مهندسی صنایع و مدیریت شریف، (زمستان ۱۴۰۰)
دوره‌ی ۱-۳۷، شماره‌ی ۲، ص. ۷۹-۹۰، (پژوهشی)

Original Article

مقایسه‌ی عملکرد روش‌های اندازه‌گیری شباهت طولانی‌ترین زیردنباله‌ی مشترک و چرخش زمانی پویا در داده‌کاوی سری‌های زمانی

غلامرضا سلیمانی (دانشجوی دکتری)
مسعود عابسی* (استادیار)
گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه یزد

تکنیک‌های داده‌کاوی به‌طور خاص برای داده‌های ثابت طراحی شده‌اند. لذا به‌کارگیری آنها برای داده‌های سری زمانی نیازمند اعمال تغییراتی (روش اندازه‌گیری شباهت) است.
بر اساس تحقیقات اخیر، روش‌های طولانی‌ترین زیردنباله‌ی مشترک و چرخش زمانی پویا، از پرکاربردترین و کارآترین این روش‌ها محسوب می‌شود. در این تحقیق، قصد داریم تا عملکرد این روش‌ها را در تکنیک‌های نزدیک‌ترین همسایگی و خوشه‌بندی کامدوید مورد ارزیابی و مقایسه قرار داده تا بتوان از آنها با دقت بهتری در این تکنیک‌ها و در مسائلی نظیر قسمت‌بندی مشتریان، زمان‌بندی کارگاه و ... استفاده کرد. به همین منظور از ۶۳ مجموعه داده سری زمانی از بانک اطلاعاتی UCR، استفاده می‌شود. نتایج نشان می‌دهد که تأثیر آنها در دقت تشخیص درست دسته‌ی سری زمانی و دقت خوشه‌بندی، به‌طور معناداری تفاوت دارد، ولی تأثیر آنها در تعیین تعداد خوشه و نماینده‌ی خوشه، تفاوت معناداری ندارد.

واژگان کلیدی: داده‌کاوی سری‌های زمانی، خوشه‌بندی، نزدیک‌ترین همسایگی، طولانی‌ترین زیردنباله‌ی مشترک، چرخش زمانی پویا.

gholam_soleimani@yahoo.com
mabessi@gmail.com

۱. مقدمه

داده‌های سری زمانی، دنباله‌یی از اعداد هستند که در فواصل زمانیِ غالباً یکسان اندازه‌گیری شده است.[۱] سری‌های زمانی تقریباً در همه حیطه‌ها وجود دارد (برای مثال: در حیطه‌ی پزشکی به‌صورت نوار ضربان قلب، نوار تنفس و نوار مغز برای یک دوره‌ی زمانی؛ در حیطه‌ی آب و هوا نظیر دمای یک مکان یا میزان رطوبت یک مکان در فواصل زمانی مشخص؛ در حیطه‌ی فروش نظیر ارزش فروش در فواصل زمانی مشخص و ...). سری‌های زمانی دارای دو ویژگی برجسته‌اند: ۱. طول سری زمانی؛ یک سری زمانی می‌تواند طولی برابر با صدها بلکه هزارها داشته باشد. این ویژگی در میزان اشغال فضای حافظه و سخت شدن محاسبات تأثیر به‌سزایی دارد. ۲. وابستگی بین اعداد؛ مقادیر یک سری زمانی به هم وابسته‌اند و همین موضوع، ضرورت توجه به انتخاب ابزارهای مناسب برای انجام محاسبات و تحلیل آنها را مشکل‌تر می‌کند.[۲-۵]

برای کشف دانش از انبوهی از داده، «داده‌کاوی» ابزاری مفید محسوب می‌شود. تکنیک‌های مختلفی برای داده‌کاوی وجود دارد، نظیر دسته‌بندی، خوشه‌بندی، کشف الگوهای جذاب و تکراری و در نهایت کشف دورافتاده‌ها. جالب توجه است که همه این تکنیک‌ها برای داده‌های ثابت طراحی شده‌اند[۶] و به‌کارگیری آنها برای داده‌های سری زمانی، منجر به ایجاد نتایج معتبر نخواهد شد. بنابراین به منظور قابل استفاده کردن این تکنیک‌ها برای داده‌های سری زمانی بایستی تغییراتی در الگوریتم‌های داده‌کاوی اعمال کرد. یکی از این تغییرات، انتخاب و استفاده از یک روش اندازه‌گیری شباهت مناسب برای داده‌های سری زمانی است.[۷] از سوی دیگر، روش‌های بسیاری برای این منظور ارائه شده که هرکدام از یک سطح کیفی برخوردارند. از میان این روش‌ها، روش طولانی‌ترین زیردنباله‌ی مشترک[۱] و چرخش زمانی پویا[۲] پرکاربردترین و کارآترین روش‌های اندازه‌گیری شباهت سری‌های زمانی هستند.[۸]

در این تحقیق قصد داریم به ارزیابی عملکرد این دو روش در مورد برخی از تکنیک‌های داده‌کاوی بپردازیم تا مشخص شود که تأثیر آنها در کدام تکنیک بهتر است؟ به منظور این ارزیابی از تکنیک‌های نزدیک‌ترین همسایگی و خوشه‌بندی کامدوید استفاده می‌کنیم.

تکنیک نزدیک‌ترین همسایگی یکی از تکنیک‌های دسته‌بندی است که کاربرد گسترده‌یی دارد. این تکنیک در حوزه‌های مختلف نظیر پزشکی ـ در تعیین وضعیت بیماری مراجعین در مقایسه با وضعیت سلامتی بیماران قبلی ـ کاربرد دارد. در حوزه‌ی اقتصادی نیز از این تکنیک برای تعیین وضعیت ارزش سهام یک شرکت در مقایسه با وضعیت ارزش سهام شرکت‌های موجود استفاده می‌شود. به‌طور ساده این تکنیک، عبارت است از جست‌وجو برای یافتن رده‌ی یک شیء در مقایسه با رده‌ی موجود از دیگر اشیاء. تکنیک نزدیک‌ترین همسایگی برای سری‌های زمانی نیز

* نویسنده مسئول
تاریخ: دریافت ۱۳۹۹/۷/۹، اصلاحیه ۱۴۰۰/۲/۶، پذیرش ۱۴۰۰/۴/۲۳.
DOI:10.24200/J65.2021.55318.2100

عبارت است از: «تعیین رده‌ی یک سری زمانی در مقایسه با رده‌ی سایر سری‌های زمانی موجود در یک مجموعه داده.»

تکنیک خوشه‌بندی یک روش یادگیری بدون نظارت است و عبارت است از تقسیم‌بندی مجموعه‌یی از اعضا به گروه‌هایی موسوم به «خوشه». عناصر موجود در هر خوشه با یکدیگر بسیار شبیه‌اند ولی با عناصر سایر خوشه‌ها شباهت کم‌تری دارند. طبق تعریف، خوشه‌بندی سری زمانی به معنای تقسیم کردن یک مجموعه سری زمانی به چند خوشه و با شرایط فوق است. باید توجه داشت که خوشه‌بندی سری زمانی به شدت به روش اندازه‌گیری شباهت وابسته است.[۹-۱۷] تکنیک‌های خوشه‌بندی به چند گروه تقسیم می‌شوند: خوشه‌بندی قسمت‌بندی، خوشه‌بندی سلسله مراتبی، خوشه‌بندی مبتنی بر چگالی، خوشه‌بندی مبتنی بر شبکه.[۱۸] در اکثر این تکنیک‌ها، تعداد خوشه توسط کاربر تعیین می‌شود و به‌عنوان ورودی الگوریتم است. در این تحقیق از الگوریتم خوشه‌بندی کامدوید که زیرمجموعه‌ی تکنیک خوشه‌بندی قسمت‌بندی است استفاده می‌شود. علت انتخاب این الگوریتم نیز عدم نیاز به میانگین‌گیری در مراحل اجرای الگوریتم است.

درادامه، ابتدا ضمن معرفی مفهوم شباهت در سری‌های زمانی، انواع رویکردهای اندازه‌گیری شباهت سری‌های زمانی معرفی می‌شود. سپس، روش‌های اندازه‌گیری شباهت LCSS و DTW تشریح خواهد شد. در بخش سوم، تکنیک‌های نزدیک‌ترین همسایگی و خوشه‌بندی کامدوید معرفی می‌شود. در قسمت چهارم، به تشریح رویکرد ارزیابی به‌کار رفته در این تحقیق می‌پردازیم و در قسمت آخر، نتایج اجرای تکنیک‌های فوق تحت روش‌های LCSS و DTW ارائه و مورد بحث قرار می‌گیرد.

۲. مرور ادبیات

۱.۲. تعریف شباهت و انواع معیارهای اندازه‌گیری شباهت

چنان که پیش‌تر نیز اشاره شد، یکی از مهم‌ترین موضوعات مطرح در داده‌کاوی سری زمانی، مسئله‌ی اندازه‌گیری شباهت سری‌های زمانی است. بر اساس تحقیقات مرتبط، شباهت در سری‌های زمانی به‌صورت شباهت نقطه به نقطه یا شباهت منعطف (یک نقطه به چند نقطه یا چند نقطه به یک نقطه) تعریف می‌شود. شباهت نقطه به نقطه به معنای آن است که هر داده از یک سری زمانی فقط می‌تواند با یک داده از سری زمانی دیگر مشابه باشد و برعکس ـ نظیر روش اندازه‌گیری شباهت اقلیدسی و طولانی‌ترین زیردنباله‌ی مشترک. شباهت یک نقطه به چند نقطه و برعکس نیز، به معنای آن است که یک داده از یک سری زمانی می‌تواند با بیش از یک داده از سری زمانی دیگر شباهت داشته باشد و برعکس، نظیر روش اندازه‌گیری شباهت چرخش زمانی پویا.[۱۹]

از منظر دیگر نیز می‌توان شباهت سری‌های زمانی را به‌صورت شباهت در زمان، شباهت در شکل و شباهت در مدل تعریف کرد. شباهت در زمان به معنای شباهت بین دو سری زمانی بر مبنای میزان شباهت داده‌ها در زمان‌های یکسان است. به عبارت دیگر در این تعریف، زمان وقوع اهمیت دارد. به‌عنوان مثال مقایسه‌ی سری زمانی فروش یک شرکت با سری زمانی فروش شرکت دیگر در همان دوره‌ی زمانی. شباهت در شکل به معنای شباهت بین دو سری زمانی بر مبنای شباهت بین زیردنباله‌های آنهاست. در این نوع شباهت، زمان اهمیتی ندارد و فقط شکل سری زمانی اهمیت دارد. حتی ابعاد طولی و عرضی هم اهمیت ندارد. شباهت در مدل نیز به معنای میزان یکسان بودن پارامترها و مدل برازش شده به سری‌های زمانی است.[۲۰]

برای اندازه‌گیری شباهت دو سری زمانی، دو رویکرد کلی وجود دارد: انطباق کامل[۳] و انطباق زیردنباله.[۴] در رویکرد انطباق کامل، کل طول سری‌های زمانی در فرایند اندازه‌گیری شباهت استفاده می‌شود، به‌طوری که اگر طول هر دو سری زمانی برابر m باشد آنگاه همه‌ی m داده‌ی سری زمانی اول و همه‌ی m داده‌ی سری زمانی دوم در فرایند اندازه‌گیری شباهت این دو سری زمانی کاربرد دارند؛ نظیر روش اندازه‌گیری شباهت اقلیدسی، طولانی‌ترین زیردنباله‌ی مشترک و چرخش زمانی پویا. در رویکرد انطباق زیردنباله، سری‌های زمانی می‌توانند طول‌های متفاوتی داشته باشند، به‌طوری که اگر طول یک سری زمانی برابر n و طول سری زمانی دیگر برابر m باشد و $m > n$ باشد، آنگاه شباهت زیردنباله‌هایی از سری زمانی بزرگ‌تر (با اعضای متوالی) با سری زمانی کوچک‌تر اندازه‌گیری می‌شود و شباهت دو سری زمانی برابر با حداکثر شباهت حاصل از فرایند فوق خواهد بود.[۲۰]

به‌طورکل معیارهای اندازه‌گیری مسافت (شباهت) را می‌توان به چهار گروه کلی تقسیم کرد:[۲۰]

۱. معیارهای مسافت مبتنی بر شکل؛[۵]

۲. معیارهای مسافت مبتنی بر ویرایش؛[۶]

۳. معیارهای مسافت مبتنی بر ویژگی؛[۷]

۴. معیارهای مسافت مبتنی بر مدل.[۸]

در قسمت بعدی از میان تمامی روش‌های اندازه‌گیری شباهت موجود برای داده‌های سری زمانی، به دو روش مطرح DTW و LCSS اشاره می‌شود.

