نشریه علمی - پژوهشی علوم و فنون نقشه برداری، دوره هشتم، شماره ۴، خرداد ماه ۱۳۹۸

ارزیابی عملکرد الگوریتم‌های خوشه‌بندی در استخراج خطوط سیر مکانی متشابه

علی مویدی[۱]، رحیم علی‌عباسپور*[۲]، علیرضا چهرقان[۳]

[۱] دانشجوی کارشناسی ارشد سیستم‌های اطلاعات مکانی - دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی - پردیس دانشکده‌های فنی - دانشگاه تهران
alimoayedi2013@ut.ac.ir

[۲] استادیار دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی - پردیس دانشکده‌های فنی - دانشگاه تهران
abaspour@ut.ac.ir

[۳] استادیار دانشکده مهندسی معدن - دانشگاه صنعتی سهند
chehreghan@sut.ac.ir

(تاریخ دریافت تیر ۱۳۹۷، تاریخ تصویب دی ۱۳۹۷)

چکیده

در سالهای اخیر، رشد بالا و روزافزون داده‌های خطوط‌سیر مکانی و لزوم پردازش و استخراج اطلاعات مفید و الگوهای معنی‌دار از آن‌ها منجر به جلب توجه محققان بسیاری در زمینه خوشه‌بندی خطوط‌سیر مکانی-زمانی شده‌است. تاکنون توابع شباهت و الگوریتم‌های خوشه بندی مختلفی برای طبقه‌بندی خطوط‌سیر ارائه شده‌اند. گستردگی الگوریتم‌های خوشه‌بندی و نتایج منحصربه‌فرد هر یک بر لزوم توجه و بررسی نقاط ضعف و قوت آن‌ها تاکید می‌کند. در این تحقیق، الگوریتم‌های خوشه‌بندی در خطوط‌سیر مکانی که تعمیم یافته از الگوریتم-های خوشه‌بندی داده‌های نقطه‌ای هستند به چهار دسته‌ی کلی روش‌های افرازی، سلسله‌مراتبی، چگالی مبنا و مبتنی بر بهینه‌سازی تقسیم شدند و پرکاربردترین الگوریتم‌ها در هر دسته پیاده‌سازی و مورد ارزیابی قرار گرفتند. فرایند ارزیابی بر روی دو مجموعه داده با پیچیدگی متفاوت و در سه حالت بدون خطا، خطا با توزیع گوسین و وجود داده پرت انجام گرفته تا توانایی روش‌ها در شرایط مختلف بررسی گردد. از شاخص سیلووت و زمان محاسباتی به عنوان دو پارامتر برای مقایسه و ارزیابی استفاده شده است. با توجه به نتایج به‌دست‌آمده توجه به داده و ویژگی‌های آن در انتخاب روش مناسب خوشه‌بندی حائز اهمیت است. با این حال در مجموع بهترین نتایج از لحاظ کیفیت خوشه‌بندی به ترتیب از دسته‌های مبتنی بر بهینه‌سازی، افرازی، سلسله مراتبی و چگالی مبنا و از لحاظ سرعت محاسبات به ترتیب دسته‌های چگالی‌مبنا، سلسله‌مراتبی، افرازی و مبتنی بر بهینه‌سازی حاصل شده است. دسته افرازی (صرفا زیر دسته طیفی) بالاترین مقاومت در برابر داده پرت و روش‌های چگالی مبنا و مبتنی بر بهینه‌سازی بالاترین مقاوت در برابر نویز را از خود نشان داده‌اند.

واژگان کلیدی: خطوط‌سیر مکانی، خوشه‌بندی، شاخص سیلووت، زمان محاسباتی

* نویسنده رابط

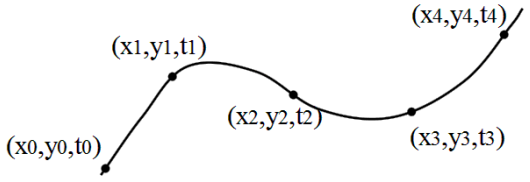
۱- مقدمه

در سالهای اخیر و به‌ویژه در دهه گذشته، با گسترش روند ثبت موقعیت اجسام متحرک نظیر وسایط نقلیه، افراد و حتی حیوانات و به دلیل توسعه سامانه‌های تعیین موقعیت و گسترش استفاده از سرویس‌های مکان‌مبنا حجم بالایی از داده‌های خطوط‌سیر مکانی-زمانی در حال تولید هستند. پردازش و تحلیل این خطوط‌سیر منجر به استخراج اطلاعات سودمندی برای حل چالش‌های متعددی مانند مدیریت ترافیک [۱]، حمل‌ونقل هوشمند [۲]، نظارت و امنیت [۳] و مطالعات زیست‌شناسی [۴] شده است.

خط‌سیر مکانی نوع خاصی از سری زمانی است که در آن، دنباله‌ای از موقعیت‌ها به همراه زمان برداشت آن ثبت می‌شود. متداول‌ترین روش تولید این نوع از داده‌ها، برداشت اطلاعات مکانی به واسطه وسیله نقلیه مجهز به GPS یا فرد حامل گوشی هوشمند است. در حالت کلی خط سیر T را می‌توان به صورت رابطه (۱) نشان داد.

$$T: ((x_1, y_1, t_1), \dots, (x_n, y_n, t_n)) \qquad (۱)$$

که در این رابطه (x_i, y_i) مختصات برداشت‌شده در زمان t_i و n طول خط‌سیر را نشان می‌دهد (شکل ۱).



شکل۱- خط سیر T تعریف شده در رابطه (۱)

خوشه‌بندی از مهمترین ابزارها برای استخراج الگو و اطلاعات سودمند از خطوط‌سیر است که تحقیقات زیادی را به خود اختصاص داده است. خوشه‌بندی فرایند گروه‌بندی مجموعه داده به زیرکلاس‌های معنی‌دار است که هر یک از این زیرکلاس‌ها خوشه نامیده می‌شود. در خوشه‌بندی اشیا موجود در یک خوشه با یکدیگر کمترین تفاوت و با اشیا خوشه‌های دیگر کمترین شباهت را دارند [۵]. الگوریتم‌های خوشه‌بندی زیادی در تحقیقات پیشین ارائه و برای آن‌ها دسته‌بندی‌های مختلفی پیشنهاد شده است. در یکی از این دسته‌بندی‌ها، الگوریتم‌های خوشه‌بندی به پنج دسته شامل افرازی، سلسله مراتبی[۱]، گرید مبنا، مدل مبنا و چگالی مبنا[۲] تقسیم‌بندی شده‌است [۶]. خوشه‌بندی از مهم‌ترین روش‌ها برای استخراج الگو از خطوط‌سیر، کاهش حجم آن‌ها، کشف داده‌های پرت در خطوط‌سیر، شاخص‌گذاری و نمایش بصری ساده آن‌ها است. در خوشه‌بندی مکانی-زمانی، اشیا بر اساس اطلاعات موقعیت و زمان رخداد آن‌ها دسته‌بندی می‌شوند. این نوع خوشه‌بندی در علوم مکانی به علت گستردگی ابزارهای اندازه‌گیری موقعیت و تحلیل‌های زمانی- مکانی موردنظر این علم بسیار موردتوجه است.

تاکنون روش‌های مختلفی برای خوشه‌بندی خطوط-سیر ارائه شده‌است. درک ویژگی‌های این روش‌ها و نقاط قوت و ضعف آن‌ها می‌تواند در تصمیم‌گیری برای انتخاب روش مناسب سودمند باشد. بعضی از روش‌های خوشه‌بندی فقط در مجموعه داده با حجم کم نتیجه‌بخش هستند؛ دسته‌ای از توابع خوشه‌بندی فقط قادر به استخراج خوشه‌ها با شکل محدب هستند، ولی بعضی دیگر، خوشه با هر نوع شکل را استخراج می‌کنند؛ برخی توابع خوشه‌بندی نیازمند تعیین پارامتر اولیه مثل تعداد خوشه‌ها از طرف کاربر هستند و دسته دیگر به این پارامترها نیاز ندارند؛ و تشخیص داده پرت در همه الگوریتم‌های خوشه‌بندی امکان‌پذیر نیست. از اینرو، مقایسه و ارزیابی روش‌های خوشه‌بندی خطوط‌سیر از اهمیت بالایی برخوردار است. در این تحقیق دسته‌بندی برای روش‌های خوشه‌بندی خطوط‌سیر که تعمیم‌یافته از خوشه‌بندی داده‌های نقطه‌ای هستند ارائه شده و مهمترین و پرکاربردترین الگوریتم در هر دسته بر روی دو مجموعه داده با ویژگی حرکتی متفاوت و در حضور نویز و داده پرت پیاده‌سازی شده‌اند تا مزایا و معایب هر روش با توجه به میزان نویز و ویژگی‌های حرکتی داده مشخص گردد.

در ادامه این تحقیق در بخش دو مروری بر تحقیقات عمده در زمینه خوشه‌بندی داده‌های خط‌سیر انجام گرفته است. در بخش سه گام‌های اصلی صورت‌گرفته در این تحقیق شامل پیش‌پردازش، انتخاب پارامترهای اولیه هر الگوریتم، انتخاب تابع شباهت، تعیین الگوریتم خوشه‌بندی و در نهایت ارزیابی به صورت نظری تشریح می‌شود. در

۱ Hierarchical
۲ Density-based

بخش چهار پیاده‌سازی الگوریتم‌ها و مقایسه آن‌ها و در بخش آخر هم نتیجه‌گیری از تحقیق انجام می‌شود.

