی و مقایسه شد. همان‌طورکه در جدول ۷ مشاهده می‌شود با افزایش تعداد نقاط پرت عملکرد نمودار T^2 خوشه‌بندی در شناسایی صحیح تمامی نقاط پرت به مراتب خیلی بهتر از نمودار T^2 معمولی است. اما نرخ سیگنال خطا در نمودار T^2 خوشه‌بندی بالاتر بوده که با

افزایش تعدادمتغیرها (P) سیگنال خطا کاهش یافته است.

برای تحلیل عملکرد نمودار T^2 خوشه‌بندی، مقایسه‌ی عملکرد آن با نمودار T^2 معمولی روی داده‌های واقعی فسفر و هاوکینز نیز انجام شد. در مجموعه داده‌ی فسفر یک داده پرت وجود دارد و عملکرد نمودار T^2 معمولی مشابه نمودار کنترل T^2 خوشه‌بندی است، در حالی که در مجموعه داده‌ی هاوکینز که ۱۴ داده‌های پرت وجود دارد نمودار T^2 معمولی فقط سه داده‌ی پرت را شناسایی کرده است اما نمودار کنترل T^2 خوشه‌بندی هر ۱۴ داده پرت شناسایی کرده است، بنابراین با توجه به مقایسات با نمودارهای دیگر در بخش اول ارزیابی و مقایسات مختلف نمودار با نمودار T^2 معمولی می‌توان نتیجه گرفت که عملکرد نمودار کنترل T^2 خوشه‌بندی کاراتر از نمودار T^2 معمولی است.

۶. نتیجه‌گیری

در بسیاری از شرایط عملی، اغلب لازم است به طور همزمان ویژگی‌های چند کیفیت کنترل شود. شناسایی پرت‌های بالقوه یا مشاهدات مؤثر در مجموعه داده‌های چندگانه، قبل از شروع به تحلیل‌های پیشرفته‌تر آماری، یک وظیفه‌ی بسیار مهم است. نمودارکنترل چندمتغیره رایج‌ترین ابزار آماری مورد استفاده در چنین شرایط است، اما نمودار کنترل چندمتغیره‌ی کلاسیک نمی‌تواند به طور مؤثری داده‌های پرت را تشخیص دهد. تاکنون روش‌های متعددی برای بهبود عملکرد نمودارهای کنترل چندمتغیره ارائه شده است. از جمله می‌توان به برآوردگر باثبات حداقل حجم بیضی‌وار، حداقل دترمینان واریانس، برآوردگر سالیوان ـ وودال و حداقل دترمینان واریانس بازموزون اشاره کرد. همه‌ی این روش‌ها به دنبال به دست آوردن برآوردی از میانگین و ماتریس کوواریانس مجموعه داده‌های چندمتغیره‌اند که نسبت به نقاط غیرمعمول و دورافتاده، مقاوم بوده و تحت تأثیر این نقاط قرار نگیرند. در حالی که این روش‌ها یعنی MVE و MCD طبق اظهارنظر فان و همکارانش (۲۰۱۲) نیازمند تلاش‌های محاسباتی فوق‌العاده زیادی هستند. به همین علت، مطلوب است که نمودارهای کنترل قوی برای موقعیت‌های چندمتغیره ایجاد شود که همچنین تعادل خوبی بین قابلیت تشخیص و تلاش محاسبات را حفظ کند. لذا در این تحقیق نیز برای این منظور از تکنیک خوشه‌بندی سلسه‌مراتبی در طراحی نمودار کنترل T^2 هتلینگ در فاز اول استفاده شد. به عبارتی در این تحقیق با استفاده از رویکردهای خوشه‌یی توان نمودار را در تشخیص داده‌های پرت افزایش داده شد. عملکرد روش پیشنهادی به دو صورت مورد ارزیابی قرار گرفت. با استفاده از شاخص عدم مرکزیت نمودار کنترل T^2 معمولی و T^2 خوشه‌بندی در ابعاد ۳، ۵ و ۱۰ و با تعداد داده‌های ۳۰، ۵۰ و ۱۰۰ مقایسه شد. نتایج به طورکلی نشان داد که با افزایش شاخص عدم مرکزیت عملکرد نمودارکنترل T^2 خوشه‌بندی افزایش یافته است. همچنین نتایج ارزیابی نیز با استفاده از روش الفارو همکاران نشان داد که نمودارکنترل T^2 خوشه‌بندی در شناسایی تعییرات بسیارکاراتر بوده اما نرخ خطا بالاست که نرخ خطا هم با افزایش تعداد متغیرها کاهش یافته است. مطابق جدول ۷ نمودارکنترل T^2 خوشه‌بندی در ۲۵ داده‌ی ۳ متغیره با ۳ داده پرت، نسبت تعداد دفعاتی است که روش توانسته است کلیه‌ی داده‌های پرت با نرخ خطای ۰٫۶ را شناسایی کند معادل ۳٫۹۸ بوده است در حالی که نمودار T^2 معمولی قادر به شناسایی نقاط پرت نبوده است. در مرحله‌ی دوم با استفاده از یک مجموعه داده‌ی فسفر و هاوکینز نمودارکنترل T^2 خوشه‌بندی طراحی و با نتایج به دست آمده از نمودارکنترل T^2 معمولی مقایسه شد. نتایج اجرای نمودارهای کنترل نشان می‌دهد که نمودارکنترل T^2 خوشه‌بندی در شناسایی نقاط پرت کاراتر است.

