Archive of SID

سیزدهمین کنفرانس آمار ایران
۴-۲ شهریور ۱۳۹۵

تطبیق الگوریتم خوشه‌بندی هارتیگان-وانگ برای داده‌هایی از آمار شکل

حامد جعفری [۱] **و موسی گل‌علی‌زاده**

گروه آمار، دانشگاه تربیت مدرس

چکیده: در دنیای امروز کاربرد روزافزون روش‌های آماری در علوم مختلف ایجاب می‌کند تا علم آمار نیز روش‌های خود را متناسب با نیاز آن‌گونه مسائل توسعه دهد. یکی از زمینه‌های کاربردی جدید که کمتر از نیم قرن است توجه دانشمندان علم آمار را به خور معطوف کرده علومی است که با تحلیل آماری داده‌های معرف شکل یک شیء سر و کار دارند که تحلیل آن‌ها با داده‌های چندمتغیره معمولی متفاوت است. در این میان یکی از تحلیل‌های مور توجه تعمیم روش‌های خوشه‌بندی چندمتغیره برای داده‌های شکل است. در این نوشتار سعی شده‌است یکی از الگوریتم‌های معروف در خوشه‌بندی داده‌های اقلیدسی، برای داده‌های شکل تطبیق داده شود. برای این منظور، پس از بررسی نظری مسأله کاربست آن در مثال واقعی نیز تشریح می‌شود.

واژه‌های کلیدی: خوشه‌بندی، آمارشکل، k-میانگین، الگوریتم هارتیگان-وانگ، داده‌های جمجمه گوریل.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H30، 62H99.

۱ مقدمه

به نظر می‌رسد، تشخیص تمایز اشکال از هم و پیدا کردن مشابه‌ها و متفاوت‌ها همواره و در همه سطوح سنی به نوعی مطرح بوده است. برای مثال معمولاً یکی از معیارهای جالبی که از طریق آن، سطح هوشی افراد سنجیده می‌شود، توانایی تشخیص اشکال مشابه است. البته کاربرد این مسأله به همین‌جا ختم نمی‌شود؛ هرکجا که لازم باشد اشکال، چهره‌ها، استخوان‌ها، گونه‌های گیاهی و جانوری، ریخت و بافت لایه‌های مختلف کره‌ی زمین و یا کرات دیگر و بقیه اجرام آسمانی و به‌طور خلاصه هرکجا که صحبت از تمایز اشیاء برحسب شکل آن‌ها مطرح باشد، مسأله «خوشه‌بندی شکل‌ها» نیز مطرح است. به طورکلی، مسأله خوشه‌بندی به صورت عام برای هر نوع داده‌ای مطرح می‌شود.

خوشه‌بندی گرچه خود به‌تنهایی سابقه‌ای دیرین دارد و دست‌کم به بیش از نیم‌قرن پیش برمی‌گردد (جانسون و ویچرن، ۱۹۹۲)، اما به‌کارگیری روش‌های آن در حوزه آمار شکل، عمری کوتاه داشته طوری که سابقه آن به آغاز قرن ۲۱ برمی‌گردد. مرور نوشتگان پیشینیان نشان می‌دهد که خوشه‌بندی داده‌های شکل، نخستین بار توسط له‌له و ریچسمییر (۲۰۰۱) مورد مطالعه قرارگرفت. آن‌ها ابتدا ماتریسی به

[۱] **حامد جعفری**: hamed.jafari@modares.ac.ir

www.SID.ir

Archive of SID

نام ماتریس فاصله را محاسبه کرده و سپس آن را در روش خوشه‌بندی سلسله مراتبی مورد استفاده قرار دادند. برای ارائه شاهد مثال‌هایی از ادامه چنین تلاش‌هایی در حوزه آمار شکل می‌توان به فعالیت سیرواستاوا و همکاران (۲۰۰۵) که در آن مروری کلی بر روش‌های خوشه‌بندی صورت گرفته است اشاره کرد. سالیان بعد، آمارال و همکاران (۲۰۱۰) درباره به‌کارگیری خوشه‌بندی برای تشخیص دو گونه ماهی از طریق شکل آن‌ها، ایشیهارا و همکاران (۲۰۱۱) برای خوشه‌بندی سلسله مراتبی چراغ جلوی خودروها و وینوِ و همکاران (۲۰۱۳) درباره خوشه‌بندی نقاط شاخص سه‌بعدی و به‌کارگیری آن در مسئله‌ای مربوط به صنعت نساجی مطالعه کردند.