۲- مروری بر تحقیقات پیشین

تحقیقات صورت‌گرفته در زمینه خوشه‌بندی خطوط سیر که تعمیم خوشه‌بندی داده‌های نقطه‌ای هستند را می‌توان به چهار دسته کلی افرازی، سلسله مراتبی، چگالی مبنا و مبتنی بر بهینه‌سازی تقسیم‌بندی کرد. روش‌های K-means و طیفی دو الگوریتم مهم در دسته روش‌های افرازی هستند که بیشتر از سایر روش‌های این دسته در تحقیقات گذشته به آن‌ها اشاره شده است. Tork در [۷] از اولین افرادی بود که خوشه‌بندی برای داده‌های مکانی-زمانی با استفاده از K-means را توسعه داد. Nanni در [۸] دو روش خوشه‌بندی K-means و سلسله‌مراتبی را برای خطوط‌سیر تعمیم داد. Vlachos و همکاران نیز نسخه‌ای از K-means را برای خوشه‌بندی سری‌های زمانی بر اساس تقریب اولیه داده خام به وسیله طول‌موج ارائه کردند [۹]. برای خوشه‌بندی سری‌های زمانی با K-means معمولاً فاصله اقلیدسی به‌عنوان تابع شباهت در نظر گرفته شده است. با این‌حال فاصله DTW[۱] که در اندازه‌گیری شباهت سری‌های زمانی و خطوط‌سیر نتیجه بهتری از فاصله اقلیدسی می‌دهد را نیز می‌توان در K-means به کار برد. برای اولین بار در [۱۰] از DTW در K-means استفاده و عملکرد خوبی نیز حاصل شد. Atev و همکاران [۱۱] با بهره‌گیری از خوشه‌بندی طیفی دیگر زیر مجموعه دسته افرازی و تابع شباهتی بر اساس فاصله هاسدورف[۲] چارچوبی برای خوشه‌بندی خطوط‌سیر ارائه دادند. همچنین Atev و همکاران در [۱۲] به استخراج الگو از خطوط‌سیر ترافیک با استفاده از خوشه‌بندی طیفی پرداختند. دومین دسته پرکاربرد در خوشه‌بندی خطوط-سیر الگوریتم‌های سلسله‌مراتبی هستند. Fu و همکاران در [۱۳] با استفاده از خوشه‌بندی سلسله‌مراتبی و طیفی چارچوبی برای شناسایی آنامولی و طبقه‌بندی خطوط سیر وسایط نقلیه در ترافیک پیشنهاد دادند. Rodrigues و همکاران در [۱۴] چارچوبی برای خوشه‌بندی سلسله-مراتبی سری‌های زمانی پیشنهاد کردند. از بین دسته‌های اشاره‌شده الگوریتم‌های چگالی‌مبنا را می‌توان پراستفاده‌ترین آن‌ها نامید. T-OPTICS [۱۵] بسطی از OPTICS[۳] و ST-DBSCAN [۱۶] و Tra-DBScan [۱۷] تعمیم‌هایی از DBSCAN[۴] هستند که برای خطوط‌سیر مناسب‌سازی شده‌اند. Palma و همکاران در [۱۸] به کشف نقاط موردعلاقه با استفاده از خوشه‌بندی چگالی مبنا پرداختند. Nanni و همکاران [۱۵]، Lee و همکاران [۱۹] و Akasapu و همکاران [۲۰] از دیگر افرادی بودند که هریک به نوعی از خوشه‌بندی چگالی مبنا برای گروه‌بندی خطوط‌سیر استفاده کردند. آخرین دسته روش‌های بر مبنای بهینه‌سازی هستند. Ahmadyfard و Modares با ترکیب K-means و PSO و Lu و همکاران با ترکیب K-means و الگوریتم ژنتیک روش ترکیبی جدید برای خوشه‌بندی داده‌ها ارائه و بررسی آن‌ها برتری این روش را بر K-means اثبات کرد [۲۱, ۲۲]. Izakian و Mesgari [۲۳] با بهره‌گیری از PSO روشی برای خوشه‌بندی سری‌های زمانی پیشنهاد کردند. همچنین Izakian و همکاران در [۲۴] روشی خودکار برای خوشه‌بندی خطوط‌سیر با استفاده از PSO ارائه کردند.

در سنجش کارایی توابع شباهت مختلف و مقایسه آن‌ها تحقیقاتی توسط Zhang و همکاران [۲۵]، Wang و همکاران [۲۶] و Bailer [۲۷] صورت گرفته، اما کمتر مقایسه‌ای بین عملکرد دسته‌های مختلف روش‌های خوشه‌بندی در خوشه-بندی خطوط‌سیر مکانی به چشم می‌خورد. Morris و Trivedi در [۲۸] به مقایسه توابع خوشه‌بندی با روش‌های مستقیم، سلسله‌مراتبی، ترکیبی، مبتنی بر گراف و طیفی پرداخته‌اند. اما تأثیر عوامل مهمی همچون نویز و داده پرت در کارایی و عملکرد توابع موردبررسی قرار نگرفته است. Atev و همکاران در [۱۱] فقط مقایسه‌ای بین دو گروه خوشه‌بندی طیفی و سلسله‌مراتبی انجام دادند.

در این مقاله تلاش شده کاستی‌های موجود در تحقیقات گذشته پوشش داده شود و با دسته‌بندی روش-های مختلف خوشه‌بندی مقایسه‌ای بین آن‌ها صورت گیرد و نقاط ضعف و قوت هر دسته مشخص شود. استفاده از مجموعه داده با پارامترهای حرکتی و پیچیدگی مختلف، بررسی تاثیر وجود نویز و داده پرت با میزان مختلف،

۱ Dynamic time warping
۲ Hausdorff

۳ Ordering points to identify the clustering structure
۴ Density-based spatial clustering of applications with noise

استفاده از شاخص ارزیابی کیفیت خوشه‌بندی برای بررسی کارایی توابع شباهت و مقایسه سرعت توابع شباهت و بررسی تاثیر افزایش حجم داده در زمان خوشه‌بندی با توابع شباهت مختلف از نقاط قوت این تحقیق نسبت به تحقیق های پیشین است.

۳- خوشه‌بندی خطوط‌سیر مکانی

اکثر روش‌های خوشه‌بندی به‌کارگرفته‌شده برای خطوط-سیر تعمیمی از روش‌های خوشه‌بندی مورد استفاده برای داده‌های نقطه‌ای با استفاده از تعریف تابع شباهت مناسب برای خطوط‌سیر است. K-means، BIRCH، DBSCAN، OPTICS و STING ازجمله الگوریتم‌هایی هستند که برای خوشه‌بندی خطوط‌سیر تعمیم یافته‌اند. در این تحقیق روش‌های خوشه‌بندی خطوط‌سیر به دسته‌های افرازی، سلسله مراتبی، چگالی مبنا و مبتنی بر بهینه‌سازی تقسیم‌بندی شده‌اند که هر یک در ادامه تشریح می‌شوند. روند کلی انجام مقایسه دسته‌های مهم خوشه‌بندی در این مقاله در شکل (۲) دیده می‌شود. در اولین مرحله، پیش-پردازش اولیه بر روی داده‌ها انجام می‌گیرد. در گام بعد پارامترهای اولیه برای هر روش خوشه‌بندی تعیین می‌شود. گام سوم مربوط به محاسبه شباهت بین همه جفت داده‌های خط‌سیر است. در گام چهارم معروف‌ترین الگوریتم(ها) برای هر دسته انتخاب و خوشه‌بندی با آن(ها) انجام می‌گیرد. در آخرین مرحله عملکرد روش‌های مختلف مورد ارزیابی قرار می‌گیرد. هر یک از این مراحل در ادامه تشریح می‌شود.

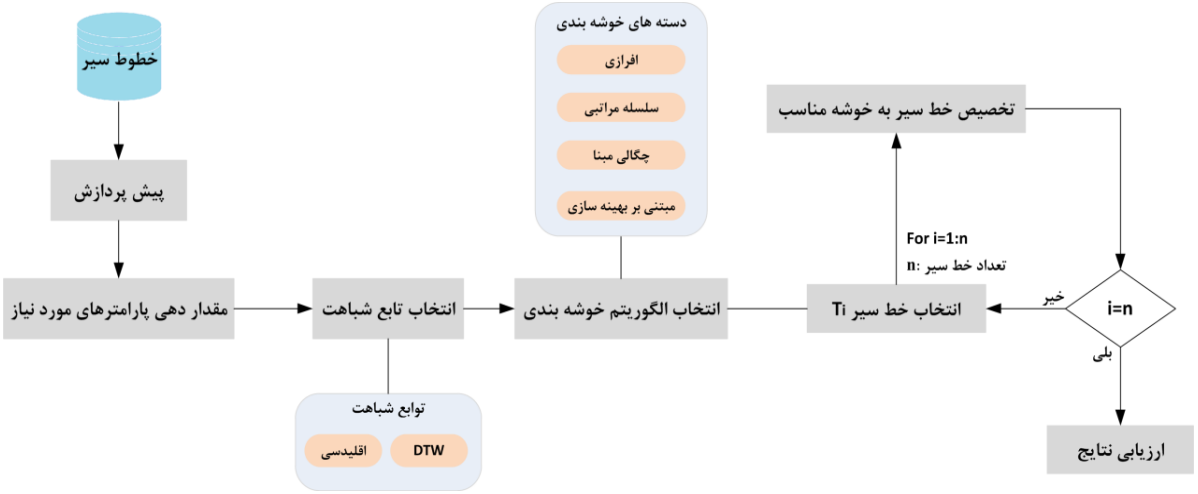
۳-۱- پیش‌پردازش

برای محاسبه شباهت بین دو خط‌سیر با استفاده از تابع اقلیدسی، لازم است تا تعداد نقاط آن‌ها برابر باشند. برای خطوط‌سیر T_1 و T_2 با تعداد نقاط به ترتیب p و q به نحوی که $q < p$ نقطهٔ $T_2(i)$ در خط‌سیر T_2 معادل نقاط در بازهٔ $(i-1)\times p/q$ تا $i\times p/q$ در خط سیر T_1 می‌باشد، بنابراین فاصله $T_2(i)$ تا نقطه معادل آن در خط سیر T_1 از رابطه (۲) به دست می‌آید.