۷. ملاحظات

این مقاله مستخرج از طرح شماره ۴۹۴۰۹ مورخه ۹۷٫۱۲٫۲۷ با حمایت معاونت پژوهشی دانشگاه فردوسی مشهد است.

پانوشت‌ها

1. masking
2. swamping
3. trimming
4. stalactite
5. minimum volume ellipsoid (MVE)
6. minimum covariance determinant
7. Subsampling
8. reweighted minimum covariance determinant

منابع (References)

1. Fazel Zarandi, M.H., Alaeddini, A. and Turksen, I.B. "A hybrid fuzzy adaptive sampling-run rules for Shewhart control charts", *Information Sciences*, **178**, pp. 1152-1170 (2008).
2. De Vries, A. and Conlin, B.J. "A comparison of the performance of statistical quality control charts in a dairy production system through stochastic simulation", *Agricultural Systems*, **85**, pp. 317-341 (2005).
3. Shewhart. W.A., *Statistical Methods From the Viewpoint of Quality Control*, Republished in 1986 by Dover Publications, New York, NY (1939).
4. Runger, G.C. "Multivariate statistical process control for autocorrelated processes", *International Journal of Production Research*, **34**, pp. 1715-1724 (1996).
5. Mason, R.L., Champ, C.W., Tracy, N.D. and et al. "Assessment of multivariate process control techniques", *Journal of Quality Technology*, **29**(2), pp. 140-143 (1997a).
6. Mason, R.L., Tracy, N.D. and Young, J.C. "A practical approach for interpreting multivariate T2 control chart signals", *Journal of Quality Technology*, **29**(4), pp. 396-406 (1997b).