۲.۲. روش چرخش زمانی پویا (DTW)

در توضیح این روش باید عنوان کرد که گاهی اوقات برخی از سری‌های زمانی نسبتاً مشابه‌اند ولی این شباهت نه در زمان یکسان، بلکه با یک تأخیر زمانی اتفاق افتاده است. لذا برای شناسایی شباهت این‌چنین سری‌های زمانی باید روش متفاوتی ایجاد شود؛ این روش متفاوت همان روش چرخش زمانی پویا (DTW) است[۲۱] که می‌تواند شباهت را بین سری‌های زمانی با طول غیریکسان اندازه‌گیری کند. این روش با کشیدن یکی از سری‌های زمانی یا هر دو سری زمانی انجام می‌شود و سعی دارد با این کار، شکل دو سری زمانی را به یکدیگر مشابه کند. این روش که از مفهوم شباهت یک نقطه به چند نقطه و برعکس بهره می‌برد از جمله روش‌های شباهت مبتنی بر شکل محسوب می‌شود و در آن از انطباق کامل در تعیین شباهت دو سری زمانی استفاده می‌شود. در صورتی که $TS_x = (x_1, x_2, \ldots, x_n)$ و بیانگر دو سری زمانی به‌ترتیب با طول‌های n و m $TS_y = (y_1, y_2, \ldots, y_m)$ باشند آنگاه مسافت بین دو سری زمانی طبق روش DTW به‌صورت TS_x, TS_y DTW نمایش داده می‌شود و برابراست با $M(n,m)$ به‌طوری‌که مقدار $M(n,m)$ از رابطه‌ی برگشتی ۱ محاسبه می‌شود.

$$\mathrm{M(i,j)} = \begin{cases} (x_i - y_j)^2 & ;i=1, j=1 \\ (\mathrm{x_i - y_j})^2 + M(\mathrm{i, j-1}) & ;i=1, j\geq 2 \\ (\mathrm{x_i - y_j})^2 + M(\mathrm{i-1, j}) & ;i\geq 2, j=1 \\ (\mathrm{x_i - y_j})^2 + Min \begin{cases} M(\mathrm{i-1, j}) \\ M(\mathrm{i, j-1}) \\ M(\mathrm{i-1, j-1}) \end{cases} & ;i\geq 2, j\geq 2 \end{cases} \quad (1)$$

این روش، دارای دو خروجی است. یکی از این خروجی‌ها، دنباله‌یی با عناصر زوجی و با طول r است به‌طوری که r بزرگ‌تر یا مساوی با $max(n,m)$ است. این دنباله‌ی زوجی را R نامگذاری می‌کنیم. خروجی دیگر، مقدار مسافت (عدم شباهت) دو سری زمانی مورد نظر یا همان $\mathrm{DTW}\ (TS_x, TS_y)$ است که بیان‌گر مجموع مربعات تفاضل عناصر زوجی دنباله‌ی R است.

از آن‌جاکه نمی‌توان میزان عدم تشابه دو سری زمانی TS_x و TS_y با میزان عدم تشابه دو سری زمانی TS_U و TS_V را فقط به استناد مقادیر $\mathrm{DTW}\ (TS_x, TS_y)$ و $\mathrm{DTW}\ (TS_U, TS_V)$ تعیین کرد (زیرا طول این سری‌های زمانی می‌تواند با هم متفاوت باشد و هر مقدارکه طول آنها بیشتر باشد ممکن است مقدار DTW آنها نیز بزرگ‌تر باشد)، لذا از روابط $dissim(TS_x, TS_y) = \sqrt{\frac{\mathrm{DTW}(TS_x, TS_y)}{||r^{۲}||}}$ و $dissim(TS_U, TS_V) = \sqrt{\frac{\mathrm{DTW}(TS_U, TS_V)}{||r^{۲}||}}$ استفاده می‌شود، به‌طوری که $r^{۱}$ و $r^{۲}$ به‌ترتیب طول دنباله‌های ایجاد شده ناشی از اجرای روش DTW برای دو سری زمانی TS_x و TS_y و برای دو سری زمانی TS_U و TS_V است. سپس مقدار $dissim(TS_x, TS_y)$ با مقدار $dissim(TS_U, TS_V)$ مقایسه می‌شود و هرکدام که کم‌تر بود به معنای تشابه بیشتر آن دو سری زمانی نسبت به یکدیگر در مقایسه با تشابه دو سری زمانی دیگر است.

از این روش اندازه‌گیری شباهت در تحقیق‌های بسیاری استفاده شده است، به‌عنوان مثال از روش DTW به‌عنوان روش اندازه‌گیری شباهت در بخش‌بندی بازار از طریق داده‌کاوی،[۲۲] برای انتخاب ویژگی‌ها به منظور دسته‌بندی سریع سری‌های زمانی از طریق الگوریتم ژنتیک کارا،[۲۳] در تشخیص سیگنال‌های زیست‌شناختی به‌عنوان روش اندازه‌گیری شباهت،[۲۴] برای خوشه‌بندی سری‌های زمانی چندمتغیر مبتنی بر امواج با به‌کارگیری رویکرد SPCA،[۲۵] برای بخش‌بندی صفحه[۲۶] استفاده شده است. به طور کلی این روش نسبت به روش‌های دیگر اندازه‌گیری شباهت سری‌های زمانی از عملکرد بهتری برخوردار است،[۲۷] ضمن این که به‌طور وسیع در تحقیقات مورد استفاده قرار می‌گیرد.

۳.۲. روش LCSS

این روش که مسئله‌یی کلاسیک در علم رایانه محسوب می‌شود، در ابتدا برای مقایسه‌ی دنباله‌یی از کاراکترها ایجاد شد. مهم‌ترین ویژگی این روش آن است که می‌تواند از مقادیر دورافتاده و نویزدار چشم‌پوشی کند. شباهت در این روش به‌صورت «عین هم بودن» دو کاراکتر از دو دنباله تعریف می‌شود. در صورتی که S_x و S_y، بیان‌گر دو دنباله به‌ترتیب با طول‌های n و m باشند، آن‌گاه طولانی‌ترین زیردنباله‌ی مشترک آنها به‌صورت $\mathrm{LCSS}(S_x, S_y)$ نمایش داده می‌شود و برابر است با $M(n,m)$ مقدار $M(n,m)$ نیز از رابطه‌ی برگشتی ۲ محاسبه می‌شود و همواره عددی صحیح، بین صفر و $min(n,m)$ است.

$$\mathrm{M(i,j)} = \begin{cases} ۰ & ; i = ۰.or.j = ۰ \\ ۱ + M(\mathrm{i}-۱, \mathrm{j}-۱) & ; \mathrm{x_i} = \mathrm{y_j}, i \geq ۱.or.j \geq ۱ \\ Max \begin{cases} M(\mathrm{i}-۱, \mathrm{j}) \\ M(\mathrm{i}, \mathrm{j}-۱) \end{cases} & ; \mathrm{x_i} \neq \mathrm{y_j}, i \geq ۱.or.j \geq ۱ \end{cases} \quad (۲)$$

روش LCSS دارای دو خروجی است. خروجی اول، طولانی‌ترین زیردنباله‌ی مشترک دو دنباله است که با LCSS نشان داده می‌شود و خروجی دیگر، دنباله‌یی با طول LCSS که بیان‌گر عناصر مشترک بین دو دنباله‌ی مورد نظر است.

از آن‌جاکه نمی‌توان میزان شباهت دو دنباله‌ی S_x و S_y را با میزان تشابه دو دنباله‌ی دیگر نظیر S_U و S_V (فرض کنید طول دنباله‌های S_U و S_V به‌ترتیب برابر با $n^{۱}$ و $m^{۱}$ باشد) فقط به استناد مقادیر $\mathrm{LCSS}(S_x, S_y)$ و $\mathrm{LCSS}(S_U, S_V)$ مشخص کرد (زیرا طول این دنباله‌ها می‌تواند با هم متفاوت باشد و هر مقدارکه طول آنها بیشتر باشد امکان این که مقدار LCSS آنها بزرگ‌تر باشد نیز وجود دارد)، لذا از روابط $sim(S_x, S_y) = \frac{۲*\mathrm{LCSS}(S_x, S_y)}{m+n}$ و $sim(S_U, S_V) = \frac{۲*\mathrm{LCSS}(S_U, S_V)}{m^{۱}+n^{۱}}$ استفاده می‌شود. بدین معنا که مقدار $Sim(S_x, S_y)$ با مقدار $sim(S_U, S_V)$ مقایسه می‌شود و هرکدام که بیشتر بود به‌معنای تشابه بیشتر آن دو دنباله نسبت به هم، در مقایسه با تشابه دو دنباله‌ی دیگر نسبت به هم است.

برای این که بتوان از این روش نیز برای داده‌های سری زمانی استفاده کرد، از حد مجاز شباهت ($\in$) برای تعیین شباهت دو داده از دو سری زمانی استفاده می‌شود.[۲۸] با این توضیحات، رابطه‌ی ۲ به‌صورت رابطه‌ی ۳ بازنویسی می‌شود.

$$\mathrm{M(i,j)} = \begin{cases} ۰ & ; i = ۰.or.j = ۰ \\ ۱ + M(\mathrm{i}-۱, \mathrm{j}-۱) & ; |\mathrm{x_i} - \mathrm{y_j}| \leq \in, i \geq ۱.or.j \geq ۱ \\ Max \begin{cases} M(\mathrm{i}-۱, \mathrm{j}) \\ M(\mathrm{i}, \mathrm{j}-۱) \end{cases} & ; \in < |\mathrm{x_i} - \mathrm{y_j}|, i \geq ۱.or.j \geq ۱ \end{cases} \quad (۳)$$

چنان که مشاهده می‌شود، هرگاه قدرمطلق تفاضل دو داده از دو سری زمانی، کوچک‌تر یا مساوی با مقدار حد مجاز شباهت باشد آن‌گاه دو داده با یکدیگر مشابه‌اند. در این شرایط یک واحد طول به طول طولانی‌ترین زیردنباله‌ی مشترک آنها تا قبل از این دو داده اضافه خواهد شد. در غیر این صورت طول طولانی‌ترین زیردنباله‌ی مشترک آنها برابر با طول طولانی‌ترین زیردنباله‌ی مشترک آنها در شرایطی است که طول آن دنباله‌ها یکی کم‌تر باشد.

نتیجه‌ی حاصل از به‌کارگیری این روش به شدت به مقدار حد مجاز شباهت وابسته است. به بیان واضح، اگر مقدار حد مجاز شباهت کوچک (سخت‌گیرانه) باشد، آن‌گاه طول طولانی‌ترین زیردنباله‌ی مشترک نیز کوتاه‌تر خواهد بود و اگر مقدار حد مجاز شباهت بزرگ (سهل‌گیرانه) باشد، آن‌گاه طول طولانی‌ترین زیردنباله‌ی مشترک نیز بلندتر خواهد بود. واقعیت آن است که به علت عدم آگاهی از ماهیت مجموعه داده، انتخاب بهترین مقدار حد مجاز شباهت بسیار دشوار است و این امر باعث می‌شود که در صورت انتخاب نامناسب مقدار حد مجاز شباهت، خطای نتیجه‌گیری نیز افزایش یابد.[۲۹]

به‌طورکلی، این روش نیز نسبت به روش‌های دیگر اندازه‌گیری شباهت سری‌های زمانی از عملکرد بهتری برخوردار است[۲۷] و در تحقیقات مختلف با اهداف متنوعی استفاده شده است. از این روش به منظور دسته‌بندی متن‌های داده‌یی نیز استفاده شده است.[۳۰] همچنین به منظور طراحی یک سیستم پیش‌بینی کارا برای داده‌های سری زمانی، از این روش با به کارگیری الگوریتم‌های انطباق الگویی استفاده شده است.[۳۱] از این روش برای طراحی یک رویکرد ابتکاری برای بچینگ و زمان‌بندی ماشین‌های منفرد به منظور کمینه‌سازی هزینه‌های راه‌اندازی[۳۲] و نیز برای طراحی یک الگوریتم ممتیک مبتنی بر یادگیری برای مسئله‌ی برداشت و تحویل چندین وسایل باربری منطبق با LIFO [۳۳] استفاده شده است. در برآورد یک الگوریتم توزیع برای کارگاه‌های فلوشاپ به منظور کمینه‌سازی زمان جریان کل[۳۴] و نیز برای سرعت بخشیدن زنجیره‌های تأمین با بهینه سازی کلونی مورچه‌یی در طیف وسیعی از راه حل‌های سخت‌افزاری،[۳۵] از این روش استفاده شده است.