در نوشتار حاضر ضمن نقد روش آمارال و همکاران (۲۰۱۰) در مورد خوشه‌بندی داده‌های مربوط به ریخت‌شناسی، روش آن‌ها طوری تطبیق داده می‌شود تا بتواند برای خوشه‌بندی داده‌هایی از این جنس، مناسب‌تر باشد. انتظار می رود روش پیشنهادی در مسائلی ازاین‌دست که علاوه بر هندسه با اندازه شکل اشیاء نیز مرتبط است، عملکرد بهتری داشته باشد. بر این اساس ابتدا در بخش ۲ اندکی درباره برخی مفاهیم مورد نیاز بحث می‌شود. سپس در بخش ۳ روش پیشنهادی مورد اشاره، مطرح می‌شود و در بخش ۴ عملکرد پیشنهاد مذکور در یک مثال واقعی مورد ارزیابی قرار می‌گیرد.

۲ مروری بر روش های خوشه‌بندی و تحلیل شکل در آمار

پیش از آن‌که به روش استفاده‌شده در این پژوهش پرداخته شود در این بخش اندکی درباره برخی مفاهیم مورداستفاده، یعنی «خوشه‌بندی» و «شکل» بحث می‌شود تا مبنای روش معرفی‌شده تا حد زیادی روشن‌تر شود. در حوزه روش‌های خوشه‌بندی چندمتغیره مطابق جانسون و ویچرن (۱۹۹۲) به یک اعتبار می‌توان گفت دو رهیافت عمومی وجود دارند؛ رهیافتی که داده‌ها را تولیدشده از توزیع احتمالی می‌داند و بنابراین هر خوشه شامل مشاهده‌های مشابهی است که از یک توزیع تولید شده‌اند. این رهیافت به روش «خوشه‌بندی مدل پایه [۱]» موسوم است. رهیافت دیگر بر پایه‌ی هیچ فرض توزیعی روی مشاهدات نیست و به این ترتیب ناپارامتری است. در بین روش‌های ناپارامتری نیز یکی از معروف‌ترین روش‌ها k-میانگین [۲] است که مبنای پژوهش حاضر بر آن استوار است. مبنای روش k-میانگین انتخاب k نقطه در دامنه مشاهدات به‌عنوان مرکز خوشه [۳] و سپس قرار دادن مرکز خوشه‌ها در کنار نزدیک‌ترین مشاهدات به آن‌ها، در یک خوشه است. مک کویین در سال ۱۹۶۷ برای اولین بار این روش را مطرح کرده است (مک‌کویین، ۱۹۶۷). سال‌ها بعد هارتیگان (هارتیگان، ۱۹۷۵) و در ادامه ایشان به همراه وانگ روشی را برای داده‌های اقلیدسی ارائه و تکمیل کردند (هارتیگان و وانگ، ۱۹۷۹) که عملکرد بهتری دارد و در منابع آماری به الگوریتم هارتیگان-وانگ [۴] (H-W) موسوم است. توجه پژوهش حاضر معطوف به تطبیق مراحل الگوریتم آن‌ها برای تحلیل داده‌های شکل بوده‌است. اما به دلیل محدودیت فضای مقاله از ارائه گام‌های این الگوریتم خودداری شده‌است. توصیه می‌شود خواننده علاقه‌مند مراحل الگوریتم را در منبع مورد اشاره مطالعه کند.

درباره شکل نیز باید گفت سابقه تحلیل آماری شکل به تلاش‌های کندال در نیمه دوم قرن بیستم باز می‌گردد (کندال، ۱۹۷۷) و کلیه‌ی محققینی که در این حوزه فعالیت می‌کنند، به‌نوعی وی را پایه‌گذار این شاخه از آمار می‌دانند.