$$Dist = (dist_1 + dist_2)/2 \qquad (۲)$$

که در این رابطه $dist_1$ و $dist_2$ فاصله اقلیدسی بین نقطه i ام در T_2 به ترتیب تا نقاط round((i-1)p/q) ام و round(i×p/q) ام در T_1 هستند. در نتیجه شباهت دو خط‌سیر از رابطه (۳) به دست می‌آید.

$$D_{ec} = \left(\sum_{i=1}^{q} Dist(i)\right)\Big/ q \qquad (۳)$$



شکل ۲- فلوچارت کلی مقایسه دسته های مختلف خوشه بندی خطوط سیر

۳-۲- مقدار دهی پارامترهای اولیه

برای مقایسه صحیح‌تر بین الگوریتم‌ها نیازمند انتخاب دقیق پارامترهای هر الگوریتم هستیم. دست‌یابی به بهترین نتیجه بر اساس معیار ارزیابی کیفیت خوشه‌ها، حصول تعداد خوشه مورد نظر و ثبات در اعضای خوشه‌ها از معیارهای مهم در انتخاب پارامتر های اولیه است.

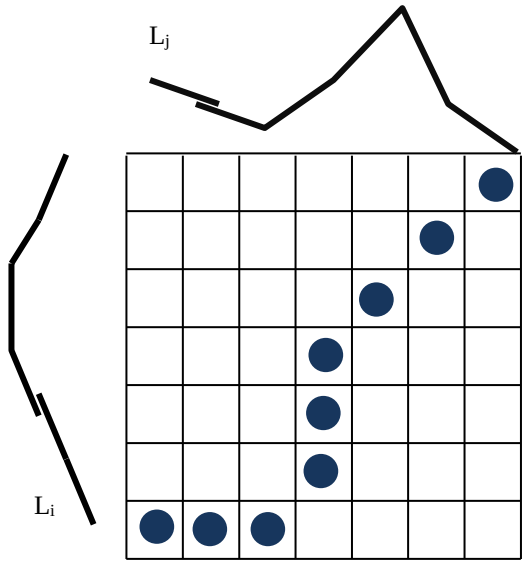
۳-۳- توابع شباهت

قبل از خوشه‌بندی خطوط‌سیر باید شباهت (یا فاصله آن‌ها) محاسبه گردد. با توجه به پیچیدگی، ابعاد بالا و خاصیت دنباله‌ای داده‌های خط‌سیر محاسبه شباهت در آن‌ها نسبت به داده‌های نقطه‌ای چالش برانگیزتر است. علاوه بر این برای محاسبه شباهت در خطوط‌سیر توجه به ماهیت داده و همچنین کاربرد مورد نظر بسیار مهم است. توابع مختلفی برای شباهت‌سنجی خطوط‌سیر پیشنهاد شده است که هر یک ویژگی‌های منحصربه‌فردی دارند. تابع اقلیدسی از قدیمی‌ترین و البته ساده‌ترین و سریع‌ترین این توابع است. تابع هاسدورف دیگر تابع شباهت است که مبنای آن شکل هندسی دو خط‌سیر است. DTW از پرکاربردترین توابع شباهت است که مستقل از انتقال زمان محلی و کشیدگی زمانی است. بر خلاف تابع اقلیدسی، الزام به برابر بودن طول دو خط‌سیر برای محاسبه شباهت در این تابع وجود ندارد. برای مقابله با تأثیر نویز در محاسبه شباهت خطوط‌سیر توابع LCSS[۱] و EDR[۲] معرفی شدند. ERP[۳] از دیگر توابع مهم شباهت است که از ترکیب دو تابع L1 و تصحیح فاصله و برای بهره‌گیری از مزایای هر دو روش پیشنهاد شده است.

در این تحقیق از تابع فاصله اقلیدسی به دلیل پیاده سازی ساده و سرعت بالا و تابع شباهت DTW به علت نتایج قابل قبول آن در تحقیقات پیشین استفاده شده است. فاصله اقلیدسی بین دو خط‌سیر L_i و L_j با تعداد نقاط برابر و مساوی n از رابطه (۴) به دست می‌آید که در این رابطه a_k^m بعد m ام از نقطه k ام L_i و b_k^m بعد m ام از نقطه k ام L_j است. در حالت دو بعدی p=2 است و a_k^1 و a_k^2 به ترتیب معادل x و y نقطه k ام خط سیر L_i است.

$$D_E(L_i, L_j) = \frac{1}{n}\sum_{k=1}^{n}\sqrt{\sum_{m=1}^{p}(a_k^m - b_k^m)^2} \qquad (۴)$$

DTW یا پیچش زمانی پویا از الگوریتم‌های محاسبه فاصله در سری‌های زمانی از طریق یافتن کوتاه ترین مسیر بین دو دنباله با پیچ‌وتاب در آن‌ها است. دو خط سیر را می توان در دو طرف یک شبکه گریدی مطابق شکل(۳) قرار داد به نحوی که داخل هر سلول فاصله عناصر متناظر دو توالی قرار گیرد. برای پیدا کردن بهترین تطابق یا هماهنگی بین این دو دنباله، باید مسیری را از طریق این شبکه گریدی یافت (ضمن در نظر گرفتن یک سری قیود محدود کننده) که فاصله کل بین آنها را به حداقل برسد.



شکل ۳- محاسبه فاصله DTW

DTW را به صورت ریاضی برای دو خط سیر L_i و L_j به ترتیب با تعداد نقاط n و m میتوان مطابق رابطه (۵) نشان داد. در رابطه $R(L_i)$ و $R(L_j)$ به ترتیب خطوط‌سیر باقی مانده از L_i و L_j بعد از حذف نقاط ابتدایی آن‌ها است.

$$D_{DTW}(L_i, L_j) = \begin{cases} 0 & m = n = 0 \\ \infty & m = 0 \,||\, n = 0 \\ \text{euclidean}(a_i^k, b_j^k) + & \text{other} \\ \min\begin{Bmatrix} D_{DTW}(R(L_i), R(L_j)), \\ D_{DTW}(R(L_i), L_j), \\ D_{DTW}(L_i, R(L_i)) \end{Bmatrix} & \end{cases} \qquad (۵)$$

۴-۳- روش‌های خوشه‌بندی خطوط‌سیر

مسئله خوشه‌بندی خطوط‌سیر را می‌توان بدین صورت تعریف کرد: در مجموعه داده $T=\{tr_1,tr_2,\ldots,tr_n\}$ فرایند گروه‌بندی T به زیر مجموعه‌های $C=\{C_1,C_2,\ldots,C_k\}$ بر اساس یک تابع شباهت را خوشه‌بندی گویند به‌نحوی‌که، $T=\cup_{i=1}^{k} C_i$ و $C_i \cap C_j = \emptyset$ اگر $i \neq j$.

۱ Longest Common Sub-Sequence
۲ Edit Distance for Real Sequences
۳ Edit distance with real penalty

۳-۴-۱- خوشه‌بندی افرازی

این روش خوشه‌بندی یک روش یک سطحی در گروه بندی داده‌ها است به نحوی که داده‌ها را در k گروه متمایز تقسیم‌بندی می‌کند و هر گروه حداقل باید یک شیء را در برگیرد و هر شیء باید فقط در یک خوشه قرار داشته باشد[۲۹]. روش‌های K-means و طیفی دو روش عمده در این گروه هستند.

۳-۴-۱-۱- الگوریتم K-means

فرایند خوشه‌بندی در این روش با انتخاب تصادفی مراکز و اختصاص اشیا به مراکز طبق تابع شباهت آغاز می‌شود. در تکرارهای مختلف مراکز خوشه‌ها که از میانگین‌گیری اعضا خوشه به دست می‌آیند به‌روز می‌شوند و این فرایند تا دستیابی به شرط پایان ادامه می‌یابد. شرط پایان می‌تواند تعداد تکرار مشخص یا دستیابی به ثبات در اعضای خوشه‌ها در نظر گرفته شود [۲۹]. پیاده‌سازی ساده و انعطاف‌پذیری بالا از مزایا K-means است که آن را به محبوب‌ترین روش خوشه‌بندی افرازی تبدیل کرده است. K-means الگوریتمی کارا ازلحاظ زمان اجرا تلقی می‌شود اما وابستگی شدید نتایج به انتخاب اولیه مراکز از نقاط ضعف این روش به‌حساب می‌آید. انتخاب نامناسب مراکز اولیه منجر به دستیابی به بهینه محلی و در نتیجه خوشه‌بندی نامناسب و همچنین افزایش زمان اجرا می‌شود. نیاز به تعیین تعداد خوشه‌ها قبل از شروع خوشه‌بندی از دیگر نقاط ضعف این الگوریتم است. در عمل تعداد خوشه‌ها از پیش معلوم نیست در نتیجه احتمال دستیابی به تعداد خوشه غیر بهینه در این روش افزایش می‌یابد. علاوه بر این با توجه به محدب نبودن بخش‌هایی از خطوط‌سیر امکان دستیابی به نتایج بی‌معنی در این روش وجود دارد.