۳. تکنیک‌های داده‌کاوی

۱.۳. تکنیک نزدیک‌ترین همسایگی

اگر TS بیان‌گر یک سری زمانی، DB بیان‌گر یک مجموعه داده سری زمانی و d بیان‌گر روش اندازه‌گیری شباهت باشد، تکنیک نزدیک‌ترین همسایگی، به دنبال یافتن یک سری زمانی از مجموعه‌ی DB است به‌گونه‌یی که دارای کم‌ترین مقدار عدم شباهت (بیشترین شباهت) با سری زمانی TS باشد. البته این تکنیک را می‌توان به‌صورت «یافتن مجموعه‌یی از سری‌های زمانی از مجموعه DB که مقدار عدم شباهت آنها با سری زمانی TS از مقدار مشخصی کم‌تر باشد» نیز تعریف کرد. در تعریفی دیگر، این تکنیک عبارت است از یافتن تعداد مشخصی از سری‌های زمانی از مجموعه DB که دارای کم‌ترین مقدار عدم شباهت با سری زمانی TS باشند.

۲.۳. تکنیک خوشه‌بندی با الگوریتم کامدوید

الگوریتم خوشه‌بندی کامدوید یکی از تکنیک‌های خوشه‌بندی قسمت‌بندی است. در این الگوریتم، مجموعه داده به چندین گروه تقسیم می‌شود، به‌طوری که عناصر هر گروه شباهت بالایی به هم دارند و با عناصر سایر گروه‌ها از شباهت کم‌تری برخوردارند. این الگوریتم در هر مرحله، از مدوید هر گروه به‌عنوان مرکز (نماینده) آن گروه استفاده می‌کند.

مشهورترین الگوریتم خوشه‌بندی کامدوید تحت عنوان الگوریتم قسمت‌بندی از مدوید[9] یا PAM نام دارد. این الگوریتم از جست‌وجوی حریصانه در فرایند قسمت‌بندی استفاده می‌کند و این امکان وجود دارد که به جواب بهینه منجر نشود بلکه به یک جواب بهینه محلی برسد، زیرا به جای جست‌وجوی تمام فضای جواب، فقط بخشی از فضای جواب را جست‌وجو می‌کند. ضمن این که همین امر موجب کاهش شدید زمان دست‌یابی به جواب هم خواهد شد. در این تحقیق از الگوریتم PAM استفاده شده است.

در صورتی که k بیان‌گر تعداد خوشه و n تعداد اعضای مجموعه داده باشد، مراحل اجرای این الگوریتم عبارت است از:

گام ۱. به‌صورت تصادفی، k عضو از مجموعه داده انتخاب می‌شود. هر عضو تصادفی به‌عنوان یکی از مراکز خوشه خواهد بود؛

گام ۲. سایر اعضای مجموعه داده به نزدیک‌ترین مرکز خوشه اختصاص داده می‌شود؛

گام ۳. یک عضو غیر مدوید به‌صورت تصادفی انتخاب می‌شود؛

گام ۴. هزینه‌ی کل (S) که بیان‌گر هزینه‌ی ناشی از جابه‌جایی عضو غیرمدوید انتخاب شده با مدوید همان خوشه است، محاسبه می‌شود. یعنی عضو غیر مدوید به‌عنوان مرکز خوشه‌یی در نظرگرفته می‌شود که این عضو از آن خوشه انتخاب شده است. سپس سایر اعضا به نزدیک‌ترین مرکزخوشه اختصاص می‌یابند. هزینه‌ی خوشه‌بندی وضعیت جدید محاسبه و از هزینه‌ی وضعیت قبلی خوشه‌بندی کسر و به‌عنوان هزینه‌ی کل در نظرگرفته می‌شود؛

گام ۵. اگر S منفی شده باشد، آن‌گاه جابه‌جایی عضو غیرمدوید به‌عنوان مدوید آن خوشه قطعی می‌شود، در غیر این‌صورت هیچ تغییری در ساختار خوشه‌ها اتفاق نمی‌افتد؛

گام ۶. اگر شرط توقف احراز شده باشد، عمل خوشه‌بندی خاتمه می‌پذیرد در غیر اینصورت به گام ۲ مراجعه می‌شود.

شروط توقف می‌تواند یکی یا چند حالت از حالات زیر باشد:

۱. رسیدن به تعداد مشخصی جابه‌جایی؛

۲. رسیدن به تعداد مشخصی جابه‌جایی موفق؛

۳. رسیدن به درصد خاصی از کاهش در هزینه‌ی کل.

۴. رویکرد ارزیابی عملکرد

در این تحقیق، عملکرد روش‌های اندازه‌گیری شباهت LCSS و DTW در مورد تکنیک‌های نزدیک‌ترین همسایگی و خوشه‌بندی با الگوریتم کامدوید روی داده‌های سری زمانی محاسبه و مقایسه می‌شوند. در این تحقیق از ۶۳ مجموعه داده سری زمانی متعلق به بانک اطلاعاتی UCR (با آدرس زیر) که به‌طور تصادفی از میان آنها انتخاب شده‌اند، استفاده شده است.

(http://www/cs.ucr.edu/~eamonn/time_series_data_2018)

نام و مشخصات این مجموعه داده در جدول ۱ ارائه شده است. هر مجموعه داده دارای دو زیرمجموعه است، یکی تحت عنوان مجموعه داده آموزشی و دیگری تحت عنوان مجموعه داده آزمایشی. طول سری‌های زمانی، تعداد خوشه و کلاسه هر سری زمانی این مجموعه‌ها مشخص است. به‌عنوان نمونه، مجموعه داده «Statistical Control» دارای سری‌های زمانی با طول ۶۰، تعداد ۶ خوشه است و تعداد اعضای مجموعه داده آموزشی و آزمایشی آن به‌ترتیب ۳۰۰ و ۳۰۰ سری زمانی است.

از مجموعه داده آموزشی برای کشف الگوهای موجود در مجموعه داده و از مجموعه داده آزمایشی برای ارزیابی کیفیت الگوهای شناسایی شده استفاده می‌شود.

در توضیح مجموعه‌های انتخاب شده، باید عنوان کرد که گروه‌بندی این مجموعه‌ها به علت تأثیر بسیار ویژگی‌هایی نظیر تعداد خوشه، طول سری زمانی و تعداد اعضای مجموعه‌های آموزشی و آزمایشی بر روی نتایج داده‌کاوی میسر نیست و تاکنون در هیچ تحقیقی، این گروه‌بندی انجام نشده است. ضمن این که ماهیت خوشه‌ها از بابت ویژگی‌های پراکندگی و تراکم ـ که به‌ترتیب بیان‌گر میزان دوری یا نزدیکی اعضای یک خوشه به یکدیگر و بیان‌گر میزان دوری یا نزدیکی مراکز خوشه‌ها با یکدیگر است ـ نیز از عوامل بسیار اثرگذار بر خروجی داده‌کاوی است. لذا در این تحقیق، ما نیز از گروه‌بندی این مجموعه‌های داده و در نتیجه، گروه‌بندی نتایج حاصله از داده‌کاوی خودداری می‌کنیم و نتایج را به‌صورت کلی و از دید معیارهای ارزیابی داده‌کاوی بررسی می‌کنیم.

چنان که قبلاً نیز اشاره شد، در این تحقیق از تکنیک نزدیک‌ترین همسایگی به منظور شناسایی رده‌ی یک سری زمانی استفاده می‌شود. هدف از این کار، ارزیابی تأثیر روش اندازه‌گیری شباهت در نتایج حاصل است. بدین معنا که بررسی کنیم و ببینیم با تغییر روش اندازه‌گیری شباهت، نتایج تکنیک چقدر و به چه سمتی تغییر می‌کند. در این تکنیک فرض می‌شود رده‌ی سری‌های زمانی مجموعه آزمایشی مشخص نیستند و با مقایسه‌ی میزان شباهت آنها با سری‌های زمانی مجموعه آموزشی نسبت به تعیین رده‌ی آنها اقدام می‌شود. این فرایند در دو گام انجام می‌پذیرد:

گام ۱. شباهت سری زمانی مجموعه آزمایشی (TS) با تک تک سری‌های زمانی مجموعه آموزشی تحت روش اندازه‌گیری شباهت مورد نظر محاسبه می‌شود؛

گام ۲. رده‌ی سری زمانی (TS) تشخیص داده می‌شود که برابر است با رده‌ی شبیه‌ترین سری زمانی مجموعه آموزشی با آن تحت روش اندازه‌گیری شباهت مورد نظر.

جدول ۱. نام و مشخصات مجموعه داده‌های استفاده شده از بانک اطلاعاتی UCR.

ردیف	نام مجموعه داده	تعداد خوشه	طول سری زمانی	تعداد اعضای مجموعه داده آموزشی	تعداد اعضای مجموعه داده آزمایشی	ردیف	نام مجموعه داده	تعداد خوشه	طول سری زمانی	تعداد اعضای مجموعه داده آموزشی	تعداد اعضای مجموعه داده آزمایشی
۱	Statistical Control	۶	۶۰	۳۰۰	۳۰۰	۳۳	SonyII	۲	۶۵	۲۷	۹۵۳
۲	Gun-Point	۲	۱۵۰	۵۰	۱۵۰	۳۴	sonySurface	۲	۷۰	۲۰	۶۰۱
۳	CBF	۳	۱۲۸	۳۰	۹۰۰	۳۵	StarLightCurve	۳	۱۰۲۴	۱۰۰۰	۸۲۳۶
۴	ECG	۲	۹۶	۱۰۰	۱۰۰	۳۶	Two Lead ECG	۲	۸۲	۲۳	۱۱۳۹
۵	Face4	۴	۳۵۰	۲۴	۸۸	۳۷	Criket X	۱۲	۳۰۰	۳۹۰	۳۹۰
۶	Medical	۱۰	۹۹	۳۸۱	۷۶۰	۳۸	Criket Y	۱۲	۳۰۰	۳۹۰	۳۹۰
۷	Sweedian	۱۵	۱۲۸	۵۰۰	۶۲۵	۳۹	U Wave X	۸	۳۱۵	۸۹۶	۳۵۸۲
۸	OSU	۶	۴۲۷	۲۰۰	۲۴۲	۴۰	U Wave Y	۸	۳۱۵	۸۹۶	۳۵۸۲
۹	Adiac	۳۷	۱۷۶	۳۹۰	۳۹۱	۴۱	Insect Wing	۱۱	۲۵۶	۲۲۰	۱۹۸۰
۱۰	Beef	۵	۴۷۰	۳۰	۳۰	۴۲	Arrow Head	۳	۲۵۱	۳۶	۱۷۵
۱۱	Lighting	۷	۳۱۹	۷۰	۷۳	۴۳	Beetle Fly	۲	۵۱۲	۲۰	۲۰
۱۲	Fish	۷	۴۶۳	۱۷۵	۱۷۵	۴۴	Bird Chicken	۲	۵۱۲	۲۰	۲۰
۱۳	50words	۵۰	۲۷۰	۴۵۰	۴۵۵	۴۵	Ham	۲	۴۳۱	۱۰۹	۱۰۵
۱۴	Trace	۴	۲۷۵	۱۰۰	۱۰۰	۴۶	Phalanges O-C	۲	۸۰	۱۸۰۰	۸۵۸
۱۵	Lighting7	۷	۳۱۹	۷۰	۷۳	۴۷	Proximal POA	۳	۸۰	۴۰۰	۲۰۵
۱۶	Distal	۷	۸۰	۱۳۹	۴۰۰	۴۸	Proximal POC	۲	۸۰	۶۰۰	۲۹۱
۱۷	Italy power demand	۲	۲۴	۶۷	۱۰۲	۴۹	Proximal PT	۶	۸۰	۲۰۵	۴۰۰
۱۸	Middle-P-T	۷	۸۰	۱۵۴	۳۹۹	۵۰	Toe Segmentation1	۲	۲۷۷	۴۰	۲۲۸
۱۹	Plane	۷	۱۴۴	۱۰۵	۱۰۵	۵۱	Toe Segmentation2	۲	۳۴۳	۳۶	۱۸۰
۲۰	Car	۴	۵۷۷	۶۰	۶۰	۵۲	Distal POA	۳	۸۰	۱۳۹	۴۰۰
۲۱	Olive Oil	۴	۵۷۰	۳۰	۳۰	۵۳	Distal POC	۲	۸۰	۲۷۶	۶۰۰
۲۲	Diatom Size	۴	۳۴۵	۱۶	۳۰۶	۵۴	Distal PT	۶	۸۰	۱۳۹	۴۰۰
۲۳	Symbol	۶	۳۹۸	۲۵	۹۹۵	۵۵	Earth quakes	۲	۵۱۲	۱۳۹	۳۲۲
۲۴	Worms	۵	۹۰۰	۷۷	۱۸۱	۵۶	Middle POA	۳	۸۰	۱۵۴	۴۰۰
۲۵	Two Pattern	۴	۱۲۸	۱۰۰۰	۴۰۰	۵۷	Middle POC	۲	۸۰	۲۹۱	۶۰۰
۲۶	Wafer	۲	۱۵۲	۱۰۰۰	۶۱۶	۵۸	Shapelet Sim	۲	۵۰۰	۲۰	۱۸۰
۲۷	Faceall	۱۴	۱۳۱	۵۳۰	۱۶۹	۵۹	Wine	۲	۲۳۴	۵۷	۵۴
۲۸	Lighting2	۲	۶۳۷	۶۰	۶۱	۶۰	Computers	۲	۷۲۰	۲۵۰	۲۵۰
۲۹	ECGFiveday	۲	۱۳۶	۲۳	۸۶۱	۶۱	Meat	۳	۴۴۸	۶۰	۶۰
۳۰	Haptics	۵	۱۰۹۲	۱۵۵	۳۰۸	۶۲	Refrigeration	۳	۷۲۰	۳۷۵	۳۷۵
۳۱	InLineSkate	۷	۱۸۸	۱۰۰	۵۵۰	۶۳	Worm Two Class	۲	۹۰۰	۷۷	۱۸۱
۳۲	Motestrain	۲	۸۴	۲۰	۱۲۵						

پس از اجرای گام‌های فوق، شاخص «دقت» تکنیک محاسبه می‌شود. طبق تعریف، شاخص دقت عبارت است از نسبت تعداد سری‌های زمانی مجموعه آزمایشی که رده‌ی آنها به درستی تشخیص داده شده است به تعداد کل سری‌های زمانی مجموعه آزمایشی.