در ادبیات آمارشکل به نقاط مشخص‌کننده شیء، «نقاط شاخص [۵]» می‌گویند. درایدن و ماردیا (۱۹۹۸) **تعریف** ۱.۲ را برای نقطه شاخص ارائه می‌دهند:

[۱] Model based clustering
[۲] K-Means
[۳] Centroid
[۴] Hartigan-Wong (H-W)
[۵] Landmarks

www.SID.ir

Archive of SID

تعریف ۱.۲. *نقطه‌ی شاخص، نقطه‌ای ایست که میان اعضای یک جامعه و همچنین بین اعضای جوامع مختلف انطباق ایجاد می‌کند.*

همچنین اگر فرض شود یک شیء با k نقطه شاخص m بعدی در اختیار است، آنگاه k نقطه شاخص را می‌توان در یک ماتریس با بعد $k \times m$ قرار داد. در این صورت به ماتریس حاصل، «ماتریس پیکربندی[۶]» گفته می‌شود. هرچند این ماتریس در ظاهر مانند یک ماتریس چند متغیره است، اما در بیشتر کاربردها چنان چه شکل جابه‌جا شود، بچرخد و یا تغییر اندازه دهد، باز هم اطلاعات موجود در آن، یک‌چیز را نشان می‌دهد؛ بنابراین مدل ریاضی شکل باید به‌گونه‌ای باشد که در طی این سه عامل ریاضی، اطلاعاتی از ویژگی هندسی شکل از بین نرود. بر همین اساس درایدن و ماردیا (۱۹۹۸) **تعریف ۲.۲** را از کندال (۱۹۷۷)، اقتباس کرده‌اند:

تعریف ۲.۲. *شکل یک‌چیز، تمام اطلاعات هندسی به‌جامانده از آن چیز بعد از حذف اثرات انتقال، دوران و مقیاس است.*

با توجه به مطالب گفته‌شده روشن است که منظور از اطلاعات خام اولیه هندسی در این‌جا نقاط شاخص هستند؛ بنابراین برای رسیدن به اطلاعات ثانویه هندسی لازم است از طریق تبدیلاتی، این اثرات از ماتریس پیکربندی حذف شود. به این تبدیلات، «تبدیلات همانندی[۷]» می‌گویند.

۳ تطبیق الگوریتم H-W برای خوشه‌بندی داده‌های شکل

از آنچه گفته شد به این نتیجه می‌توان رسید که نمی‌توان نرم اقلیدسی بین دو ماتریس پیکربندی را به‌عنوان فاصله دو مشاهده تعریف کرد چراکه ممکن است نقاط شاخص ثبت‌شده ، تبدیل یافته نقاط شاخص اصلی باشند که در حین نگارش به این شکل ثبت شده‌اند. بنابراین داده‌های شکل در فضای اقلیدسی قرار ندارند و فاصله مورداستفاده در الگوریتم H-W برای داده‌های شکل باید تغییر کند. همچنین به دلیل ارتباط میانگین و فاصله، میانگین مورد استفاده برای محاسبه مرکز خوشه‌ها نیز باید تغییر کند. آمارال و همکاران (۲۰۱۰) برای خوشه‌بندی دو گونه ماهی این تغییرات را صورت داده‌اند و نتیجه امر را برای چهار فاصله پروکراستس تام [۸] ، پروکراستس جزئی[۹]، فاصله ریمانی[۱۰] و فاصله در مختصات فضای مماس[۱۱] مورد مقایسه قرار دادند. به عنوان مثال، برای اعمال یک تطبیق مناسب میانگین شکل مورد نیاز نیز از طریق روش فرشه (۱۹۴۸) به دست آمده که نتیجه حاصل به میانگین فرشه [۱۲]

$$\hat{\mu} = \underset{z^{(m)} \in M}{\arg\min} \left[\frac{۱}{۲n} \sum_{j=۱}^{L} dist(z^{(m)}, z_j^{(m)})^۲ \right]$$

معروف است. توجه شود که در این‌جا $dist$، فاصله مورداستفاده (اقلیدسی یا نااقلیدسی) و z ماتریس پیکربندی است که اثر مکان و مقیاس از آن حذف شده‌اند. و بنابراین همه داده‌ها با مقیاس واحد در مبدأ قرار دارند و زبروند m نشان‌دهنده بعد نقاط شاخص است.