۳-۴-۱-۲- روش طیفی

برای مجموعه اشیا مفروض، خوشه‌بندی طیفی ماتریس وابستگی A را ایجاد می‌کند و سپس با استفاده از آنالیز بردار ویژه ماتریس لاپلاسین A داده‌ها را خوشه‌بندی می‌کند [۳۰]. این روش خوشه‌بندی را می‌توان در دو حالت کلی نرمال‌شده و غیرنرمال در گراف لاپلاسین طبقه‌بندی کرد که در این تحقیق از خوشه‌بندی طیفی نرمال‌شده استفاده شده است.

در ادامه مراحل خوشه‌بندی با این روش تشریح می‌شود.

۱- ساخت ماتریس وابستگی A برای مجموعه داده X برای همه i و j ها که $i \neq j$ و $A_{ii}=0$ با استفاده از رابطه (۶).

$$A_{ij} = e^{\wedge}((-\operatorname{dist}(xi, xj))/(2\,\partial^{\wedge}2)) \qquad (۶)$$

در این رابطه، ∂ پهنای باند و از پارامترهای اولیه این روش است که توسط کاربر تعیین می‌گردد.

۲- ساخت ماتریس $L= D^{-1/2} A D^{-1/2}$ به‌نحوی‌که D یک ماتریس قطری است و $D_{ii}=\sum_{j=1}^{N} A_{ij}$.

۳- محاسبه ماتریس $E=[e_1,\ldots,e_k] \in R^{N\times N}$ به‌نحوی‌که e_k برابر k امین بردار ویژه نرمال شده ماتریس L از لحاظ بزرگی است.

۴- در نظر گرفتن هر سطر E به‌عنوان یک نقطه در R^k و خوشه‌بندی آن در k خوشه با استفاده از K-means و در نهایت تخصیص شیء اولیه به خوشه k اگر سطر i ام E به خوشهٔ k اختصاص یافته باشد.

برخلاف K-means این الگوریتم قادر به استخراج خوشه‌ها با شکل غیرمحدب است که باعث برتری نتایج آن در خوشه‌بندی خطوط‌سیر می‌شود. به علت استفاده از K-means در مراحل الگوریتم طیفی، نتایج حاصل در هر بار خوشه‌بندی با این روش می‌تواند تغییر یابد.

۳-۴-۲- روش‌های خوشه‌بندی سلسله‌مراتبی

این روش خوشه‌بندی زنجیره‌ای از گروه‌بندی مجموعه داده را ایجاد می‌کند و در دو حالت کلی بالا به پایین و پایین به بالا پیاده‌سازی می‌شود. در دسته بالا به پایین ابتدا همهٔ داده‌ها در یک خوشه قرار دارند و در هر تکرار به خوشه‌های کوچک‌تر تقسیم می‌شوند تا زمانی که در هر خوشه یک شی قرار گیرد و یا شرط پایان حلقه محقق شود. برعکس در دستهٔ پایین به بالا در ابتدا هر شی به عنوان یک کلاس در نظر گرفته می‌شود و در هر مرحله کلاس‌ها با یکدیگر ادغام می‌شوند تا کلاس‌های بزرگ‌تر ایجاد شوند و این تکرار تا دستیابی به شرط خاتمه و یا رسیدن به یک کلاس واحد که دربردارندهٔ همه اشیا است ادامه می‌یابد. این روش برای تجزیه یا ادغام کلاس‌ها نیازمند تعریف فاصله بین دو خوشه است. فاصله دو خوشه از سه روش زیر قابل محاسبه است.

- بیشترین فاصله بین المان‌ها در دو خوشه

$$\max\{\, d(x,y) : x \in A, y \in B \,\} \quad (۷)$$

- کمترین فاصله بین المان‌ها در دو خوشه

$$\min\{\, d(x,y) : x \in A, y \in B \,\} \quad (۸)$$

- میانگین فاصله بین المان‌های دو خوشه

$$\left(\sum_{x \in A}\sum_{y \in B} d(x,y)\right) \Big/ \mathrm{card}(A)\,\mathrm{card}(B) \quad (۹)$$

در این رابطه card(A) و card(B) تعداد اعضای خوشه A و B است.

از مزایای این روش نسبت به روش افرازی می‌توان به قابلیت بالاتر در نمایش و تفسیر خوشه‌ها، دستیابی به نتیجه‌ی قطعی و عدم نیاز به تعیین تعداد خوشه‌ها قبل از شروع خوشه‌بندی اشاره کرد. علاوه بر موارد اشاره‌شده برخلاف روش‌های دسته افرازی در خوشه‌بندی سلسله-مراتبی هیچ فرضی در مورد نحوه توزیع داده‌ها در نظر گرفته نمی‌شود. محدودیت اصلی این روش بار محاسباتی بالای آن است که باعث شده در خوشه‌بندی خطوط‌سیر با حجم بالا نتوان از آن بهره برد. علاوه بر این خوشه‌بندی سلسله‌مراتبی فرایندی غیرقابل‌برگشت است یعنی بعد از ترکیب یا تقسیم خوشه‌ها امکان لغو و برگشت به حالت قبل وجود ندارد.

۳-۴-۳- روش‌های خوشه‌بندی چگالی مبنا

برخلاف روش‌های قبلی که بر مبنای فاصله استوارند، ایده اصلی روش‌های چگالی مبنا، چگالی داده‌ها است. برای هر شی درون یک خوشه، در همسایگی با یک شعاع معلوم ε باید حداقل به تعداد مشخص μ شی وجود داشته باشد؛ به عبارت دیگر چگالی خوشه‌ها نباید از یک حد آستانه مشخص کمتر باشد. الگوریتم DBSCAN می‌توانند کاندید مناسبی برای دسته الگوریتم‌های چگالی مبنا باشند زیرا نیاز به پارامترهای اولیه کمی دارد و علاوه بر این به آسانی برای داده‌های پیچیده و حجیم قابل تعمیم است و سرعت بالایی دارد که برای داده خطوط‌سیر با حجم بالا بسیار حائز اهمیت است [۳۱, ۳۲].

DBSCAN همسایگی به شعاع ε همه نقاط مجموعه داده D را چک می‌کند. اگر همسایگی با شعاع ε نقطه o ($N_\varepsilon(o)$) بیشتر از μ عضو داشته باشد، o اصطلاحا هسته[۱] نامیده می‌شود و خوشه جدید C دربردارنده $N_\varepsilon(o)$ ایجاد می‌شود. سپس همسایگی ε همه نقاط p درون خوشه C که قبلا مشاهده نشده‌اند بررسی می‌گردد. اگر $N_\varepsilon(p)$ دربردارنده بیش از μ نقطه باشد همسایگی p که قبلا عضو C نبوده به آن افزوده می‌شود و برچسب مشاهده‌شده دریافت می‌کند. این فرایند تا زمانی که هیچ نقطه جدیدی به خوشه C اضافه نشود ادامه می یابد. در ادامه، الگوریتم با بررسی نقاط مشاهده‌نشده و یافتن خوشه‌های جدید ادامه می یابد.

مزایا این دسته توابع خوشه‌بندی نسبت به سایر دسته‌ها عبارت‌اند از:

- نیاز به تعیین تعداد خوشه‌ها از قبل وجود ندارد. معمولاً تعداد خوشه‌ها قبل از فرایند خوشه‌بندی مشخص نیست.
- این روش خوشه‌ها با اشکال دلخواه را استخراج می‌کند. این ویژگی بسیار مهمی در شبکه خطوط‌سیر است چون این شبکه از هیچ شکل خاصی پیروی نمی‌کنند.
- از قابلیت شناسایی داده پرت برخوردار اند. به علت تأثیر پارامترهای مختلف مثل تداخل الکترومغناطیسی، قطع لحظه‌ای سنسورها و کمبود باتری وجود داده پرت در خطوط‌سیر اجتناب‌ناپذیر است.
- برخلاف روش‌های K-means و طیفی، DBSCAN روشی قطعی است یعنی نتایج حاصل با هر بار تکرار الگوریتم با پارامترهای برابر یکسان است.

عدم دستیابی به نتیجه قابل قبول هنگام نزدیکی خوشه‌ها به یکدیگر از مشکلات این روش است. علاوه بر این تعیین پارامترهای اولیه در DBSCAN از دیگر چالش‌های استفاده از این روش است.

۳-۴-۴- خوشه‌بندی بر مبنای بهینه‌سازی

خوشه‌بندی را می‌توان در قالب یک مسئله بهینه‌سازی مدلسازی و حل کرد. همواره یافتن تعداد بهینه خوشه‌ها و مکان بهینه مراکز آن‌ها در فرایند خوشه‌بندی از نکات مورد توجه محققان بوده است که با استفاده از روش‌های

[۱] Core

بهینه‌سازی می‌توان پاسخ مناسبی برای آن‌ها یافت. بعضی از الگوریتم‌های خوشه‌بندی اشاره‌شده مثل K-means از مشکل گرفتار شدن در بهینه محلی رنج می‌برند و استفاده از الگوریتم‌های بهینه‌سازی می‌تواند در دستیابی به بهینه کلی راهگشا باشد.