تکنیک دیگری که در این تحقیق از آن استفاده شده است، تکنیک خوشه‌بندی است. هدف از این کار، بررسی تأثیر روش‌های اندازه‌گیری شباهت در خوشه‌بندی مناسب مجموعه داده است. بدین معنا که، تأثیر روش‌های اندازه‌گیری شباهت موردنظر در تعیین تعداد درست خوشه، اعضا و نماینده‌ی آنها محاسبه و ارزیابی خواهد شد. از میان تکنیک‌های مختلف خوشه‌بندی، از الگوریتم خوشه‌بندی کامدوید استفاده می‌شود، زیرا این الگوریتم نیازی به انجام عمل میانگین‌گیری از اعضای خوشه برای تعیین مرکز خوشه در هر یک از مراحل خوشه‌بندی ندارد. فرایند ارزیابی نتایج خوشه‌بندی در دو گام انجام می‌شود.

گام ۱. خوشه‌بندی مجموعه داده آموزشی: عمل خوشه‌بندی با الگوریتم کامدوید با مقادیر مختلف تعداد خوشه اجرا و برای هر مقدار از تعداد خوشه، ۵۰۰ بار عمل خوشه‌بندی تکرار می‌شود و در هر بار تکرار نیز امکان جابه‌جایی نماینده خوشه به تعداد ۵۰۰ مرتبه وجود دارد. به عنوان مثال اگر تعداد خوشه برابر ۲ باشد، عمل خوشه‌بندی به تعداد ۵۰۰ مرتبه اجرا می‌شود و در هر بار تکرار این امکان وجود دارد که برای بهتر شدن مقدار تابع هدف تا ۵۰۰ مرتبه نماینده‌های خوشه به‌هنگام شوند. بعد از اجرای خوشه‌بندی، کم‌ترین مقدار تابع هدف در هر حالت (مقدار تعداد خوشه) به‌عنوان بهترین وضعیت آن حالت در نظر گرفته می‌شود و مشخصات خوشه‌بندی بهینه شامل نماینده خوشه‌ها و اعضای آنها ذخیره می‌شود. به‌عنوان مثال، مشخص می‌شود که بهترین مقدار تابع هدف خوشه‌بندی در حالتی که تعداد خوشه برابر ۲ است درکدام وضعیت خوشه‌بندی (نماینده و اعضای خوشه) ایجاد شده است. همین روال برای تعیین بهترین مقدار تابع هدف خوشه‌بندی در حالتی که تعداد خوشه برابر ۳ و ... باشد نیز انجام می‌شود. سپس با به‌کارگیری شاخص زاویه[۱۰] بهترین مقدار تعداد خوشه از میان تعداد خوشه‌های ممکن انتخاب خواهد شد. در انتها نیز «دقت» خوشه‌بندی که

عبارت است از نسبت تعداد سری‌های زمانی که به درستی به خوشه درست خود اختصاص یافته‌اند به کل سری‌های زمانی مجموعه داده آموزشی، محاسبه می‌شود. خروجی این گام، تعیین بهترین مقدار تعداد خوشه برای مجموعه داده، تعیین دقت و مشخصات بهترین خوشه‌بندی است؛

گام ۲. گروه‌بندی مجموعه داده آزمایشی: در این گام از نتایج گام اول (تعداد خوشه بهینه و نماینده آنها) برای گروه‌بندی مجموعه داده آزمایشی استفاده می‌شود. بدین معنا که، شبیه‌ترین نماینده خوشه به هر سری زمانی مجموعه داده آزمایشی تعیین می‌شود و سری زمانی مذکور به آن خوشه اختصاص داده می‌شود. در انتها «دقت» گروه‌بندی که عبارت است از نسبت تعداد سری‌های زمانی‌یی که به درستی به خوشه درست خود اختصاص یافته‌اند به کل سری‌های زمانی مجموعه داده آزمایشی، محاسبه می‌شود. خروجی گام دوم، ارزیابی مناسب بودن کیفیت نماینده خوشه حاصل از گام اول خوشه‌بندی است.

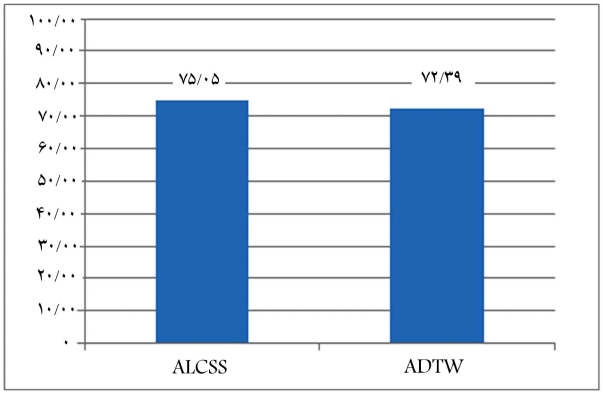
۵. نتایج ارزیابی و بحث

۱.۵. نتایج اجرای الگوریتم نزدیک‌ترین همسایگی

نتایج تکنیک نزدیک‌ترین همسایگی تحت مقادیر مختلف حد مجاز شباهت روش LCSS و DTW در جدول ۲ ارائه شده است. به‌عنوان مثال در مورد مجموعه داده‌ی «Statistical Control» وقتی حد مجاز شباهت ($\in$) برابر ۰٫۰۵ است آن‌گاه رده‌ی ۶۹٫۶۷٪ از سری‌های زمانی آزمایشی به درستی شناسایی شده است. دقت این تکنیک برای این مجموعه داده با تغییر مقدار حد مجاز شباهت تغییر می‌کند، به‌طوری که با افزایش مقدار آن از ۰٫۰۵ تا ۰٫۳۵، منجر به افزایش دقت تکنیک از ۶۹٫۶۷٪ تا ۹۴٫۶۷٪ می‌شود. متوسط و انحراف معیار مقادیر این دقت‌ها برای این مجموعه به‌ترتیب برابر ۸۷٫۹۵٪ و ۸٫۶۳٪ است. روند تأثیر تغییرات حد مجاز شباهت بر دقت این تکنیک برای تمامی مجموعه‌های داده به صورتی که اکنون توضیح داده شد نیست بلکه به‌عنوان مثال، این تأثیر در مورد مجموعه داده «Gun Point» با افزایش مقدار حد مجاز شباهت، در ابتدا افزایشی و سپس کاهشی خواهد بود، به‌طوری که بیشینه‌ی آن در مقدار حد مجاز شباهت ۰٫۱۵ رخ می‌دهد و متوسط و انحراف معیار دقت آن نیز به‌ترتیب برابر ۹۳٫۳۳٪ و ۵٫۰۵٪ است.

به‌طور کل، متوسط دقت و انحراف معیار این تکنیک برای همه مجموعه داده‌ها به‌ترتیب برابر ۷۳٫۴۱٪ و ۴٫۸۰٪ است، به‌طوری که این موضوع بیانگر وابستگی دقت این تکنیک به مقدار حد مجاز شباهت است. لذا انتخاب نادرست مقدار حد مجاز می‌تواند منجر به نتایج نامناسبی شود. شایان ذکر است که بیشترین مقدار دقت متوسط کل در حد مجاز شباهت ۰٫۲۵ رخ می‌دهد و برابر با ۷۵٫۰۵٪ است.

در جدول ۲، دقت این تکنیک تحت روش DTW نیز ارائه شده است. به‌عنوان مثال در مورد مجموعه داده «Statistical Control»، با دقت ۹۷٫۳۳٪، رده‌ی سری‌های زمانی آزمایشی به درستی شناسایی شده است. اما متأسفانه، این دقت برای برخی مجموعه‌های داده بسیار نامطلوب است، نظیر مجموعه داده‌ی OSU و Middle-P-T به‌ترتیب با مقادیر دقت ۴۶٫۲۸٪ و ۵۸٫۴٪ به عبارت دیگر، عملکرد روش DTW در برخی از مجموعه‌های داده بسیار خوب و در برخی دیگر متوسط و در برخی موارد نامناسب است. در نهایت این که، دقت این تکنیک برای کلیه مجموعه‌های داده تحت روش DTW برابر ۷۲٫۳۸٪ است. نمودار ۱ بیانگر متوسط دقت تکنیک نزدیک‌ترین همسایگی تحت روش‌های LCSS و DTW است.



شکل ۱. متوسط کل دقت تکنیک نزدیک‌ترین همسایگی تحت روش‌های LCSS و DTW.

برای مقایسه‌ی دقت این تکنیک تحت روش‌های فوق‌الذکر از آزمون مقایسات جفتی t استفاده می‌کنیم تا از این طریق و دقت این تکنیک برای یک مجموعه داده سری زمانی تحت دو روش مورد نظر، بتوانیم تشخیص دهیم که تأثیر کدامیک در تکنیک نزدیک‌ترین همسایگی بهتر است.

بنابراین اگر A_{LCSS} بیانگر دقت تکنیک نزدیک‌ترین همسایگی تحت روش LCSS و A_{DTW} بیانگر دقت تکنیک نزدیک‌ترین همسایگی تحت روش DTW باشد، آنگاه فرضیات این آزمون به شرح زیر خواهد بود.

$$A_{\mathrm{LCSS}} = A_{\mathrm{DTW}} : H_0$$

$$A_{\mathrm{LCSS}} > A_{\mathrm{DTW}} : H_1$$

نتایج این آزمون در جدول ۳ و برای سطوح مختلفی از قابلیت اطمینان نمایش داده شده است. بر اساس این نتایج، می‌توان با ۹۲٫۵٪ اطمینان ادعا کرد که دقت تکنیک نزدیک‌ترین همسایگی تحت روش LCSS از دقت این تکنیک تحت روش DTW بهتر است.

این مقایسه را می‌توان، از منظر آمار توصیفی نیز انجام داد که در جدول ۴ نمایش داده شده است. بر اساس اطلاعات مندرج در این جدول، دقت تکنیک نزدیک‌ترین همسایگی تحت روش LCSS نسبت به دقت این تکنیک تحت روش DTW در ۷۱٫۴۲٪ حالات بهتر، در ۱٫۵۸٪ از حالات برابر و فقط در ۲۶٫۹۸٪ حالات بدتر است.

به‌طور خلاصه این بررسی‌ها نشان می‌دهد که روش LCSS نسبت به روش DTW تأثیر بهتری در دستیابی به دقت بالاتر دارد.

۲.۵. نتایج اجرای الگوریتم خوشه‌بندی کامدوید

در این قسمت نتایج حاصل از اجرای الگوریتم خوشه‌بندی کامدوید تحت روش‌های LCSS و DTW طبق فرایند ارزیابی توضیح داده شده در قسمت ۴ ارائه خواهد شد. این نتایج در دو قسمت، نتایج خوشه‌بندی و نتایج گروه‌بندی ارائه می‌شود. هدف از این ارزیابی پاسخ به سؤالات زیر است:

۱. آیا دقت تکنیک خوشه‌بندی تحت روش LCSS با دقت این تکنیک تحت روش DTW تفاوت دارد؟

۲. آیا تأثیر روش LCSS با تأثیر روش DTW در تعیین تعداد خوشه بهینه حاصل از خوشه‌بندی تفاوت دارد؟

جدول ۲. دقت تکنیک نزدیک‌ترین همسایگی تحت روش‌های LCSS و DTW (برحسب درصد).