هرکدام از فاصله‌های ذکرشده نیز با در نظر گرفتن ملاحظاتی در فضای شکل تعریف می‌شود. برای آشنایی بیشتر با این فاصله‌ها و تفاوت‌هایشان به درایدن و ماردیا (۱۹۹۸) مراجعه شود ولی وجه اشتراک بین این فاصله‌ها این است که همه به‌نوعی سعی دارند اثرات هر سه تبدیل همانندی را از بین ببرند؛ یعنی، همه با مقیاس واحد در مکان واحد قرار دارند و طوری چرخانده می‌شوند تا بیشترین انطباق

[۶] Configuration Matrix
[۷] Symilarity Transformations
[۸] Full Procrustes
[۹] Partial Procrustes
[۱۰] Reimannian
[۱۱] Tangent Space Coordinates
[۱۲] Fréchet mean

www.SID.ir

Archive of SID

را بر هم داشته باشند. در برخی موارد لازم است اثر مقیاس حذف نشود چراکه به نظر می‌رسد مقیاس خود جزو ویژگی‌های متمایزکننده شکل‌ها از یکدیگر است. به نظر می‌رسد بسیاری از مسائل مربوط به ریخت‌شناسی مانند بررسی اسکلت جانداران از این دسته باشد. به همین دلیل این انتقاد به فعالیت آمارال و همکاران (۲۰۱۰) وارد می‌شود که "چرا از فاصله‌ای که با این رویکرد محاسبه می‌شود استفاده نکرده‌اند؟" درواقع شاید یکی از موارد تمایز شکل دو گونه مورد بررسی، اندازه اسکلت آن‌ها باشد. نتایج استفاده از این فاصله بر روی داده‌های به‌دست‌آمده از اسکلت سر تعدادی گوریل در بخش پنجم مبین این مطلب است که استفاده از این فاصله درباره داده‌هایی با این ویژگی بسیار می‌تواند بهتر عمل کند. این فاصله به صورت ۱.۳ تعریف می‌شود.

تعریف ۱.۳. *فاصله فضای اندازه-شکل بین دو ماتریس پیکربندی* $\mathbf{X}_1^\circ$ *و* $\mathbf{X}_2^\circ$ *که* X_1 *و* X_2 *تنها اثر مکان از آن‌ها حذف‌شده است، عبارت است از:*

$$d_S^2(X_1^\circ, X_2^\circ) = \inf_{\Gamma \in SO(m)} \|X_1 - X_2\Gamma\|^2$$

.