از بین الگوریتم‌های فراابتکاری متعدد پیشنهاد شده همچون ازدحام ذرات، کلونی مورچه‌ها، علف‌های هرز، تکامل تفاضلی و شبیه‌سازی تبرید، الگوریتم ژنتیک [۳۳, ۳۴] بیشتر از سایرین برای دستیابی به نتایج بهتر در خوشه‌بندی به کار گرفته شده است.

در تحقیقات زیادی تلاش شده است قابلیت همگرایی به جواب بهینه در K-means را با ترکیب آن با یک الگوریتم بهینه‌سازی بهبود بخشند [۳۵, ۳۶]. در این تحقیق نیز از الگوریتم ژنتیک برای بهینه کردن نتایج حاصل از K-means در خوشه‌بندی خطوط‌سیر مکانی استفاده شده است.

۳-۵- معیارهای ارزیابی

با توجه به حجم بالای داده‌های مربوط به خط‌سیر، برای ارزیابی الگوریتم‌های خوشه‌بندی در کنار کیفیت خوشه‌ها توجه به زمان محاسباتی نیز از اهمیت بالایی برخوردار است. در این تحقیق از دو معیار کیفیت خوشه‌ها و زمان محاسباتی برای مقایسه عملکرد الگوریتم‌های خوشه‌بندی استفاده شده است.

ارزیابی خوشه‌ها از مهمترین مسائل در خوشه‌بندی است که به شناسایی بهترین گروه‌بندی برای داده‌ها کمک کند. روش‌های مختلفی برای ارزیابی الگوریتم‌های خوشه-بندی در تحقیقات پیشین مورد استفاده قرار گفته‌اند. در این تحقیق از معیار فشردگی و تفکیک‌پذیری خوشه‌ها برای ارزیابی استفاده شده است. معیار فشردگی نشان-دهنده نزدیکی اعضای یک گروه به یکدیگر و تفکیک-پذیری نشان‌دهنده دوری اعضای یک خوشه از اعضای سایر خوشه‌ها است. شاخص سیلووت[۱] از معروف‌ترین معیارها برای سنجش همزمان فشردگی و تفکیک‌پذیری خوشه‌ها است [۳۷]. سیلووت برای هر یک از اعضای خوشه (X_j (j = 1,…, c یک پارامتر کیفیت به نام پهنای سیلووت تعریف می‌کند که مقدار آن بین ۱ و ۱- است و از رابطه (۱۲) محاسبه می‌شود. نزدیکی شاخص سیلووت به ۱ نشان‌دهنده خوشه بندی با کیفیت مناسب است.

$$s(i) = (b(i) - a(i))/\max\{a(i), b(i)\} \qquad (۱۲)$$

در این رابطه $a(i)$ برابر میانگین فاصله بین المان i ام تا دیگر المان‌ها در خوشه X_j است و $b(i)$ حداقل میانگین فاصله بین المان i ام تا همه المان‌های خوشه‌های X_k به‌نحوی‌که $k = 1, \ldots, c; k \neq j$ و c تعداد کل خوشه‌ها می‌باشد. رابطه (۱۳) شاخص سیلووت مربوط به هر خوشه با تعداد اعضای m و رابطه (۱۴) شاخص سیلووت کلی را نشان می‌دهد.

$$S_j = \left(\sum_{i=1}^{m} s(i)\right)/m \qquad (۱۳)$$

$$GS_u = \left(\sum_{j=1}^{c} S_j\right)\Big/ n \qquad (۱۴)$$

۴- پیاده‌سازی و ارزیابی

۴-۱- معرفی مجموعه داده‌های مورد استفاده

مقایسه توابع خوشه‌بندی مختلف بر روی دو مجموعه داده با ویژگی‌های حرکتی متفاوت صورت گرفته است. این دو مجموعه داده در شکل (۴) نشان داده شده اند.

مجموعه داده‌های CVRR شامل چهار گروه داده است که در این تحقیق از میان آن‌ها، مجموعه داده CROSS و I5 استفاده شده است. CROSS شامل ۱۹۰۰ خط‌سیر است که شبیه‌سازی خطوط‌سیر در یک تقاطع چهارراه را در برمی‌گیرد. در مقابل I5 دارای ۸۰۶ خط سیر است که توسط یک ردیاب بینایی[۲] ساده استخراج شده است. خلاصه‌ای از ویژگی‌های مجموعه داده‌ها در جدول (۱) آورده شده است. پیچیدگی شکل آورده شده در این جدول توسط رابطه‌ی (۱۵) بدست می‌آید [۲۵].

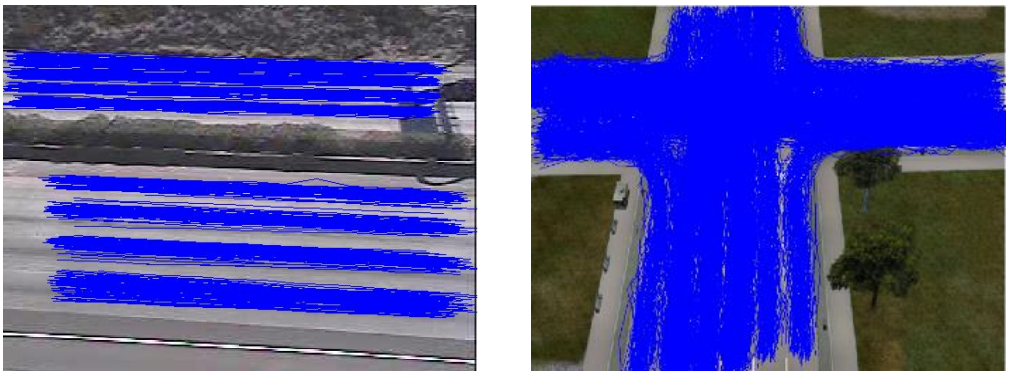
$$\varepsilon = d_{euc}(T_n, T_1)\Big/\sum_{i=1}^{n-1} d_{euc}(T_{i+1}, T_i) \qquad (۱۵)$$

که در این رابطه T_i مختصات نقطه i ام در خط سیر T و n تعداد نقاط آن است.

---

۱ Silhouettes

۲ Visual tracker



الف ب

شکل ۴- مجموعه داده به‌کاررفته (الف)CROSS و (ب)I5

جدول ۱- ویژگی‌های دو مجموعه داده CROSS و I5

داده	تعداد خط سیر	میانگین سرعت	انحراف معیار سرعت	میانگین طول	انحراف معیار طول	پیچیدگی شکل
CROSS	۱۹۰۰	۳۲/۶۱	۱۶/۹۴	۴۲۳/۳۲	۸۸/۶۱	۰/۸۱
I5	۸۰۶	۱۴/۲۸	۶/۸۵	۲۷۹/۵۴	۳۷/۷۳	۱/۰۰

۴-۲- ارزیابی دسته‌های خوشه‌بندی بدون افزودن خطا

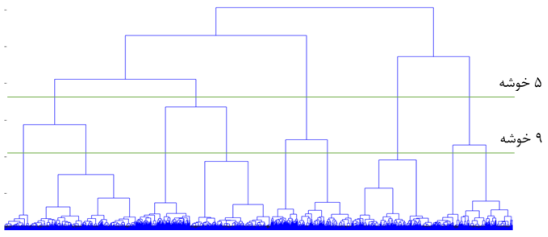
در این مقاله، به منظور ارزیابی عملکرد دسته الگوریتم‌های اشاره‌شده در بخش قبل در خوشه‌بندی خطوط‌سیر، معروف‌ترین الگوریتم(ها) در هر دسته پیاده‌سازی شده است. الگوریتم‌ها در نرم‌افزار متلب ۲۰۱۷ و با استفاده از سیستم با حافظه ۸ گیگابایت و پردازنده ۲/۳۹ گیگاهرتز پیاده‌سازی شده‌اند. پارامترهای به کار رفته برای هر الگوریتم مطابق جدول (۲) در نظر گرفته شده است. پارامترهای انتخابی به نحوی انتخاب شده اند تا تعداد خوشه مورد نظر حاصل شود و بیشترین شاخص سیلووت به دست آید. تعداد خوشه به صورت دلخواه و برابر ۶ و ۱۰ خوشه در نظر گرفته شده است تا تاثیر تعداد خوشه‌ها نیز بر نتایج مورد ارزیابی قرار گیرد. پهنای باند در روش طیفی به نحوی انتخاب شده است که بیشترین شاخص سیلووت حاصل شود. پارامترها در روش چگالی مبنا بر مبنای دستیابی به تعداد مورد نظر خوشه و همچنین حداکثر شاخص سیلووت در نظر گرفته شده‌اند. تعداد تکرار در هر الگوریتم نیز به حدی است که در اعضای خوشه‌ها به ثبات دست یابیم.