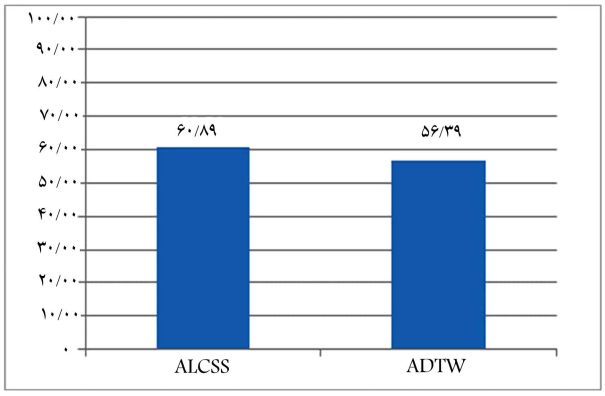
ردیف	نام مجموعه داده	تحت روش LCSS با مقادیر مختلف حد مجاز شباهت									تحت روش DTW
		حد مجاز شباهت (ε)							میانگین (Mean)	انحراف معیار (Std)	بهترین مقدار ε =۰/۲۵
		۰,۰۵	۰,۱۰	۰,۱۵	۰,۲۰	۰,۲۵	۰,۳۰	۰,۳۵			
۱	statistical control	۶۹,۶۷	۸۶,۶۷	۸۷	۹۱,۶۷	۹۳	۹۳	۹۴,۶۷	۸۷,۹۶	۸,۶۳	۹۳	۹۷,۳۳
۲	Gun-Point	۹۲,۶۷	۹۸	۹۸,۶۷	۹۷,۳۳	۹۲,۶۷	۸۸,۶۷	۸۵,۳۳	۹۳,۳۳	۵,۰۵	۹۲,۶۷	۹۰,۶۷
۳	CBF	۹۶,۸۹	۹۸,۱۱	۹۹,۵۶	۹۹,۴۴	۹۹,۶۷	۹۹,۷۸	۹۹,۸۹	۹۹,۰۵	۱,۱۲	۹۹,۶۷	۹۹
۴	ECG	۷۷	۸۵	۸۶	۸۷	۹۰	۹۱	۹۱	۸۶,۷۱	۴,۹۲	۹۰	۷۹
۵	Face4	۸۱,۸۲	۸۹,۷۷	۹۲,۰۵	۹۳,۱۸	۹۴,۵	۹۲,۰۵	۹۴,۳۲	۹۱,۱۰	۴,۴۰	۹۴,۵	۸۱,۸۲
۶	Medical	۵۳,۹۲	۵۷,۷۶	۶۰,۱۳	۶۱,۸۴	۶۲,۳۳	۶۳,۱۶	۶۱,۷۱	۶۰,۱۲	۳,۲۶	۶۲,۳۳	۶۵,۳۹
۷	Sweedian	۴۳,۲	۷۵,۲	۸۲,۴	۸۳,۵	۸۴,۶۲	۸۴,۹۶	۸۲,۷۲	۷۶,۶۶	۱۵,۱۱	۸۴,۶۲	۷۲,۱۶
۸	OSU	۶۲,۴	۶۷,۳۶	۶۸,۶	۶۹,۴۲	۶۹,۸۳	۶۸,۱۸	۶۸,۱۸	۶۷,۷۱	۲,۴۸	۶۹,۸۳	۴۶,۲۸
۹	Adiac	۷۷,۹۲	۶۶,۸۸	۵۵,۱۹	۴۸,۸۶	۴۲,۷	۳۳,۳۸	۲۶,۲۳	۵۰,۱۷	۱۸,۲۰	۴۲,۷	۷۴,۰۳
۱۰	Beef	۵۶,۶۷	۷۰	۷۰	۷۶,۶۷	۷۳	۶۳,۳۳	۵۶,۶۷	۶۶,۶۲	۷,۸۹	۷۳	۶۳,۳۳
۱۱	Lighting	۶۰,۲۷	۵۶,۱۶	۶۹,۸۶	۷۱,۳۳	۷۵,۳۴	۷۵,۳۴	۶۸,۴۹	۶۸,۱۱	۷,۳۳	۷۵,۳۴	۶۸,۴۹
۱۲	Fish	۸۴	۸۸	۸۹,۱۴	۸۴,۵۷	۸۱,۶۹	۷۷,۱۴	۷۳,۱۴	۸۲,۵۳	۵,۷۴	۸۱,۶۹	۷۵,۴۳
۱۳	50words	۶۰,۵	۶۸,۳۵	۷۳,۱۱	۷۳,۵۱	۷۳,۹۵	۷۵,۳۵	۷۷,۰۳	۷۱,۶۹	۵,۶۱	۷۳,۹۵	۶۶,۳۹
۱۴	Trace	۹۲	۹۷	۱۰۰	۹۹	۹۸	۹۶	۹۴	۹۶,۵۷	۲,۸۲	۹۸	۱۰۰
۱۵	Lighting7	۶۰,۲۷	۵۶,۱۶	۶۹,۸۶	۷۱,۲۳	۷۵,۳۴	۷۵,۳۴	۶۸,۴۹	۶۸,۱۰	۷,۳۲	۷۵,۳۴	۶۸,۴۹
۱۶	Distal	۶۹,۵	۷۳,۵	۷۴	۷۵,۲۵	۷۵,۵	۷۶	۷۶,۲۵	۷۴,۲۹	۲,۳۴	۷۵,۵	۷۰,۵۰
۱۷	Italy power demand	۷۸,۵۲	۸۲	۸۷,۵۶	۹۰,۱۸	۹۱,۷۴	۹۲,۴۲	۹۳,۷۲	۸۸,۰۲	۵,۷۳	۹۱,۷۴	۹۳,۹۷
۱۸	Middle-P-T	۵۸,۴	۶۱,۶۵	۵۶,۶۴	۵۹,۹	۶۲,۴۱	۶۱,۹	۶۱,۱۵	۶۰,۲۹	۲,۱۱	۶۲,۴۱	۵۸,۴۰
۱۹	Plane	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۰	۱۰۰	۱۰۰
۲۰	Car	۸۸,۳۳	۸۵	۸۵	۸۳,۴	۸۱,۶۷	۷۳,۳۳	۷۳,۳۳	۸۱,۴۴	۵,۸۹	۸۱,۶۷	۷۱,۶۷
۲۱	Olive Oil	۷۷,۳۳	۵۶,۶۷	۵۳,۳۳	۴۸	۴۵	۴۰	۴۰	۵۱,۴۸	۱۳,۰۱	۴۵	۸۳,۳۳
۲۲	Diatom Size	۹۵,۱۰	۹۵,۷۵	۹۶,۴۱	۹۴,۴۴	۹۳,۴۵	۹۱,۸۳	۸۱,۳۷	۹۲,۶۲	۵,۱۹	۹۳,۴۵	۹۶,۴۱
۲۳	Symbol	۹۵,۰۸	۹۵,۸۸	۹۶,۲۸	۹۶,۸۸	۹۵,۴۸	۹۳,۹۷	۹۲,۴۶	۹۵,۱۵	۱,۵۰	۹۵,۴۸	۹۴,۲۷
۲۴	Worms	۴۴,۱۸	۴۵,۳۳	۴۵,۲۹	۴۸,۰۷	۵۱,۳۷	۵۱,۹۳	۵۴,۱۵	۴۸,۶۲	۳,۸۹	۵۱,۳۷	۴۵,۳۳
۲۵	Two Pattern	۹۳,۷	۹۹	۹۹,۶	۹۹,۶	۹۹,۲	۹۹,۲	۹۹,۲	۹۸,۵۰	۲,۱۳	۹۹,۲	۱۰۰
۲۶	Wafer	۹۸,۰۴	۱۰۰	۱۰۰	۹۸,۷۲	۹۸,۷۲	۹۸,۷۲	۹۸,۳۵	۹۸,۹۳	۰,۷۷	۹۸,۷۲	۹۸,۰۴
۲۷	Faceall	۵۶,۹۸	۷۳,۶۱	۷۸,۴۶	۸۴,۰۳	۸۸,۱۷	۹۳,۰۲	۹۳,۰۲	۸۱,۰۴	۱۲,۸۳	۸۸,۱۷	۸۴,۷۴
۲۸	Lighting2	۷۰,۴۹	۷۳,۷۷	۷۷,۰۵	۷۸,۶۹	۷۸,۶۹	۷۸,۶۹	۷۷,۰۵	۷۶,۳۵	۳,۱۲	۷۸,۶۹	۸۶,۹۰
۲۹	ECGFiveday	۷۰,۷۳	۷۶,۷۷	۷۶,۷۷	۸۰,۰۲	۸۱,۴۲	۸۴,۰۳	۸۶,۵۳	۷۹,۴۹	۵,۲۸	۸۱,۴۲	۷۵,۸۴
۳۰	Haptics	۳۹,۳۹	۳۹,۲۹	۴۵,۱۳	۴۳,۱۸	۴۳,۱۸	۴۱,۲۳	۴۱,۲۳	۴۱,۷۹	۲,۱۷	۴۳,۱۸	۴۱,۲۳
۳۱	InLineSkate	۳۵,۴۵	۳۵,۸	۳۶,۴	۳۷,۲۷	۳۲,۲۷	۴۰	۳۸	۳۷,۱۳	۱,۵۴	۳۲,۲۷	۳۸,۹۰
۳۲	Motestrain	۹۱,۰۵	۹۱,۰۵	۹۱,۲۹	۹۲,۰۹	۹۲,۸۱	۹۲,۶۵	۹۱,۵۴	۹۱,۷۸	۰,۷۴	۹۲,۸۱	۸۱,۱۶
۳۳	SonyII	۷۰,۸۳	۸۰,۸۰	۸۰,۸۰	۸۱,۲۲	۸۲,۴۸	۸۳,۳۲	۸۳,۳۲	۸۰,۳۹	۴,۳۶	۸۲,۴۸	۸۴,۵۸
۳۴	sonySurface	۶۰,۰۷	۶۱,۰۷	۶۲,۰۶	۶۵,۷۳	۶۶,۰۶	۶۷,۷۲	۶۸,۲۲	۶۴,۴۲	۳,۳۰	۶۶,۰۶	۶۵,۲۳
۳۵	StarLightCurve	۶۸	۷۲	۷۶,۰۱	۷۶,۰۱	۷۸	۷۲	۶۴	۷۲,۲۹	۴,۹۶	۷۸	۶۶,۱۰
۳۶	Two Lead ECG	۸۶,۸۳	۹۲,۱	۸۹,۶۵	۸۹,۶۵	۸۶,۸۳	۸۵,۹۵	۸۱,۵۶	۴۵,۹	۵۲,۵۶	۸۶,۸۳	۸۹,۴۷
۳۷	Criket X	۳۴,۶۲	۳۹,۲۳	۴۵,۰۵	۴۴,۳۶	۴۵,۹۰	۵۲,۵۶	۵۲,۵۶	۴۴,۴۷	۶,۶۳	۴۵,۹۰	۶۹,۹۹
۳۸	Criket Y	۳۸,۹۸	۴۵,۳۸	۴۸,۴۶	۵۶,۴۱	۵۶,۴۱	۵۶,۴۱	۵۴,۸۷	۵۰,۹۹	۶,۹۰	۵۶,۴۱	۵۶,۴۰
۳۹	U Wave X	۵۶,۲۵	۵۶,۲۵	۵۷,۷۹	۵۷,۷۹	۵۹,۳۸	۵۹,۳۸	۶۵,۶۳	۵۸,۹۳	۳,۲۲	۵۹,۳۸	۶۵,۶۱
۴۰	U Wave Y	۴۰,۹۰	۴۰,۹۰	۴۰,۹۰	۴۰,۹۰	۴۳,۱۹	۵۰	۵۲,۲۶	۴۴,۱۵	۴,۸۹	۴۳,۱۹	۴۰,۹۳
۴۱	Insect Wing	۳۰,۲۶	۵۱,۵۲	۵۴,۶۰	۵۶,۱۶	۵۳,۰۸	۵۳,۰۸	۵۳,۰۸	۵۰,۲۵	۸,۹۴	۵۳,۰۸	۲۷,۲۷
۴۲	Arrow Head	۶۵,۷۲	۶۸,۵۷	۶۸,۵۷	۷۳,۷۱	۷۰,۸۶	۶۷,۴۳	۶۷,۴۳	۶۸,۹۰	۲,۶۴	۷۰,۸۶	۷۰,۳۱
۴۳	Beetle Fly	۸۰	۸۰	۸۰	۸۰	۸۰	۸۰	۸۰	۸۰	۰	۸۰	۷۰
۴۴	Bird Chicken	۷۵	۸۵	۸۵	۸۵	۹۰	۹۰	۸۵	۸۵	۵	۹۰	۷۵
۴۵	Ham	۶۰	۶۵,۷۱	۶۵,۷۱	۷۱,۴۳	۷۴,۲۹	۶۸,۵۷	۶۵,۷۱	۶۷,۳۵	۴,۶۳	۷۴,۲۹	۴۲,۸۶
۴۶	Phalanges O-C	۶۶,۳۲	۶۸,۳۰	۶۹,۱۱	۷۰,۵۲	۷۱,۱۰	۷۱,۱۰	۶۸,۱۸	۶۹,۲۳	۱,۷۸	۷۱,۱۰	۶۹,۲۳
۴۷	Proximal POA	۷۸,۵۴	۷۸,۵۴	۷۸,۵۴	۷۹,۵۱	۷۹,۵۱	۷۷,۵۶	۷۷,۵۶	۷۸,۵۴	۰,۸۰	۷۹,۵۱	۷۴,۶۴
۴۸	Proximal POC	۸۵,۵۷	۸۹	۹۵,۸۸	۹۴,۸۵	۹۱,۷۵	۸۹,۳۵	۸۷,۹۷	۹۰,۶۲	۳,۷۳	۹۱,۷۵	۸۷,۶۳
۴۹	Proximal PT	۷۴,۵	۷۴,۵	۷۸,۲۵	۷۹	۷۸,۵	۷۵,۷۵	۷۴,۷۵	۷۶,۴۶	۲,۰۴	۷۸,۵	۷۴
۵۰	Toe Segmentation1	۷۶,۷۶	۷۶,۷۶	۷۸,۰۷	۸۲,۴۶	۸۷,۲۸	۸۲,۴۶	۸۰,۷۱	۸۰,۶۴	۳,۸۲	۸۷,۲۸	۷۶,۳۱
۵۱	Toe Segmentation2	۹۳,۸۹	۹۳,۸۹	۹۵,۵۶	۹۶,۶۷	۹۶,۶۷	۹۵,۵۶	۹۵,۵۶	۹۵,۴۰	۱,۱۴	۹۶,۶۷	۸۶,۱۱
۵۲	Distal POA	۷۸	۸۰,۵	۷۹,۵	۷۸	۷۷,۵	۷۴,۵	۷۴	۷۷,۴۳	۲,۴۱	۷۷,۵	۸۳
۵۳	Distal POC	۷۳,۳۴	۷۳,۳۴	۷۳,۳۴	۷۳,۳۴	۷۶,۶۷	۷۵	۷۵	۷۴,۲۹	۱,۳۱	۷۶,۶۷	۷۱,۸۳
۵۴	Distal PT	۷۳	۷۳,۵	۷۵,۵	۷۵,۵	۷۵,۵	۷۵,۵	۷۷,۷۵	۷۵,۱۸	۱,۵۶	۷۵,۵	۷۲,۷۵
۵۵	Earth quakes	۶۲,۱۱	۶۲,۱۱	۶۵,۸۴	۶۵,۸۴	۶۸,۰۱	۷۱,۷۴	۷۱,۷۴	۶۶,۷۷	۴	۶۸,۰۱	۶۲,۴۲
۵۶	Middle POA	۶۵,۷۵	۶۵,۷۵	۶۸	۷۲	۶۷	۶۶,۵	۶۶,۵	۶۷,۳۶	۲,۱۹	۶۷	۶۴,۵۰
۵۷	Middle POC	۷۱,۱۷	۷۶,۶۷	۷۷,۱۷	۸۲,۸۴	۸۵,۳۳	۸۲,۸۴	۸۲,۸۴	۷۹,۸۳	۴,۹۹	۸۵,۳۳	۷۹
۵۸	Shapelet Sim	۷۳,۸۹	۷۳,۸۹	۸۶,۶۷	۸۸,۸۹	۹۱,۶۷	۹۲,۷۸	۹۲,۷۸	۸۵,۸۰	۸,۴۳	۹۱,۶۷	۵۶,۱۱
۵۹	Wine	۵۷,۴۱	۵۱,۸۵	۴۸,۱۵	۴۸,۱۵	۵۰	۵۰	۵۰	۵۰,۷۹	۳,۱۸	۵۰	۵۷,۴۰
۶۰	Computers	۶۹,۲	۶۹,۲	۶۲	۶۲	۶۲	۶۲	۶۲	۶۴,۰۶	۳,۵۱	۶۲	۵۴,۸۰
۶۱	Meat	۹۰	۹۳,۳۳	۶۶,۶۷	۵۳,۳۳	۴۸,۳۳	۴۵	۳۳,۳۳	۶۱,۴۳	۲۲,۹۴	۴۸,۳۳	۹۳,۳۳
۶۲	Refrigeration	۴۹,۰۷	۴۵,۰۷	۴۵,۰۷	۳۹,۴۸	۳۷,۳۳	۳۹,۲	۳۷,۳۳	۴۱,۷۹	۴,۵۹	۳۷,۳۳	۳۳,۳۴
۶۳	Worm Two Class	۶۲,۹۸	۷۱,۸۲	۷۴,۵۹	۷۴,۵۹	۷۴,۵۹	۷۱,۸۲	۷۴,۵۹	۷۲,۱۴	۴,۲۴	۷۴,۵۹	۷۱,۲۷
	متوسط دقت	۶۹,۸۵	۷۳,۰۵	۷۴,۱۳	۷۴,۷۵	۷۵,۰۵	۷۴,۱۴	۷۲,۹۰	۷۳,۴۱	۴,۸۰	۷۵,۰۵	۷۲,۳۸