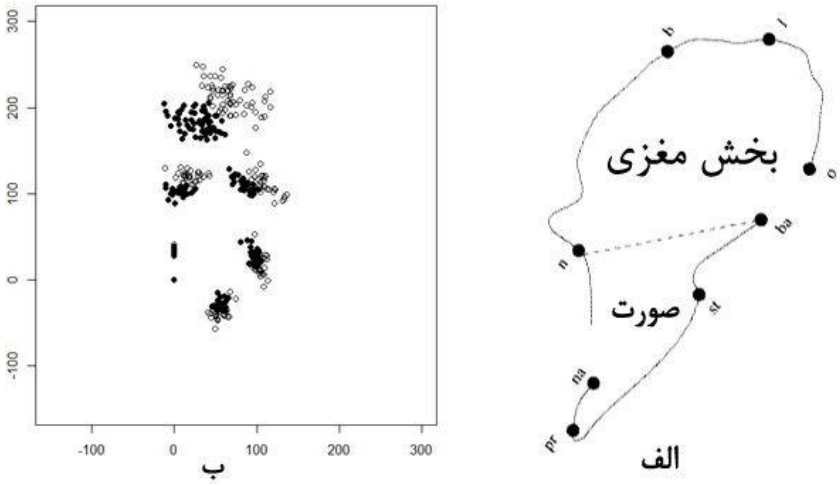
علاوه بر نقد به نوع فاصله مورد استفاده نحوه انتخاب نقطه شروع الگوریتم نیز چالش برانگیز است. ازآنجاکه در شروع تصادفی ممکن است دو نقطه انتخابی از یک خوشه باشند و درنتیجه الگوریتم باید نسبت به ماهیت داده‌ها از دقت کافی برخوردار باشد تا بتواند تفکیک مناسب را انجام دهد، یک راه ساده برای اطمینان از قرارگرفتن دو نقطه شروع در دو خوشه مجزا استفاده از دو نقطه با بیشترین فاصله است. به طور دقیق‌تر، فاصله‌ی دوبه‌دوی نقاط محاسبه شود و سپس دونقطه با بیشترین فاصله در یک خوشه قرار گیرند.

۴ تحلیل یک مثال واقعی

به‌منظور ارزیابی پیشنهادهای ارائه شده دراین مقاله و به ویژه تطبیق الگوریتم H-W براساس روش پیشنهادی آمارال و همکاران (۲۰۱۰)، از داده‌های جمجمه گوریل‌ها استفاده شده است. این داده‌ها در پی یک تحقیق به‌منظور بررسی شباهت یا عدم شباهت جمجمه گوریل‌های نر و ماده به‌صورت ۸ نقطه شاخص آناتومیک از اسکلت سر ۵۹ گوریل شامل ۲۹ نر و ۳۰ ماده مطابق شکل ۱ (الف) جمع‌آوری شده‌اند که اولین بار اُهیگینس (۱۹۹۸) به مطالعه آن‌ها پرداخته است. مختصات نقاط این ۵۹ گوریل در دو بعد نیز در شکل ۱ (ب) نشان داده شده است. همانطور که مشاهده می‌شود اندکی دو گروه با هم متفاوت اند. اگرچه این فرض پذیرفته شده است که داده‌ها تفکیک شده‌اند، هدف خوشه‌بندی این داده‌ها به دو گروه نر و ماده است. چنانچه فرض شود که جنسیت گوریل‌ها با شکل آن‌ها ارتباط معنی‌دار دارد، می‌توان با خوشه‌بندی آن‌ها از روی شکل، داده‌ها را به دو گروه نر و ماده تقسیم کرد. البته ازآنجاکه تا به امروز هنوز آزمون مورد قبولی برای فرضیه‌هایی این‌چنین- ارتباط بین متغیر شکل و متغیر تبیینی- ارائه نشده است، امکان آزمون کردن این فرضیه وجود ندارد. بنابراین یک حالتی که ممکن است رخ دهد این است که مثلاً شکل اسکلت یک یا چند گوریل نر شبیه اسکلت‌های گروه ماده باشد و یا بالعکس. می‌توان پیش از هر چیز این فرضیه را مورد آزمون قرار داد که "آیا اساساً دو گروه گوریل‌های نر و ماده به لحاظ شکلی نیز دو گروه هستند یا خیر؟". این فرضیه از طریق آزمون اختلاف میانگین شکل معرفی‌شده در آمارال و همکاران (۲۰۰۷) قابل انجام است. برای هدف این مقاله، چهار آزمون مرسوم در آمار شکل انجام و نتایج آن‌ها به صورت جدول ۱ گزارش شده است. همانطورکه ملاحظه می‌شود با توجه به نتایج حاصل و به لحاظ شکل دو گروه مجزا وجود دارد. از آن‌جا که معیار عدم شباهت و اصولاً اساس خوشه‌بندی k-میانگین بر اساس فاصله است، می‌توان از طریق فاصله این موضوع را مورد بررسی شهودی و تصویری قرار داد. برای این امر به این صورت عمل می‌کنیم :

www.SID.ir

Archive of SID



شکل ۱: ۸ نقطه شاخص آناتومیک در جمجمه گوریل (الف) و نقاط شاخص گوریل‌های نر(نقاط خالی) و گوریل های ماده (نقاط پر) در کنار هم (ب).