جدول ۲- پارامترهای اولیه در نظر گرفته شده برای روش‌های مختلف و در توابع فاصله مختلف

روش خوشه‌بندی	تابع فاصله	نوع پارامتر	مقدار پارامتر	
			CROSS	I5
طیفی	اقلیدسی	پهنای باند	۶۰	۹۰
	DTW	پهنای باند	۳۰	۴۰
چگالی مبنا	اقلیدسی	ε	۶۶ ۸۲	۱۶ ۱۰
		minpoint	۵۰	۲ ۵
	DTW	ε	۶۵۶ ۷۳۰	۹۰ ۱۲۸
		minpoint	۲۵ ۵۰	۵
مبتنی بر بهینه سازی	اقلیدسی و DTW	تعداد جمعیت	۳۰	۳۰
		درصد تقاطع	۵۰٪	۵۰٪
		درصد جهش	۶٪	۶٪

قبل از مقایسه کلی الگوریتم‌ها به ویژگی شاخص دسته سلسله مراتبی که آن را از سایر دسته‌ها متمایز می‌سازد پرداخته می‌شود. در روش سلسله مراتبی نیاز به تعیین تعداد خوشه‌ها وجود ندارد. در شکل (۵) نمودار

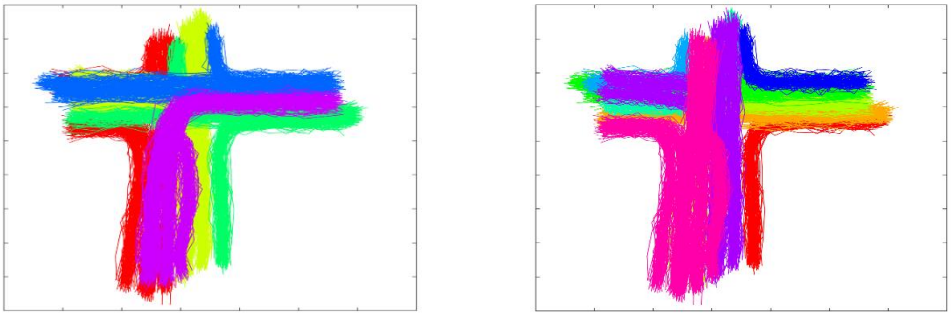
درختواره حاصل از این خوشه‌بندی نمایش داده شده است. درختواره برای نمایش ساختار سلسله‌مراتبی خوشه‌ها به کار می‌رود. دو خط افقی متقاطع با درختواره از بالا به پایین به ترتیب محل قرار گیری ۵ خوشه و ۹ خوشه را برای مجموعه داده CROSS نمایش می‌دهد.



شکل ۵- نمودار درختواره در خوشه بندی سلسله مراتبی

همانطور که در شکل (۶) دیده می‌شود، در این روش ذات سلسله مراتبی و توالی خوشه‌بندی در نظر گرفته شده است.

هر روش خوشه‌بندی خوشه‌ها با شکل خاص خود ایجاد می‌کند. با توجه به مطالب بیان شده در بخش معیارهای ارزیابی برای مقایسه کارایی دسته‌های مختلف می‌توان از دو معیار فشردگی و تفکیک‌پذیری خوشه‌ها بهره برد. با توجه به حجم بالای داده‌های خطسیر زمان محاسباتی موردنیاز هر روش نیز دیگر معیار مهم مقایسه بین روش‌ها است. برای مقایسه صحیح‌تر و جلوگیری از تاثیرپذیری از خاصیت تصادفی نتایج در بعضی روش‌ها هر الگوریتم ۳۰ بار اجرا شده است.



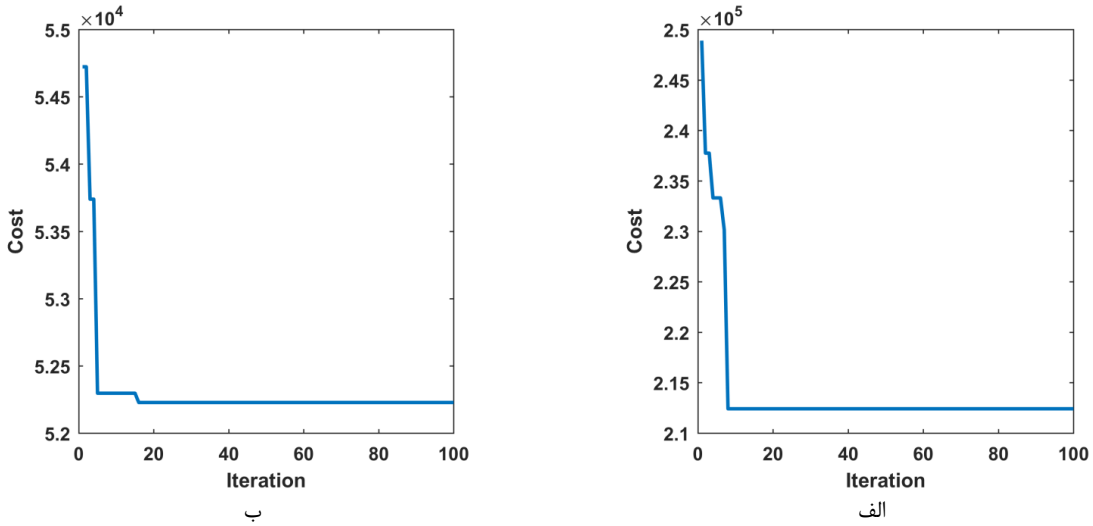
ب الف

شکل ۶- نمایش سلسله مراتبی خوشه‌ها (الف) خوشه بندی با ۹ خوشه و (ب) خوشه بندی با ۵ خوشه

شاخص سیلووت حاصله از روش‌های مختلف و استفاده از دو تابع فاصله اقلیدسی و DTW و بر روی هر دو مجموعه داده در جدول (۳) نمایش داده شده است. به صورت کلی نتایج به دست آمده از دسته مبتنی بر بهینه‌سازی با انحراف معیار مقادیر شاخص سیلووت از ۱ برابر ۰/۰۰۶ بر بقیه نتایج برتری دارد. با اضافه کردن الگوریتم ژنتیک به K-means همه نتایج آن در هر دو مجموعه داده و استفاده از هر دو نوع تابع فاصله بهبود یافته است. نمودار همگرایی مربوط به این الگوریتم در شکل (۷) نشان داده شده است. الگوریتم ژنتیک در K-means باعث دستیابی به مکان بهینه خوشه‌ها و حل مشکل گرفتار شدن آن در کمینه محلی شده است. بعد از دسته مبتنی بر بهینه‌سازی دسته افرازی با انحراف معیار مقادیر شاخص سیلووت از ۱ برابر ۰/۰۱۰ در رده دوم بهترین نتایج قرار می‌گیرد. زیردسته طیفی در این گروه با انحراف معیار مقادیر از ۱ برابر ۰/۰۰۹ نسبت به زیردسته K-means با انحراف معیار مقادیر از ۱ برابر ۰/۰۱۱ از برتری نسبی برخوردار است و دلیل آن را می‌توان در قابلیت بالای خوشه-بندی طیفی در استخراج خوشه با شکل غیرمحدب جستجو کرد. دسته سلسله مراتبی پیاده شده در این تحقیق برای مجموعه داده CROSS نتایج بسیار خوب و برای مجموعه داده I5 نتایج ضعیفی ارائه کرده است که نشان از وابستگی بالای این روش به نوع داده است.

جدول ۳- مقدار شاخص سیلووت برای روش‌های مختلف خوشه‌بندی در استفاده از توابع فاصله اقلیدسی و DTW و بر روی دو مجموعه داده

دسته روش خوشه‌بندی		CROSS				I5			
		۶ خوشه		۱۰ خوشه		۶ خوشه		۱۰ خوشه	
		اقلیدسی	DTW	اقلیدسی	DTW	اقلیدسی	DTW	اقلیدسی	DTW
افرازی	K-means	۰/۹۹۵۹	۰/۹۹۶۲	۰/۹۹۴۴	۰/۹۹۴۲	۰/۹۹۱۰	۰/۹۹۱۹	۰/۹۷۸۵	۰/۹۸۶۳
	طیفی	۰/۹۹۶۴	۰/۹۹۶۵	۰/۹۹۲۷	۰/۹۸۹۲	۰/۹۹۳۳	۰/۹۹۳۸	۰/۹۸۶۵	۰/۹۸۶۰
سلسله مراتبی		۰/۹۹۶۸	۰/۹۹۸۲	۰/۹۹۷۴	۰/۹۹۷۳	۰/۹۶۰۹	۰/۹۵۷۲	۰/۹۱۹۶	۰/۹۱۹۱
چگالی مبنا		۰/۹۹۴۵	۰/۹۹۲۹	۰/۹۹۶۸	۰/۹۹۶۷	۰/۸۴۹۲	۰/۹۴۳۲	۰/۷۲۰۸	۰/۸۳۶۸
مبتنی بر بهینه‌سازی		۰/۹۹۸۴	۰/۹۹۸۷	۰/۹۹۷۷	۰/۹۹۸۳	۰/۹۹۴۱	۰/۹۹۶۵	۰/۹۸۷۸	۰/۹۹۳۱



ب الف

شکل ۷- نمودار همگرایی خوشه‌بندی با روش مبتنی بر بهینه‌سازی الف) بر روی داده CROSS در ۱۰ خوشه ب) بر روی داده I5 در ۶ خوشه

در کل انحراف معیار مقادیر شاخص سیلووت از ۱ در این روش برابر ۰/۰۴۵ است که در رده سوم بالاترین شاخص سیلووت بعد از دسته افرازی و مبتنی بر بهینه-سازی قرار می‌گیرد. آخرین دسته روش‌های چگالی مبنا با انحراف معیار مقادیر شاخص سیلووت از ۱ برابر ۰/۱۲۸ است که دلیل آن تفاوت طول خطوط‌سیر و همچنین نزدیکی خوشه‌ها به هم در مجموعه داده I5 است. تفاوت نتایج حاصل برای داده‌های مختلف مبین لزوم توجه به داده در انتخاب تابع خوشه‌بندی مناسب است. برای داده CROSS بهترین نتایج به ترتیب از دسته‌های مبتنی بر بهینه‌سازی، سلسله مراتبی، چگالی مبنا و افرازی حاصل شده است. با این حال برای داده I5 بهترین نتایج به ترتیب از دسته‌های مبتنی بر بهینه‌سازی، افرازی، سلسله مراتبی و چگالی مبنا به دست آمده است.