جدول ۳. نتایج آزمون مقایسات جفتی t برای دقت تکنیک نزدیک‌ترین همسایگی‌تحت روش‌های LCSS و DTW.

	Paired Differences		T	Df	sig. (۲ − tailed)	Correlation
	Mean	Std.Deviation				
$A_{LCSS}-A_{DTW}$	۲٫۶۵۹۲۵	۱۲٫۹۳۵۵۵	۱٫۶۳۲	۶۲	۰٫۱۰۸	$A_{LCSS}\&A_{DTW}=$ ۰٫۷۲۸
	Confidence Interval of the Differenc					
		٪۹۰	٪۹۲٫۵	٪۹۵	٪۹۷٫۵	٪۹۹
$A_{LCSS}-A_{DTW}$	Lower	۰٫۵۴۸۱۸	۰٫۲۸۳۷۸	-۰٫۰۶۲۰۷	-۰٫۵۹۸۵۲	-۱٫۲۳۲۵۵

جدول ۴. مقایسه‌ی آماری دقت تکنیک نزدیک‌ترین همسایگی تحت روش‌های LCSS وDTW .

ردیف	توضیحات	تعداد	درصد
۱	$A_{LCSS} > A_{DTW}$	۴۵	۷۱٫۴۲
۲	$A_{LCSS} = A_{DTW}$	۱	۱٫۵۸
۳	$A_{LCSS} < A_{DTW}$	۱۷	۲۶٫۹۸



نمودار ۲. متوسط کل دقت خوشه‌بندی مجموعه داده آموزشی تحت روش‌های LCSS و DTW.

۳. آیا تأثیر روش LCSS با تأثیر روش DTW در تعیین نماینده خوشه بهینه حاصل از خوشه‌بندی تفاوت دارد؟

۱.۲.۵. نتایج خوشه‌بندی مجموعه آموزشی

جدول ۵ نتایج اجرای تکنیک خوشه‌بندی کامدوید را تحت روش‌های DTW و LCSS نشان می‌دهد. بر اساس این نتایج و به‌عنوان مثال برای مجموعه داده‌ی «Statistical Control»، بهترین تعداد خوشه تحت روش DTW برابر ۶ خواهد بود که دارای دقت خوشه‌بندی ٪۹۷٫۶۷ است. بهترین تعداد خوشه تحت روش LCSS نیز برابر ۶ و با دقت خوشه‌بندی ٪۸۷٫۳۳ خواهد بود. یعنی این تکنیک تحت هر دو روش فوق توانسته است تعداد خوشه را به درستی تعیین کند.

با نگاه مجدد به نتایج مندرج در جدول ۵، مشاهده می‌شود که دقت خوشه‌بندی تحت روش‌های فوق، متفاوت است که برای نتیجه‌گیری درست باید به‌طور آماری بررسی شود. از سوی دیگر، دقت تکنیک خوشه‌بندی برای کلیه مجموعه‌های داده تحت روش DTW برابر ٪۵۶٫۵۵ و تحت روش LCSS برابر ٪۶۰٫۸۹ است. نمودار ۲ بیانگر متوسط دقت تکنیک نزدیک‌ترین همسایگی تحت روش‌های LCSS و DTW است.

به منظور پاسخ به سؤال اول، از آزمون مقایسات زوجی t استفاده می‌شود.

بنابراین، اگر A_{LCSS} بیانگر دقت تکنیک خوشه‌بندی تحت روش LCSS و A_{DTW} بیانگر دقت تکنیک خوشه‌بندی تحت روش DTW باشد، آنگاه فرضیات این آزمون به شرح زیر خواهد بود.

$$A_{LCSS} = A_{DTW} : H_0$$

$$A_{LCSS} > A_{DTW} : H_1$$

نتایج این آزمون در جدول ۶ و برای سطوح مختلفی از قابلیت اطمینان نمایش داده شده است. بر اساس این نتایج، می‌توان حداقل با ٪۹۹ اطمینان ادعا کرد که دقت خوشه‌بندی تحت روش LCSS از دقت این تکنیک تحت روش DTW بهتر است.

این مقایسه را می‌توان از منظر آمار توصیفی نیز انجام داد. نتایج این مقایسه در جدول ۷ نمایش داده شده است. بر اساس اطلاعات مندرج در این جدول، دقت خوشه‌بندی تحت روش LCSS نسبت به دقت خوشه‌بندی تحت روش DTW، در ٪۵۳٫۹۷ حالات بهتر، در ٪۱۲٫۷۰ از حالات برابر و فقط در ٪۳۳٫۳۳ حالات بدتر است. به‌طور خلاصه این بررسی نشان می‌دهد که روش LCSS تأثیر بهتری در دست‌یابی به دقت بالاتر نسبت به روش DTW دارد.

در ادامه و در جدول ۸، تأثیر روش‌های LCSS و DTW در تعیین درست تعداد خوشه نشان داده شده است. از اطلاعات این جدول برای پاسخ به سؤال دوم استفاده می‌کنیم. بر اساس این اطلاعات، در خوشه‌بندی تحت روش DTW و با به‌کارگیری شاخص زاویه، می‌توان تعداد خوشه را برای ۲۸ مجموعه داده به‌درستی و برای ۱۷ مجموعه داده با اختلاف یک عدد تشخیص داد. این مقادیر به‌ترتیب معادل ٪۴۴٫۴۴ و ٪۲۷ (جمعاً برابر ٪۷۱٫۴۴) از کل مجموعه داده‌ها می‌شود. از سوی دیگر، در خوشه‌بندی تحت روش LCSS و با به‌کارگیری شاخص زاویه می‌توان تعداد خوشه را برای ۲۵ مجموعه داده به‌درستی و برای ۱۹ مجموعه داده با اختلاف یک عدد تشخیص داد. این مقادیر نیز به‌ترتیب معادل ٪۳۹٫۶۸ و ٪۳۰٫۱۵ (جمعاً برابر ٪۶۹٫۸۳) از کل مجموعه داده‌ها هستند. این مطالب نشان از تأثیر نسبتاً برابر این روش‌ها در شناسایی تعداد درست خوشه دارد.

به‌طور خلاصه، روش LCSS تأثیر بهتری در تعیین اعضای هر خوشه نسبت به تأثیر روش DTW دارد، ولی از نظر تأثیر در تعیین درست تعداد خوشه، تقریباً از اثر یکسانی برخوردارند.

۲.۲.۵. نتایج گروه‌بندی مجموعه داده آزمایشی

چنان که قبلاً توضیح داده شد، پس از خوشه‌بندی مجموعه داده آموزشی و تعیین بهترین تعداد خوشه و به تبع آن عناصر خوشه و نماینده هر خوشه، نسبت به گروه‌بندی مجموعه داده آزمایشی مبتنی بر این نتایج اقدام می‌شود. بدین معنا که اعضای

جدول ۵. نتایج خوشه‌بندی کامدوید (تعداد خوشه بهینه و دقت خوشه‌بندی برحسب درصد روی مجموعه آموزشی.