جدول ۱: نتایج آزمون‌های مقایسه میانگین برای گوریل‌های نرو ماده

معیار	لامبدا	هتلینگ	جیمز	گودال
مقدار آماره	۲۴۹٫۱۸۴۶	۲۶٫۴۷۰۴۲	۳۹۱٫۸۵۷۷	۲۲٫۲۸۵۷۱
p-مقدار	۰٫۰۰۰۰	$۱٫۱۱۰۲e-۱۶$	۰٫۰۰۰۰	۰٫۰۰۰۰
p-مقدار(بازنمونه‌گیری)	۰٫۰۰۵	۰٫۰۰۵	۰٫۰۰۵	۰٫۰۰۵

- ابتدا به گوریل‌های نر برچسب‌های ۱ تا ۲۹ و به گوریل‌های ماده برچسب‌های ۳۰ تا ۵۹ را می‌دهیم.
- سپس عمل خوشه‌بندی بر اساس روش آمارال و همکاران (۲۰۱۰) برای پنج فاصله و نقطه شروع گفته‌شده به دست می‌آوریم.
- پس از آن یکی از دونقطه شروعی که بیشترین فاصله را دارند به‌عنوان مبدأ انتخاب می‌کنیم و چهار فاصله ریمانی، پروکراستس تام، پروکراستس جزئی و تقریب فضای مماس بقیه داده‌ها را از آن می‌سنجیم. بدیهی است که در این حالت باید داده‌ها به هم‌جنس خودشان نزدیک‌تر باشند. برای مثال اگر یک گوریل نر به‌عنوان مبدأ انتخاب شده است باید گوریل‌های نر به آن نزدیک‌تر باشند و گوریل‌های ماده دورتر. اگر غیر از این باشد آن‌گاه استفاده از معیار فاصله برای خوشه‌بندی مورد اشکال خواهد بود.

نتیجه عملیات بند اخیر و همچنین خوشه‌بندی طبق روش آمارال و همکاران (۲۰۱۰) برای پنج فاصله در قسمت‌های الف تا ث شکل ۲ نشان داده‌شده است. این شکل‌ها نشان می‌دهد که در معیارهای متفاوت فاصله موقعیت مشاهدات نسبت به هم چگونه است. می‌توان ملاحظه کرد که در این مطالعه معیار فاصله اندازه–شکل بهتر از بقیه فواصل عمل کرده است. این نتیجه از جدول ۲ نیز قابل استنتاج است. این فاصله در فضای اندازه شکل به‌دست آمده و ظاهراً با کمترین تداخل شکل‌ها را از یکدیگر جدا کرده است. ممکن است دلیل آن

www.SID.ir

Archive of SID

این باشد که در مطالعه گوریل‌ها بزرگی و کوچکی جمجمه هم مهم است. به‌بیان‌دیگر رابطه معنی‌داری بین اندازه جمجمه و جنسیت وجود دارد و شاید بتوان گفت رابطه معنی‌داری بین شکل و اندازه از یک‌طرف و جنسیت از طرف دیگر وجود دارد. نکته‌ی جالب توجه دیگر در این مطالعه فاصله نسبتاً زیاد بین دو گروه است و شاید همین موضوع کمک کرده تا دو شکل بهتر تفکیک شوند. مطلب دیگری که از این شکل استنتاج می‌شود این است که نقطه ۱۱ که در فاصله‌های قبلی اشتباه تشخیص داده‌شده بود، نه‌تنها در به‌کارگیری فاصله جدید درست تشخیص داده‌شده، بلکه با مرز تفکیک کننده دو گروه هم فاصله نسبی دارد. این یعنی استفاده از فاصله نامناسب می‌تواند منجر به خطاهای فاحش شود . البته از طرف دیگر نقاط ۵۷ و ۵۴ نزدیک به مرزند. بنابراین ممکن است تشخیص اشتباه در فواصل دیگر در حقیقت به دلیل اشتباه انتخاب فاصله باشد نه این‌که واقعاً نقطه ۱۱ شباهت بیشتری به گروه مخالف خود، گروه ماده، داشته باشد. این امر در موارد دیگر نیز ممکن است اتفاق بیفتد چراکه استفاده از فاصله اقلیدسی در فضای نا اقلیدسی ممکن است دو شیء نزدیک را دور و دو شیء دور را نزدیک هم قرار دهد.