علاوه بر داده انتخاب تابع فاصله نیز در نتایج حاصل از روش‌های مختلف موثر است. در مجموع فاصله اقلیدسی با انحراف معیار مقادیر شاخص سیلووت از ۱ برابر ۰/۰۷۴ نسبت به DTW با انحراف معیار مقادیر شاخص سیلووت از ۱ برابر ۰/۰۴۴ عملکرد ضعیف‌تری را نتیجه داده است. هر چند در بعضی توابع خوشه‌بندی فاصله اقلیدسی نسبت به DTW شاخص سیلووت بالاتری را نتیجه داده اما در مجموع DTW نسبت به اقلیدسی عملکرد بهتری داشته است.

۴-۳- ارزیابی دسته‌ها با افزودن خطا به داده‌ها

در جدول (۴) تأثیر داده پرت و نویز موجود در خطوط-سیر در عملکرد توابع مختلف مورد بررسی قرار گرفته است. برای این منظور دو مجموعه داده جدید یکی دربردارنده مجموعه داده I5 به همراه نویز با توزیع گوسین و سیگنال به نویز ۲ و دیگری مجموعه داده I5 با ۱۰ درصد داده پرت ایجاد شده است و در ۶ خوشه تقسیم بندی شده‌اند. مطابق این جدول در حضور نویز و داده پرت بهترین نتایج مربوط به روش مبتنی بر بهینه‌سازی با انحراف معیار مقادیر شاخص سیلووت از ۱ برابر ۰/۰۰۶ است. بعد از این روش به ترتیب روش‌های سلسله مراتبی، افرازی و چگالی مبنا بالاترین شاخص سیلووت را نتیجه داده اند.

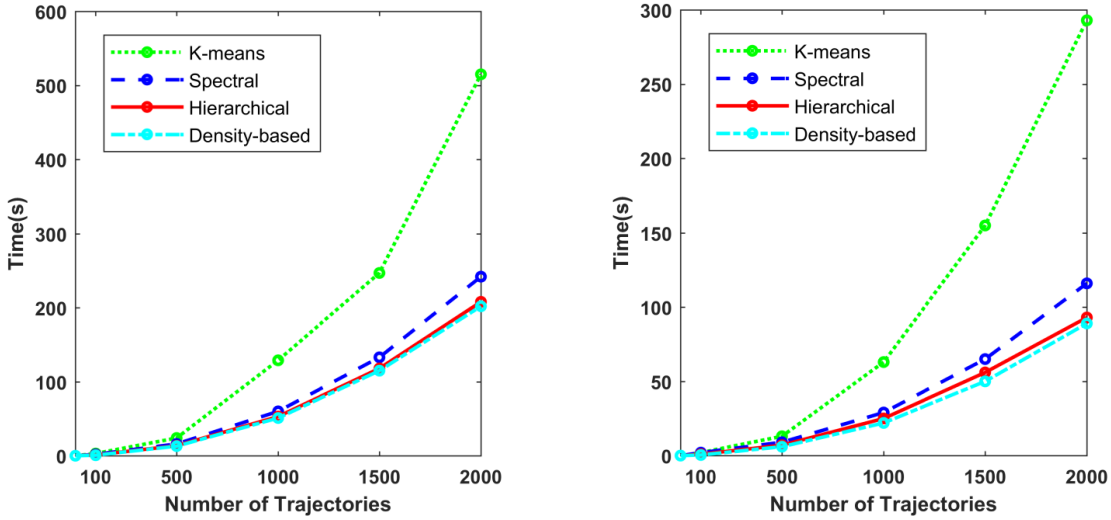
جدول ۴- مقدار شاخص سیلووت برای الگوریتم‌های مختلف در حضور خطا

روش خوشه بندی		داده پرت ۱۰٪		نویز SNR=۲	
		اقلیدسی	DTW	اقلیدسی	DTW
افرازی	K-means	۰/۹۷۶۵	۰/۹۸۶۶	۰/۹۹۱۴	۰/۹۸۶۶
	طیفی	۰/۹۹۲۴	۰/۹۹۳۰	۰/۹۹۳۶	۰/۹۹۱۱
سلسله مراتبی		۰/۹۱۹۶	۰/۹۴۰۵	۰/۹۴۷۷	۰/۹۵۷۲
چگالی مبنا		۰/۷۷۹۶	۰/۸۸۱۴	۰/۹۰۱۱	۰/۹۲۱۷
مبتنی بر بهینه‌سازی		۰/۹۹۲۶	۰/۹۹۴۰	۰/۹۹۴۹	۰/۹۹۶۰

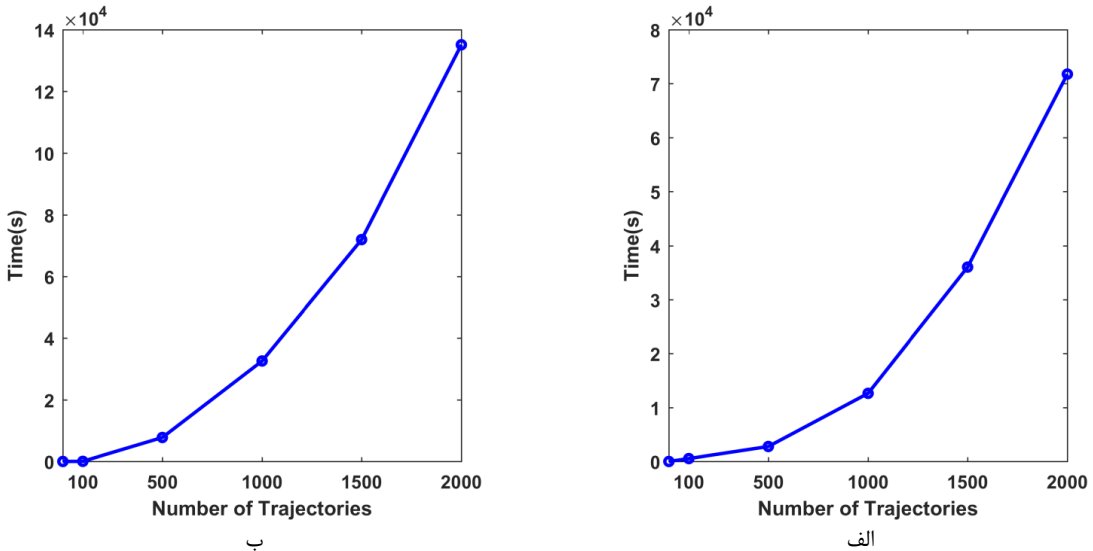
روش طیفی زیر مجموعه دسته افرازی از مقاومت بالایی در برابر داده پرت برخوردار است و شاخص سیلووت در این روش کمترین کاهش را داشته است. در مقابل روش‌های چگالی مبنا و مبتنی بر بهینه‌سازی بیشترین مقاومت را در برابر نویز داشته‌اند.

۴-۴- ارزیابی دسته‌ها از لحاظ زمان محاسباتی

علاوه بر کیفیت خوشه‌ها پارامتر زمان نیز در انتخاب تابع مناسب برای خوشه‌بندی تأثیرگذار است. در شکل (۸) روش‌های مختلف بر اساس زمان محاسباتی مورد مقایسه قرار گرفته‌اند. روش مبتنی بر بهینه‌سازی به علت زمان محاسباتی بالا به صورت جداگانه و در شکل (۹) ترسیم شده است.



شکل ۸- مقایسه زمانی الگوریتم‌ها با استفاده از فاصله الف) اقلیدسی ب) DTW



شکل ۹- زمان مبتنی بر بهینه‌سازی با استفاده از فاصله الف) اقلیدسی ب) DTW

روش مبتنی بر بهینه‌سازی برای ۲۰۰۰ خط سیر در تابع اقلیدسی حدودا به ۷۰۰۰۰ ثانیه و در تابع DTW حدودا به ۱۳۵۰۰۰ ثانیه زمان نیاز داشته است. در مقابل روش های چگالی مبنا، سلسله مراتبی و طیفی در استفاده از تابع DTW و برای ۲۰۰۰ خط به ترتیب به ۲۰۲، ۲۰۸ و ۲۴۲ ثانیه زمان نیاز داشته‌اند. مطابق شکل (۷) زمان لازم برای تابع K-means در هر دو تابع اقلیدسی و DTW بیشتر از سایر الگوریتم‌ها است. سریع‌ترین الگوریتم‌ها به ترتیب چگالی‌مبنا، سلسله‌مراتبی، طیفی، K-means و در آخر مبتنی بر بهینه‌سازی است.