ردیف	نام مجموعه داده	تحت روش DTW		تحت روش LCSS	
		دقت	تعداد خوشه	دقت	تعداد خوشه
۱	statistical control	۹۷,۶۷	۶	۸۷,۲۳	۶
۲	Gun-Point	۵۶	۲	۵۶	۴
۳	CBF	۹۶,۶۷	۳	۹۶,۶۷	۳
۴	ECG	۶۰	۲	۷۳	۲
۵	Face4	۶۶,۶۷	۳	۸۳,۳۳	۵
۶	Medical	۳۲,۲۹	۶	۳۱,۷۶	۶
۷	Sweedian	۵۲,۴۰	۱۹	۶۶,۴۰	۱۳
۸	OSU	۴۳	۵	۴۷,۵	۱۰
۹	Adiac	۴۲,۳۱	۳۰	۳۷,۱۷	۳۳
۱۰	Beef	۴۳,۳۳	۶	۴۶,۶۷	۷
۱۱	Lighting	۵۵,۷۱	۶	۶۰	۱۰
۱۲	Fish	۵۹,۵۷	۹	۷۲,۵۷	۶
۱۳	50words	۶۴,۹۰	۴۸	۴۸,۲۲	۴۵
۱۴	Trace	۷۸	۳	۵۲	۲
۱۵	Lighting7	۵۵,۷۱	۶	۵۵,۷۱	۷
۱۶	Distal	۵۹	۳	۶۷,۶۳	۳
۱۷	Italy power demand	۷۳,۱۴	۳	۶۷,۱۶	۳
۱۸	Middle-P-T	۵۵,۸۵	۲	۵۵,۸۵	۲
۱۹	Plane	۱۰۰	۷	۱۰۰	۷
۲۰	Car	۵۶,۶۷	۵	۷۰	۴
۲۱	Olive Oil	۸۶,۶۷	۴	۶۵	۳
۲۲	Diatom Size	۱۰۰	۴	۱۰۰	۴
۲۳	Symbol	۹۶	۶	۱۰۰	۶
۲۴	Worms	۲۲,۰۸	۲	۲۸,۵۷	۶
۲۵	Two Pattern	۹۷,۵۰	۴	۶۰,۳۹	۴
۲۶	Wafer	۶۲,۳۵	۲	۷۰,۴۸	۲
۲۷	Faceall	۶۷,۸۶	۱۵	۷۸,۳۹	۱۴
۲۸	Lighting2	۶۱,۶۷	۶	۴۸,۳۳	۴
۲۹	ECGFiveday	۶۵,۲۲	۲	۶۰,۸۷	۳
۳۰	Haptics	۲۸,۳۹	۴	۴۲,۵۸	۷
۳۱	InLineSkate	۳۸	۷	۳۸	۸
۳۲	Motestrain	۵۵	۳	۹۰	۲
۳۳	SonyII	۷۴,۰۷	۴	۹۲,۶۰	۲
۳۴	sonySurface	۶۰	۳	۶۵	۲
۳۵	StarLightCurve	۶۶,۰۷	۴	۷۵	۳
۳۶	Two Lead ECG	۶۰,۵۸	۴	۷۸,۲۶	۲
۳۷	Criket X	۴۱,۸۰	۱۳	۲۹,۲۳	۱۱
۳۸	Criket Y	۳۹,۲۳	۱۱	۳۲,۸۲	۱۳
۳۹	U Wave X	۴۹,۸۹	۶	۴۵,۸۶	۷
۴۰	U Wave Y	۴۸,۰۵	۵	۴۱,۹۰	۱۱
۴۱	Insect Wing	۲۵,۴۶	۱۱	۴۱,۸۲	۷
۴۲	Arrow Head	۶۱,۱۱	۲	۶۳,۸۹	۲
۴۳	Beetle Fly	۶۵	۲	۸۵	۲
۴۴	Bird Chicken	۵۵	۴	۶۵	۳
۴۵	Ham	۴۴,۹۶	۲	۵۲,۲۹	۲
۴۶	Phalanges O-C	۳۳,۸۳	۲	۴۲,۱۱	۴
۴۷	Proximal POA	۷۸,۲۵	۳	۷۴,۴۵	۲
۴۸	Proximal POC	۳۸,۳۳	۲	۳۸,۳۳	۲
۴۹	Proximal PT	۵۳,۱۷	۴	۶۷,۸۱	۲
۵۰	Toe Segmentation1	۳۵	۲	۵۲,۵	۳
۵۱	Toe Segmentation2	۸۰,۵۶	۲	۷۷,۷۸	۲
۵۲	Distal POA	۴۶,۰۴	۷	۴۸,۹۲	۶
۵۳	Distal POC	۴۳,۸۴	۲	۴۳,۴۸	۲
۵۴	Distal PT	۵۸,۹۹	۳	۶۸,۳۴	۳
۵۵	Earth quakes	۴۸,۲۰	۲	۴۳,۸۸	۳
۵۶	Middle POA	۳۷,۰۱	۳	۵۳,۲۵	۲
۵۷	Middle POC	۶۲,۸۹	۲	۶۶,۶۷	۲
۵۸	Shapelet Sim	۳۰	۲	۶۰	۳
۵۹	Wine	۴۷,۳۷	۲	۵۷,۸۹	۲
۶۰	Computers	۴۰,۴	۲	۶۲	۲
۶۱	Meat	۷۸,۳۳	۳	۵۶	۲
۶۲	Refrigeration	۴۴,۸۰	۴	۵۳,۶۰	۴
۶۳	Worm Two Class	۲۶,۶۸	۴	۴۲,۸۶	۲

جدول ۶. نتایج آزمون مقایسات جفتی t برای دقت تکنیک خوشه‌بندی تحت روش‌های LCSS و DTW.

	Paired Differences		T	Df	sig. (۲ − tailed)	Correlation
	Mean	Std.Deviation				
$A_{LCSS} - A_{DTW}$	۴٫۳۴۶۳۵	۱۲٫۴۳۴۴۵	۲٫۷۷۴	۶۲	۰٫۱۰۷	$A_{LCSS}\&A_{DTW} =$ ۰٫۷۸۹
	Confidence Interval of the Differenc					
		٪۹۰	٪۹۲٫۵	٪۹۵	٪۹۷٫۵	٪۹۹
$A_{LCSS} - A_{DTW}$	Lower	۲٫۳۱۷۰۵	۲٫۰۶۲۹۰	۱٫۷۳۰۴۴	۱٫۲۱۴۷۷	۰٫۶۰۵۳۱

جدول ۷. مقایسه‌ی آماری دقت خوشه‌بندی تحت روش‌های LCSS وDTW .

ردیف	توضیحات	تعداد	درصد
۱	$A_{LCSS} > A_{DTW}$	۳۴	۵۳٫۹۷
۲	$A_{LCSS} = A_{DTW}$	۸	۱۲٫۷۰
۳	$A_{LCSS} < A_{DTW}$	۲۱	۳۳٫۳۳

مجموعه آزمایشی به شبیه‌ترین نماینده خوشه اختصاص داده شده و دقت گروه‌بندی محاسبه می‌شود. نتایج این گروه‌بندی تحت روش‌های فوق در جدول ۹ ارائه شده است.

در این جدول و به‌عنوان مثال برای مجموعه داده‌ی «Statistical Control» اگر تعداد و نماینده خوشه بهینه حاصل از خوشه‌بندی مجموعه داده آموزشی این

جدول ۸. خلاصه تعداد حالات پیش‌بینی درست تعداد خوشه تحت روش‌های LCSS و DTW.

ردیف	توضیحات	DTW	LCSS
۱	تعداد حالاتی که تعداد خوشه به درستی پیش‌بینی شده است	۲۸	۲۵
۲	تعداد حالاتی که تعداد خوشه، با اختلاف یک عدد پیش‌بینی شده است	۱۷	۱۹

جدول ۹. دقت گروه‌بندی مجموعه داده‌های آزمایشی مبتنی بر نتایج خوشه‌بندی تحت روش‌های LCSS و DTW.

ردیف	نام مجموعه داده	DTW	LCSS	ردیف	نام مجموعه داده	DTW	LCSS	ردیف	نام مجموعه داده	DTW	LCSS
۱	statistical control	۹۶	۸۵,۵۷	۲۲	Diatom Size	۸۴,۶۴	۹۱,۸۳	۴۳	Beetle Fly	۶۰	۸۰
۲	Gun-Point	۴۸	۴۶,۶۷	۲۳	Symbol	۹۱,۷۶	۹۴,۱۷	۴۴	Bird Chicken	۴۰	۴۵
۳	CBF	۹۳,۳۳	۹۱	۲۴	Worms	۲۳,۷۶	۳۴,۲۵	۴۵	Ham	۴۵,۷۲	۴۱,۹۱
۴	ECG	۵۴	۷۱	۲۵	Two Pattern	۹۷,۱۷	۵۹,۳۲	۴۶	Phalanges O-C	۴۴,۹۹	۳۱,۳۵
۵	Face4	۴۴,۳۲	۸۶,۳۴	۲۶	Wafer	۵۹,۳۸	۶۸,۵۸	۴۷	Proximal POA	۷۶,۱۰	۸۵,۳۷
۶	Medical	۳۲,۷۶	۲۶,۴۵	۲۷	Faceall	۵۸,۸۱	۶۳,۶۷	۴۸	Proximal POC	۳۶,۴۳	۳۶,۴۳
۷	Sweedian	۵۵,۰۴	۶۳,۸۴	۲۸	Lighting2	۳۲,۷۹	۳۶,۰۷	۴۹	Proximal PT	۵۲,۵	۷۲,۲۵
۸	OSU	۳۵,۵۴	۳۹,۶۷	۲۹	ECGFiveday	۶۰,۸۶	۵۳,۶۶	۵۰	Toe Segmentation1	۴۱,۶۷	۵۶,۱۴
۹	Adiac	۳۷,۸۵	۳۵,۲۹	۳۰	Haptics	۱۹,۱۶	۲۹,۲۲	۵۱	Toe Segmentation2	۸۲,۳۱	۷۴,۶۲
۱۰	Beef	۴۶,۶۷	۵۰	۳۱	InLineSkate	۱۸	۱۷,۲۶	۵۲	Distal POA	۶۵,۷۵	۶۹
۱۱	Lighting	۵۱,۴۳	۵۰,۴۹	۳۲	Motestrain	۶۹,۵۷	۸۸,۵۸	۵۳	Distal POC	۳۷,۱۷	۳۷,۱۷
۱۲	Fish	۵۲	۷۲,۵۷	۳۳	SonyII	۶۶,۳۲	۷۴,۱۹	۵۴	Distal PT	۷۳,۲۵	۷۷,۲۵
۱۳	50words	۴۰,۶۶	۴۵,۴۹	۳۴	sonySurface	۴۴,۵۹	۴۳,۲۶	۵۵	Earth quakes	۴۷,۵۲	۴۱,۳۱
۱۴	Trace	۷۲	۴۸	۳۵	StarLightCurve	۶۷,۵۹	۷۵,۴۱	۵۶	Middle POA	۶۸,۷۵	۲۵,۵
۱۵	Lighting7	۵۲,۰۱	۴۹,۳۲	۳۶	Two Lead ECG	۶۰,۵۸	۶۶,۳۷	۵۷	Middle POC	۴۴,۱۷	۴۸,۵
۱۶	Distal	۷۳,۲۵	۷۸,۲۵	۳۷	Criket X	۳۸,۴۶	۱۹,۷۴	۵۸	Shapelet Sim	۲۹,۴۴	۴۸,۸۹
۱۷	Italy power demand	۷۳,۸۶	۶۴,۳۴	۳۸	Criket Y	۳۶,۶۷	۲۸,۲۰	۵۹	Wine	۴۸,۱۵	۵۰
۱۸	Middle-P-T	۶۱,۱۶	۶۱,۱۶	۳۹	U Wave X	۵۴,۱۰	۴۷,۶۳	۶۰	Computers	۳۸,۴۰	۶۵,۲۰
۱۹	Plane	۹۹,۰۵	۹۹,۰۵	۴۰	U Wave Y	۴۵,۰۹	۳۶,۶۸	۶۱	Meat	۷۱,۶۷	۴۶,۶۷
۲۰	Car	۴۳,۳۳	۵۱,۶۷	۴۱	Insect Wing	۱۸,۳۳	۴۲,۶۸	۶۲	Refrigeration	۴۱,۶۰	۴۶,۱۳
۲۱	Olive Oil	۸۶,۶۷	۵۶,۶۷	۴۲	Arrow Head	۴۸	۵۲,۵۷	۶۳	Worm Two Class	۳۴,۲۵	۵۶,۹۱

مجموعه داده تحت روش DTW را برای گروه‌بندی مجموعه داده‌ی آزمایشی خودش تحت روش DTW به کار ببریم، دقت گروه‌بندی برابر ۹۶٪ خواهد بود. اگر با استفاده از تعداد و نماینده خوشه بهینه حاصل از خوشه‌بندی مجموعه داده‌ی آموزشی همین مجموعه داده تحت روش LCSS، عمل گروه‌بندی اعضای مجموعه داده‌ی آزمایشی همین مجموعه داده را تحت روش LCSS انجام دهیم، آنگاه دقت گروه‌بندی برابر ۸۵٫۵۷٪ خواهد بود. این عملیات برای سایر مجموعه داده‌ها نیز انجام شده است.

بر اساس این نتایج، متوسط دقت گروه‌بندی کل برای همه مجموعه‌های داده تحت روش DTW و LCSS به‌ترتیب برابر ۵۴٫۳۵٪ و ۵۶٫۰۶٪ است که در نمودار ۳ نمایش داده شده است.