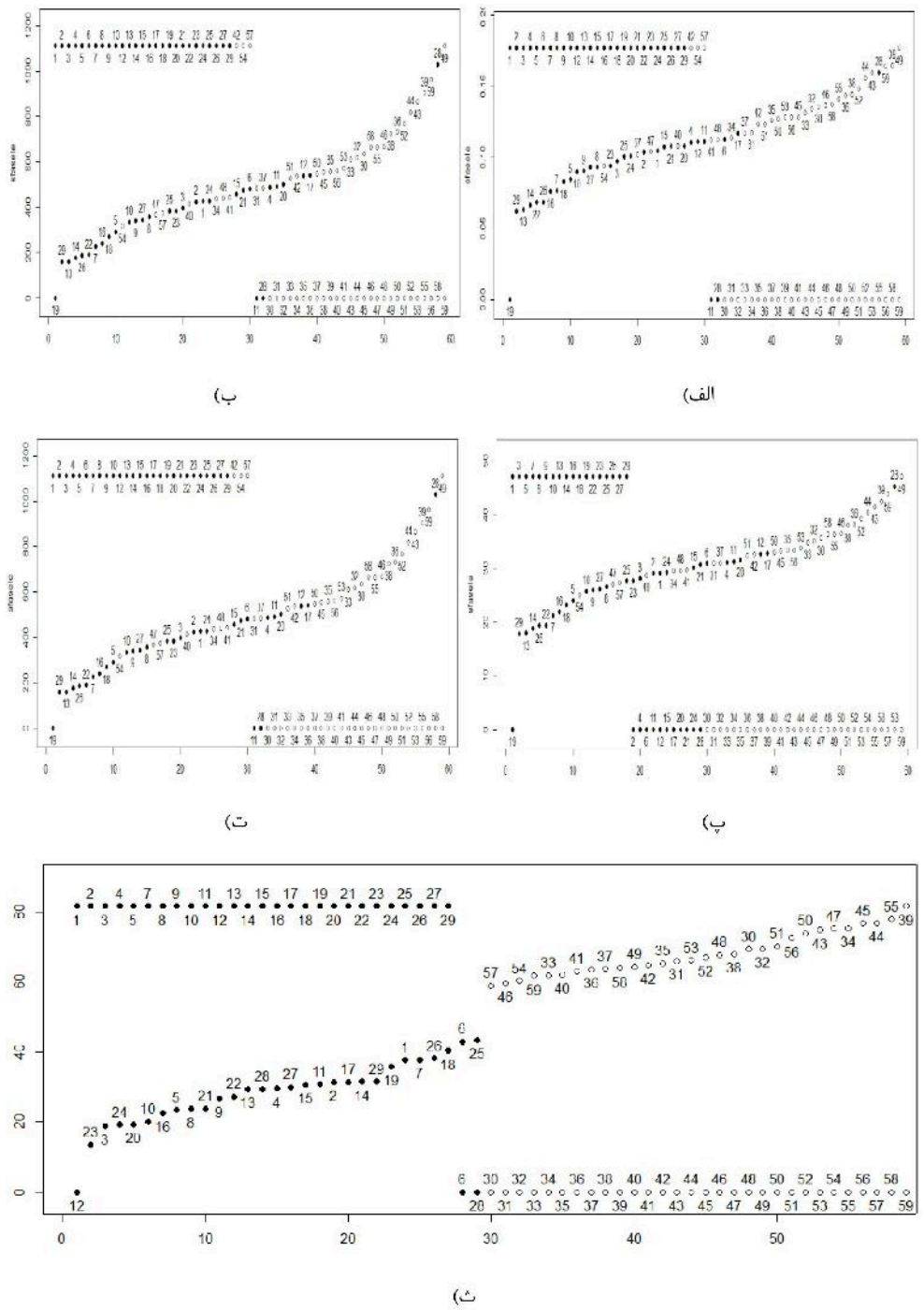
نقاط ابهامی که در تحلیل این مثال مطرح شدند خود مبانی نظری در آمارشکل دارند که نویسندگان مقاله حاضر علاقه‌مند به مطالعه آن‌ها از هر دو جنبه نظری و کاربردی هستند. با این حال، نتیجه تطبیق الگوریتم H-W برای داده‌های شکل با به کارگیری فاصله اندازه-شکل بسیار رضایت بخش بوده‌است.