7. Williams, J.D., Woodall, W.H., Birch, J.B. and et al. "Distribution of hotelling's T2 statistic based on the successive differences estimator", *Journal of Quality Technology*, **38**, pp. 217-229 (2006).
8. Fan, S.k., Huang, H.-K. and Chang, Y.-J. "Robust multivariate control chart for outlier detection using hierarchical cluster tree in SW2", *Quality and Reliability Engineering* 29(7), (2012). DOI: 10.1002/qre.1448.
9. Mason, R.L. and Young, J.C., *Multivariate Statistical Process Control With Industrial Applications American Statistical Association and the Society for Industrial and Applied Mathematics*, Philadelphia, PA (2002).
10. Ong, H.C. and Alih, E. "A control chart based on cluster-regression adjustment for retrospective monitoring of individual characteristics", *PLoS ONE*, **10**(4), e0125835 (2015). doi:10.1371/journal.pone.0125835.
11. Alfaro, J.L. and Ortega, J.F. "A comparison of robust alternatives to hotelling's T 2 control chart", *Journal of Applied Statistics*, **36**(12), pp. 1385-1396 (2009).
12. Chenouri, S.E., Steiner, S.H. and Variyath, A.M. "A multivariate robust control chart for individual observations", *Journal of Quality Technology*, **41**(3), pp. 259-271 (2009).
13. Rousseeuw, P.J. and Van Zomeren, B.C. "Unmasking multivariate outliers and leverage points", *Journal of the American Statistical Association*, **85**(411), pp. 633-639 (1990).
14. Ahsan, M., Mashuri, M., Kuswanto, H. and et al. "Outlier detection using PCA mix based T 2 control chart for continuous and categorical data", *Communications in Statistics-Simulation and Computation*, pp. 1-28 (2019).
15. Rocke, D.M. "And RQ charts: robust control charts", *Journal of the Royal Statistical Society: Series D (The Statistician)*, **41**(1), pp. 97-104 (1992).
16. Alfaro, J.L. and Ortega, J.F. "A robust alternative to Hotelling's T2 control chart using trimmed estimators", *Quality and Reliability Engineering International*, **24**(5), pp. 601-611 (2008).
17. Sullivan, J.H. and Woodall, W.H. "A Comparison of multivariate control charts for individual observations", *Journal of Quality Technology*, **28**(4), pp. 398-408 (1996).
18. Atkinson, A.C. and Mulira, H.M. "The stalactite plot for the detection of multivariate outliers", *Statistics and Computing*, **3**(1), pp. 27-35 (1993).
19. Rousseeuw, P.J. "Least median of squares regression", *Journal of the American Statistical Association*, **79**(388), pp. 871-880 (1984).
20. Kang, JI.H. and Bum Kim, S. "A clustering algorithm-based control chart for inhomogeneously distributed TFT-LCD processes", *International Journal of Production Research*, **51**(18). pp. 5645-5657 (2013).
21. Hajabbasi. "Multivariate quality control charts", Unpublished master's Thesis. Shahid Bahonar University of Kerman. Kerman (2013).
22. Vargas, J.A.N. "Robust estimation in multivariate control charts for individual observations", *Journal of Quality Technology*, **35**(4), pp. 367-376 (2003).
23. Rousseeuw, P.J. and Leroy, A.M., *Robust Regression and Outlier Detection*, **589**, John wiley & sons (2005).
24. Rousseeuw, P.J. and Driessen, K.V. "A fast algorithm for the minimum covariance determinant estimator", *Technometrics*, **41**(3), pp. 212-223 (1999).
25. Jensen, W.A., Birch, J.B. and Woodall, W.H. "High breakdown estimation methods for phase I multivariate control charts", *Quality and Reliability Engineering International*, **23**(5), pp. 615-629 (2007).
26. Willems, G., Pison, G., Rousseeuw, P.J. and et al. "A robust Hotelling test", *Metrika*, **55**(1-2), pp. 125-138 (2002).
27. Variyath, A.M. and Vattathoor, J. "Robust control charts for monitoring process mean of phase-I multivariate individual observations", *Journal of Quality and Reliability Engineering Volume 2013, Article ID 542305*, 14 pages (2013). doi.org/10.1155/2013/542305.
28. Ankara, H. and Yerel, S. "Determination of sampling errors in natural stone plates through single linkage cluster method", *Journal of Materials Processing Technology*, pp. 2483-2487 (2008). doi:10.1016/j.jmatprotec.2008.05.048.
29. Jobe, J.M. and Pokojovy, M. "A multistep, cluster-based multivariate chart for retrospective monitoring of individuals", *Journal of Quality Technology*, **41**(4), pp. 323-339 (2009).
30. Jobe, J.M. and Pokojovy, M. "A cluster-based outlier detection scheme for multivariate data", *Journal of the American Statistical Association*, **110**(512), pp. 1543-1551 (2015).
31. Huang, J., Zhu, Q., Yang, L. and et al. "A novel outlier cluster detection algorithm without top-n parameter", *Knowledge-Based Systems*, **121**, pp. 32-40 (2017).
32. https://www.mathworks.com/help/stats/hierarchical-clustering.html.
33. https://rdrr.io/cran/robustbase/man/phosphor.html.
34. Snedecor, G.W. and Cochran, W.G., *Statistical Methods*, 6th Edition. The Iowa State University Press: Ames, Iowa (1967).
35. Hawkins, D.M., Bradu, D. and Kass, G.V. "Location of several outliers in multiple-regression data using elemental sets", *Technometrics*, **26**, pp. 197-208 (1984).