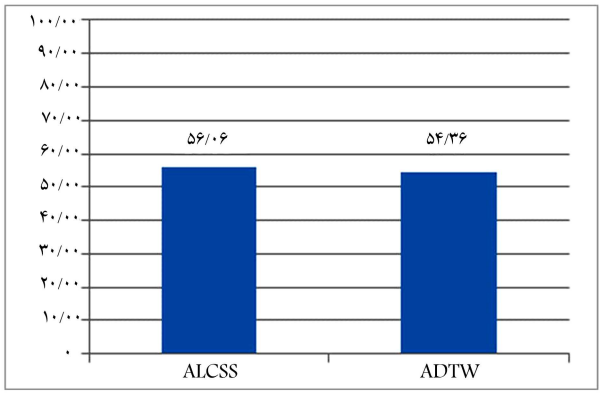
با این توضیحات و به منظور پاسخ به سؤال سوم این تحقیق، مجدداً از آزمون مقایسات زوجی t استفاده می‌کنیم.

اگر A_{LCSS} بیان‌گر دقت گروه‌بندی مجموعه داده آزمایشی تحت روش LCSS و A_{DTW} بیان‌گر دقت گروه‌بندی مجموعه داده آزمایشی تحت روش DTW باشد، آنگاه فرضیات این آزمون به شرح زیر خواهد بود.

$$A_{\mathrm{LCSS}} = A_{\mathrm{DTW}} : H_0$$

$$A_{\mathrm{LCSS}} \neq A_{\mathrm{DTW}} : H_1$$

نتایج این آزمون در جدول ۱۰ و برای سطوح مختلفی از قابلیت اطمینان نمایش



نمودار ۳. متوسط کل دقت گروه‌بندی مجموعه داده آزمایشی تحت روش‌های LCSS و DTW.

داده شده است. بر اساس این نتایج، می‌توان فرض صفر را پذیرفت و ادعا کرد که در سطوح مختلف اطمینان، دقت گروه‌بندی مجموعه داده‌ی آزمایشی تحت روش LCSS با دقت گروه‌بندی مجموعه داده‌ی آزمایشی تحت روش DTW، تفاوتی ندارد. به عبارت دیگر، کیفیت نماینده خوشه‌های حاصله از خوشه‌بندی مجموعه داده‌های آموزشی تحت روش‌های یاد شده از کیفیت یکسانی برخوردار است و تفاوت معناداری با یکدیگر ندارند.

جدول ۱۰. نتایج آزمون مقایسات جفتی t برای دقت گروه‌بندی مجموعه داده آزمایشی تحت روش‌های LCSS و DTW.

	Paired Differences		T	Df	sig. ($2-$tailed)	Correlation
	Mean	Std.Deviation				
$A_{\text{LCSS}}-A_{\text{DTW}}$	۱٫۷۱۴۲۹	۱۴٫۶۵۸۹۵	۰٫۹۲۸	۶۲	۰٫۳۵۷	$A_{\text{LCSS}}\&A_{\text{DTW}}=$ ۰٫۷۳۲
	Confidence Interval of the Differenc					
		٪۸۰	٪۹۰	٪۹۵	٪۹۷٫۵	٪۹۹
$A_{\text{LCSS}}-A_{\text{DTW}}$	Lower	-۰٫۶۷۸۰۵	-۱٫۳۶۹۶۰	-۱٫۹۷۷۵۲	-۲٫۵۲۸۲۰	-۳٫۱۹۳۶۹
	Upper	۴٫۱۰۶۶۲	۴٫۷۹۸۱۷	۵٫۴۰۶۰۹	۵٫۹۵۶۷۷	۶٫۶۲۲۲۶

جدول ۱۱. مقایسه‌ی آماری دقت خوشه‌بندی تحت روش‌های LCSS وDTW .

ردیف	توضیحات	تعداد	درصد
۱	$A_{\text{LCSS}} > A_{\text{DTW}}$	۳۵	۵۵٫۵۶
۲	$A_{\text{LCSS}} = A_{\text{DTW}}$	۴	۶٫۳۵
۳	$A_{\text{LCSS}} < A_{\text{DTW}}$	۲۴	۳۸٫۱۰

این مقایسه را می‌توان از منظر آمار توصیفی نیز انجام داد. نتایج این مقایسه در جدول ۱۱ نمایش داده شده است. بر اساس اطلاعات مندرج در این جدول، دقت گروه‌بندی تحت روش LCSS نسبت به دقت این تکنیک تحت روش DTW، در ٪۵۵٫۵۶ از حالات، بهتر، در ٪۶٫۳۵ از حالات برابر و فقط در ٪۳۳٫۳۳ از حالات، بدتر است. به‌طور خلاصه، این آمارها به همراه انحراف معیار اختلاف دقت گروه‌بندی تحت دو روش فوق نشان می‌دهد که تأثیر روش LCSS و تأثیر روش DTW در تعیین نماینده خوشه تفاوتی ندارد.

۶. نتیجه‌گیری

در یک جمع‌بندی کلی، می‌توان ادعا کرد که دقت تکنیک نزدیک‌ترین همسایگی تحت روش LCSS با اطمینان ٪۹۲٫۵ به نتایج بهتری نسبت حالتی خواهد داشت که این تکنیک تحت روش DTW اجرا شود. می‌توان ادعا کرد که با اطمینان ٪۹۹ دقت خوشه‌بندی کامدوید تحت روش LCSS نسبت به روش DTW به نتایج بهتری دست خواهیم یافت. تأثیر روش LCSS نسبت به تأثیر روش DTW در تعیین تعداد خوشه‌های مجموعه داده به‌صورت آماری یکسان است. تأثیر روش LCSS با تأثیر روش DTW در تعیین نماینده خوشه نیز تقریباً تفاوتی ندارد به‌طوری که گاه یکی از آنها برای برخی مجموعه داده‌های سری زمانی خوب وگاه دیگری برای سایر مجموعه داده‌های سری زمانی خوب است و این تأثیر حدوداً بینابین است.

لازم به ذکر است که به منظور بررسی جامع‌تر تأثیر روش‌های یاد شده در داده کاوی سری زمانی، می‌توان آنها را در تکنیک‌های دیگر داده‌کاوی نیز مورد ارزیابی قرار داد یا از تکنیک‌های آماری دیگری برای ارزیابی و مقایسه‌ی آنها استفاده کرد.

پانوشت‌ها

1. longest common sub sequence (LCSS)
2. dynamic time warping (DTW)
3. whole matching
4. subsequence matching
5. shape based distance measure
6. edit based distance measure
7. feature based distance measure
8. model based distance measure
9. partitioning algorithm of medoid (PAM)
10. elbow index

منابع (References)

1. Morris, B. and Trivedi, M. "Learning trajectory patterns by clustering: experimental studies and comparative evaluation", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 312-319 (2009).
2. Fu, T.C. "A review on time series data mining", *Engineering Applications of Artificial Intelligence*, **24**(1) pp. 164-181 (2011).
3. Keogh, E. and Kasetty, S. "On the need for time series data mining benchmarks: a survey and empirical demonstration", *Data Mining and Knowledge Discovery*, **7**(4), pp. 349-371 (2003).
4. Sangeeta, R. and Geeta, S. "Recent techniques of clustering of time series data: a survey", *International Journal of Computer Applications*, **52**(15), pp. 1-9 (2012).
5. Lin, J., Vlachos, M., Keogh, E. and et al. "Iterative incremental clustering of time series", *International Conference on Extending Database Technology, Advances in Database Technology*, pp. 106-122 (2004).
6. Gordon, A.D. "Clustering algorithms and cluster validity", Computational Statistics, PhysicaVerlag, pp. 497-512 (1994).
7. Milligan, G.W. and Cooper, M.C. "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, **50**, pp. 159-179 (1985).

8. Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.Y. "Time series clustering-A decade review", *Information Systems*, **53**, pp. 16-38 (2015).
9. Faloutsos, C. Ranganathan, M. and Manolopoulos, Y. "Fast subsequence matching in time-series databases", *In Proceedings of the ACMSIGMOD International Conference on Management of Data*, **23**, pp. 419-429 (1994).
10. Sakoe, H. and Chiba, S. "A dynamic programming approach to continuous speech recognition", *In Proceedings of the 7th International Congress on Acoustics*, **3**, pp. 65-69 (1971).
11. Sakoe, H. and Chiba, S. "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **26**(1), pp. 43-49 (1987).
12. Chen, L. and Ng, R. "On the marriage of Lp-norms and edit distance", *In Proceedings of the 30th International Conference on Very Large Data Bases*, **30**, pp. 792-803 (2004).
13. Aßfalg, J., Kriegel, H.P., Kroger, P. and et al. "Similarity search on time series based on threshold queries", In Advances in Database Technology, pp. 276-294 (2006).
14. Vlachos, M., Kollios, G. and Gunopulos, D. "Discovering similar multidimensional trajectories", *In Proceedings of the 18th International Conference on Data Engineering*, pp. 673-684 (2002).
15. Banerjee, A. and Ghosh, J. "Clickstream clustering using weighted longest common subsequences", *In Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining*, pp. 33-40 (2001).
16. Chen, L., Ozsu, M.T. and Oria, V. "Robust and fast similarity search for moving object trajectories", *In Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 491-502 (2005).
17. Keogh, E., Lonardi, S., Ratanamahatana, C.A. and et al. "Compression-based data mining of sequential data", *Data Mining and Knowledge Discovery*, **14**(1), pp. 99-129 (2007).
18. Han, J., Kamber, M. and Pei, J., *Data Mining: Concepts and Techniques*, 3rd ed (2011).
19. Rousseeuw, P.J. "Silhouettes: a graphical Aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, **20**, pp. 53-65 (1987).
20. Hubert, L. and Arabie, P. "Comparing partitions", *Journal of Classification*, **2**, pp. 193-218 (1985).
21. Notredame, C., Higgins, D.G. and Heringa, J. "TCoffee: a novel method for fast and accurate multiple sequence alignment", *Journal of Molecular Biology*, **302**(1), pp. 205-217 (2000).
22. Murray, P.W., Agard, B. and Barajas, A. "Market segmentation through data mining: a method to extract behaviors from a noisy data set", *Computers&Industrial Engineering*, **109**, pp. 233-252 (2017).
23. Ahn, G. and Hur, S. "Efficient genetic algorithm for feature selection for early time series classification", *Computers & Industrial Engineering*, **142**, 106345 (2020).
24. Luzianin, I. and Krause, B. "Similarity measurement of biological signals using dynamic time warping algorithm", *Proc. of the 4th International Conference on Applied Innovations in IT*, (2016).
25. Barragan, J.F., Fontes, C.H. and Embicucu, M. "A wavelet-based clustering of multivariate time series using a multiscale SPCA approach", *Computers and Industrial Engineering* (2016). https://doi.org/10.1016/j.cie.2016.03.003.
26. Hong, Jae.Y., Park, Seung. H. and Baek, Jun-G. "SS-DTW:Shape segment dynamic time warping", *Expert Systems with Applications*, **150**, 113291 (2020).
27. Just, W. "Computational complexity of multiple sequence alignment with SP-score", *Journal of Computational Biology*, **8**(6), pp. 615-623 (2001).
28. Hubert, L. and Schultz, J. "Quadratic assignment as a general data-analysis strategy", *British Journal of Mathematical and Statistical Psychology*, **29**, pp. 190-241 (1987).
29. Chou, C.H., Su, C. and Lai, E. "A new cluster validity measure for clusters with different densities", *IASTED International Conference on Intelligence Systems and Control*, pp. 276-281 (2003).
30. Cho, Su.G. and Kim, Seoung. B. "A Data-driven document similarity measure based on classification algorithms", *International Journal of Industrial Engineering: Theory, Applications and Practice*, **24**(3), pp.328-339(2017).
31. Sridevi, S., Parthasarathy, S. and Rajaram, S. "An Effective prediction system for time series data using pattern matching algorithms", *International Journal of Industrial Engineering: Theory, Applications and Practice*, **25**(2), pp.123-136 (2018).
32. Agnetisa, A., Alfierib, A. and Nicosia, G. "An heuristics approach to batching and scheduling a singal machine to minimize setup costs", *Copmuters & Industrial Engineering*, **46**, pp. 793-802 (2004).
33. Penga, B., Zhanga, Y., Lub, Z. and et al. "A learning based memetic algorithm for the multiple vehicle pickup and delivery problem with LIFO loading", *Copmuters & Industrial Engineering*, **142**, 106241 (2020).
34. Zhang, Y. and Xiaoping, L. "Estimation of distribution algorithm for permutation flow shops with total flowtime minimization", *Copmuters & Industrial Engineering*, **60**, pp. 706-718 (2011).
35. Dzalbs, I. and Kalganova, T. "Accelerating supply chains with ant colony optimization across a range of hardware solutions", *Copmuters & Industrial Engineering*, **147**, 106610, (2020).