جدول ۲: جدول درصد تشخیص درست الگوریتم به ازای هر روش به تفکیک جنسیت

نام روش/جنسیت	مذکر	مؤنث
پروکراستس تام	۹۳/۱۰	۹۰/۰۰
پروکراستس جزئی	۹۳/۱۰	۹۰/۰۰
ریمانی	۹۳/۱۰	۹۰/۰۰
مختصات فضای مماس	۶۲/۰۷	۱۰۰
اندازه-شکل	۹۳/۱۰	۱۰۰

www.SID.ir

Archive of SID



شکل ۲: نمودار موقعیت نسبی مشاهدات برحسب فاصله از هم به همراه خوشه‌بندی برای پنج فاصله الف)ریمانی ب)پروکراستس جزئی پ)فضای مماس ت)پروکراستس تام ث)اندازه-شکل

www.SID.ir

Archive of SID

مراجع

[1] Amaral, G. A., Dryden, I. L., and Wood, A. T. A. (2007). Pivotal Bootstrap Methods for k-sample Problems in Directional Statistics and Shape Analysis, *Journal of the American Statistical Association*, **102**, 695-707.

[2] Amaral, G. J., Dore, L. H., Lessa, R. P., and Stosic, B. (2010). K-Means Algorithm in Statistical Shape Analysis, *Communications in Statistics—Simulation and Computation*, **39**, 1016-1026.

[3] Dryden, I. L., and Mardia, K. V. (1998). *Statistical Shape Analysis*, Chichester: Wiley.

[4] Fréchet, M. (1948). les Éléments Aléatoires de Nature Quelconque dans un Espace Distancié. In *Annales de l'Institut Henri Poincaré*, **10**, pp. 215-310.

[5] Hartigan, J. A. (1975). *Clustering Algorithms*,New york: John Wiley and Sons.

[6] Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: A k-means Clustering Algorithm, *Applied Statistics*, 100-108.

[7] Ishihara, S., Ishihara, K., and Nagamachi, M. (2011). Statistical Shape Analysis of Headlights, In *Proceedings of International Conference on Biometrics and Kansei Engineering*, pp. 27-32.

[8] Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*, New Jersey: Prentice Hall.

[9] Kendall, D. G. (1977). The Diffusion of Shape, *Advances in Applied Probability*, 428-430.

[10] Lele, S. R., and Richtsmeier, J. T. (2001). *An Invariant Approach to Statistical Analysis of Shapes*, Boca Raton: Chapman and Hall.

[11] MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*,**1**, 281-297.

[12] O'Higgins, P. (1989). *A Morphometric Study of Cranial Shape in the Hominoidea*, (Doctoral dissertation, University of Leeds).

[13] Srivastava, A., Joshi, S. H., Mio, W., and Liu, X. (2005). Statistical Shape Analysis: Clustering, Learning, and Testing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 590-602.

[14] Vinué, G., Simó, A., and Alemany, S. (2013). The K-means Algorithm for 3D Shapes with an Application to Apparel Design, *Advances in Data Analysis and Classification*, DOI: 10.1007/s11634 - 014 - 0187 - 1.

www.SID.ir