۵- نتیجه‌گیری

با توجه به تعدد الگوریتم‌های مختلف با ویژگی‌های متفاوت در خوشه‌بندی خطوط‌سیر مکانی نیاز به بررسی جامع و ارائه دسته‌بندی از این روش‌ها ضروری به نظر می‌رسد. در این تحقیق عملکرد روش‌های خوشه‌بندی تعمیم یافته از داده‌های نقطه‌ای در خوشه‌بندی خطوط-سیر مورد بررسی و ارزیابی قرارگرفته است و به چهار دسته کلی تقسیم بندی شده‌اند. تأثیر نویز و داده پرت به عنوان مهم‌ترین پارامترهای دخیل در کیفیت عملکرد توابع خوشه‌بندی نیز لحاظ شده است. بنا به نتایج به‌دست‌آمده انتخاب روش مناسب خوشه‌بندی وابسته به داده و ویژگی‌های آن و همچنین تابع فاصله مورد استفاده است. با این حال در مجموع بهترین نتایج به ترتیب از دسته‌های مبتنی بر بهینه‌سازی، افرازی، سلسله مراتبی و چگالی مبنا حاصل شده است. بهبود قابل توجه نتایج نسبت به دسته‌ی افرازی (صرفاً زیر دسته K-means) و دستیابی به انحراف معیار مقادیر شاخص سیلووت از ۱ برابر ۰/۰۰۵ از نقاط قوت و زمان محاسباتی بالا از نقاط ضعف خوشه‌بندی مبتنی بر بهینه‌سازی است. با توجه به ذات داده‌های خط‌سیر وجود داده پرت و نویز در آن اجتناب ناپذیر است با این حال بعضی روش‌های خوشه-بندی همچون دسته افرازی (صرفا زیر دسته‌ی K-means) به شدت وابسته به داده پرت هستند. دسته افرازی (صرفا زیر دسته طیفی) دارای مقاومت بالا در برابر داده پرت و دسته‌های چگالی مبنا و مبتنی بر بهینه‌سازی بیشترین مقاومت را در برابر نویز داشته‌اند. انحراف معیار مقادیر شاخص سیلووت از ۱ حاصل از دسته‌ی مبتنی بر بهینه-سازی در حضور نویز و داده پرت برابر ۰/۰۰۶ است که در مجموع بهترین عملکرد در حضور نویز و داده پرت است. عدم بررسی اثر تبدیلات هندسی، تغییر در نرخ برداشت نقاط و کشیدگی زمانی بر عملکرد توابع مختلف در خوشه‌بندی خطوط‌سیر و استفاده از فقط دو مجموعه داده از محدودیت‌های اصلی این تحقیق است. علاوه بر این در این تحقیق تنها توابعی که از خوشه‌بندی داده‌های نقطه‌ای تعمیم یافته‌اند مورد ارزیابی قرار گرفته شد. پیشنهاد می‌شود برای تحقیقات آتی توابع خوشه‌بندی مختص خطوط‌سیر نیز مورد ارزیابی قرار گیرند، نقاط قوت و ضعف آن‌ها در شرایط مختلف و در حضور نویز و داده پرت بررسی شود و دسته‌بندی مناسب برای آن‌ها ارائه گردد.

مراجع

[1] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in Proceedings of the 6th ACM conference on Embedded network sensor systems, 2008, pp. 323-336: ACM.

[2] E. Kamar and E. Horvitz, "Collaboration and Shared Plans in the Open World: Studies of Ridesharing," in IJCAI, 2009, vol. 9, p. 187.

[3] M. Batty, J. DeSyllas, and E. Duxbury, "The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades," International Journal of Geographical Information Science, vol. 17, no. 7, pp. 673-697, 2003.

[4] B. Dumont, A. Boissy, C. Achard, A. Sibbald, and H. Erhard, "Consistency of animal order in spontaneous group movements allows the measurement of leadership in a group of grazing heifers," Applied Animal Behaviour Science, vol. 95, no. 1, pp. 55-66, 2005.

[5] P. Rai and S. Singh, "A survey of clustering techniques," International Journal of Computer Applications, vol. 7, no. 12, pp. 1-5 ,2010.

[6] R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Transactions on neural networks, vol. 16, no. 3, pp. 645-678, 2005.

[7] H. F. Tork, "Spatio-temporal clustering methods classification," in Doctoral Symposium on Informatics Engineering, 2012, vol. 1, no. 1, pp. 199-209.

[8] M. Nanni, Clustering methods for spatio-temporal data. SEU, 2002.

[9] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, "Indexing multi-dimensional time-series with support for multiple distance measures," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 216-225: ACM.

[10] W. Meesrikamolkul, V. Niennattrakul, and C. A. Ratanamahatana, "Shape-based clustering for time series data," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2012, pp. 530-541: Springer.

[11] S. Atev, G. Miller, and N. P. Papanikolopoulos, "Clustering of vehicle trajectories," IEEE Transactions on Intelligent Transportation Systems, vol. 11, no. 3, pp .647-657 ,2010.

[12] S. Atev, O. Masoud, and N. Papanikolopoulos, "Learning traffic patterns at intersections by spectral clustering of motion trajectories," in Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on, 2006, pp. 4851-485 :٦IEEE.

[13] Z. Fu, W. Hu, and T. Tan, "Similarity based vehicle trajectory clustering and anomaly detection," in Image Processing, 2005. ICIP 2005. IEEE International Conference on, 2005, vol. 2, pp. II-602: IEEE.

[14] P. P. Rodrigues, J. Gama, and J. P .Pedroso, "ODAC: Hierarchical clustering of time series data streams," in Proceedings of the 2006 SIAM International Conference on Data Mining, 2006, pp. 499-503: SIAM.

[15] M. Nanni and D. Pedreschi, "Time-focused clustering of trajectories of moving objects," Journal of Intelligent Information Systems, vol. 27, no. 3, pp. 267-289, 2006.

[16] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," Data & Knowledge Engineering, vol. 60, no. 1, pp. 208-221, 2007.

[17] L. X. Liu, J. T. Song, B. Guan, Z. X. Wu, and K. J. He, "Tra-dbscan: a algorithm of clustering trajectories," in Applied Mechanics and Materials, 2012, vol. 121, pp. 4875-4879: Trans Tech Publ.

[18] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares, "A clustering-based approach for discovering interesting places in trajectories," in Proceedings of the 2008 ACM symposium on Applied computing, 2008, pp. 863-868: ACM.

[19] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in Proceedings of the 2007 ACM SIGMOD international conference on Management of data, 2007, pp. 593-604: ACM.

[20] A. K. Akasapu, P. S. Rao, L. Sharma, and S. Satpathy, "Density based k-nearest neighbors clustering algorithm for trajectory data," International Journal of Advanced Science and Technology, vol. 31, no. 1, 2011.

[21] A. Ahmadyfard and H. Modares, "Combining PSO and k-means to enhance data clustering," in Telecommunications, 2008. IST 2008. International Symposium on, 2008, pp. 688-691 :IEEE.

[22] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. J. Brown, "Incremental genetic K-means algorithm and its application in gene expression data analysis," BMC bioinformatics, vol. 5, no. 1, p. 172, 2004.

[23] Z. Izakian and M. Mesgari, "Fuzzy clustering of time series data: A particle swarm optimization approach," Journal of AI and Data Mining, vol. 3, no. 1, pp. 39-46, 2015.

[24] Z. Izakian, M. S. Mesgari, and A. Abraham, "Automated clustering of trajectory data using a particle swarm optimization," Computers, Environment and Urban Systems, vol. 55, pp. 55-65, 2016.

[25] Z. Zhang, K. Huang, and T. Tan, "Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes," in Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, 2006, vol. 3, pp. 1135-1138: IEEE.

[26] H. Wang, H. Su, K. Zheng, S. Sadiq, and X. Zhou, "An effectiveness study on trajectory similarity measures," in Proceedings of the Twenty-Fourth Australasian Database Conference-Volume 137, 2013, pp. 13-22: Australian Computer Society, Inc.

[27] W. Bailer, "A comparison of distance measures for clustering video sequences," in Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on, 2008, pp. 595-599: IEEE.

[28] B. Morris and M. Trivedi, "Learning trajectory patterns by clustering: Experimental studies and comparative evaluation," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 312-319: IEEE.

[29] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM computing surveys (CSUR), vol. 31, no. 3, pp. 264-323, 1999.

[30] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Advances in neural information processing systems ,2002 ,pp. 849-856.

[31] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Kdd, 1996, vol. 96, no. 34, pp. 226-231.

[32] A. Zhou, S. Zhou, J. Cao, Y. Fan, and Y .Hu, "Approaches for scaling DBSCAN algorithm to large spatial databases," Journal of computer science and technology, vol. 15, no. 6, pp. 509-526, 2000.

[33] D. Goldberg, "Genetic algorithms in search, optimization, and machine learning (1989)," New York , Addison-Wesley.

[34] J. H. Holland, Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, 1992.

[35] Y. Lin et al., "K-means optimization clustering algorithm based on particle swarm optimization and multiclass merging," in Advances in Computer Science and Information Engineering: Springer, 2012, pp. 569-578.

[36] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. J. Brown, "FGKA: A fast genetic k-means clustering algorithm," in Proceedings of the 2004 ACM symposium on Applied computing, 2004, pp. 622-623: ACM.

[37] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of computational and applied mathematics ,vol. 20, pp. 53-65, 1987.