دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد

گرایش مهندسی نرم‌افزار

عنوان:

الگوریتم‌های تقریبی برای خوشه‌بندی نقاط در مدل جویبار داده

نگارش:

بهنام حاتمی ورزنه

استاد راهنما:

حمید ضرابی‌زاده

شهریور ۱۳۹۴

بسم الله الرحمن الرحیم

به نام خدا

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد

عنوان: الگوریتم‌های تقریبی برای خوشه‌بندی نقاط در مدل جویبار داده

نگارش: بهنام حاتمی ورزنه

کمیته‌ی ممتحنین

استاد راهنما: حمید ضرابی‌زاده امضاء:

استاد مشاور: حمید بیگی امضاء:

استاد مدعو: استاد ممتحن امضاء:

تاریخ:

چکیده

این قسمت باید تکمیل گردد.

کلیدواژه‌ها: خوشه‌بندی، kـ مرکز، جویبار داده، الگوریتم تقریبی

فهرست مطالب

۱ مقدمه **۹**

۱ـ۱ تعریف مسئله ۱۰

۱ـ۲ اهمیت موضوع ۱۳

۱ـ۳ ادبیات موضوع ۱۴

۱ـ۴ اهداف تحقیق ۱۴

۱ـ۵ ساختار پایان‌نامه ۱۵

۲ مفاهیم اولیه **۱۷**

۲ـ۱ مسائل ان‌پی‌ـسخت ۱۷

۲ـ۱ـ۱ پوشش رأسی ۱۹

۲ـ۲ الگوریتم‌های تقریبی ۲۰

۲ـ۲ـ۱ میزان تقریب‌پذیری مسائل ۲۲

۲ـ۳ الگوریتم‌های جویبارداده ۲۲

۳ کارهای پیشین **۲۳**

۳ـ۱ kـمرکز در حالت ایستا ۲۳

۳ـ۲ kـمرکز در حالت جویبار داده ۲۶

۳ـ۳ kـمرکز با داده‌های پرت ۳۳

۴ نتایج جدید ۳۴

۵ نتیجه‌گیری ۳۵

آ مطالب تکمیلی ۳۶

فهرست شکل‌ها

۱ـ۱ نمونه‌ای ازمسئله‌ی ۲ـ مرکز ۱۱
۱ـ۲ نمونه‌ای ازمسئله‌ی ۲ـ مرکز با داده‌های پرت ۱۲
۱ـ۳ نمونه‌ای ازمسئله‌ی ۲ـ مرکز در حالت پیوسته ۱۳

۳ـ۱ نمونه‌ای از تخصیص نقاط به ازای مراکز آبی رنگ. ۲۴
۳ـ۲ نمونه‌ای از تبدیل یک گراف ورودی مسئله‌ی پوشش رأسی به یک ورودی مسئله‌ی kـ مرکز(در گراف سمت چپ، یال‌های سیاه وزن ۱ و یال‌های آبی، وزن ۲ دارند) . ۲۴
۳ـ۳ نمونه‌ای از حل مسئله‌ی ۳ـ مرکز با الگوریتم گنزالز ۲۵
۳ـ۴ نمونه‌ای از توری‌بندی الگوریتم ضرابی‌زاده (نقاط آبی، مراکز به دست آمده از الگوریتم تقریبی است). پس از توری‌بندی کافی است برای هر کدام از خانه‌های شبکه‌بندی، تنها یکی را به عنوان نماینده در نظر بگیریم. ۲۷
۳ـ۵ نمونه‌ای از اجرای الگوریتم ضرابی‌زاده بر روی چهار نقطه $P_1 \cdots P_4$ که به ترتیب اندیس در جویبار داده فرا می‌رسند و دایره‌های $B_2 \cdots B_4$ دایره‌هایی که الگوریتم به ترتیب نگه می‌دارد. ۲۹
۳ـ۶ اثبات لم ۳ـ۲ . . . . ۳۱
۳ـ۷ اثبات لم ۳ـ۳ . . . . ۳۳

فهرست جدول‌ها

۲ ـ ۱ نمونه‌هایی از ضرایب تقریب برای مسائل بهینه‌سازی ۲۲

فصل ۱

مقدمه

مسئله‌ی **خوشه‌بندی**[۱] یکی از مهم‌ترین مسائل داده‌کاوی[۲] به حساب می‌آید. در این مسئله هدف، دسته‌بندی تعدادی جسم به گونه‌ای است که اجسام در یک دسته (خوشه)، نسبت به یکدیگر در برابر دسته‌های دیگر شبیه‌تر باشند (معیارهای متفاوتی برای تشابه تعریف می‌گردد). این مسئله در حوزه‌های مختلفی از علوم کامپیوتر، از جمله داده‌کاوی، جست‌وجوی الگو[۳]، پردازش تصویر[۴]، بازیابی اطلاعات[۵] و بایوانفورماتیک[۶] مورد استفاده قرار می‌گیرد [۱].

مسئله‌ی خوشه‌بندی، به‌خودی‌خود یک مسئله‌ی الگوریتمی به حساب نمی‌آید. راه‌حل‌های الگوریتمی بسیار زیادی برای خوشه‌بندی تعریف شده است. این الگوریتم‌ها را براساس رویکردهای مختلفی که به مسئله دارند، می‌توان در یکی از چهار دسته‌بندی زیر قرار داد:

- خوشه‌بندی‌های سلسه‌مراتبی[۷]
- خوشه‌بندی‌های مرکزگرا[۸]

---

[۱] Clustering
[۲] Data mining
[۳] Pattern recognition
[۴] Image analysis
[۵] Information retrieval
[۶] Bioinformatics
[۷] Hierarchical clustering
[۸] Centroid-based clustering

- خوشه‌بندی‌های مبتنی بر توزیع[۹] نقاط
- خوشه‌بندی‌های مبتنی بر چگالی[۱۰] نقاط

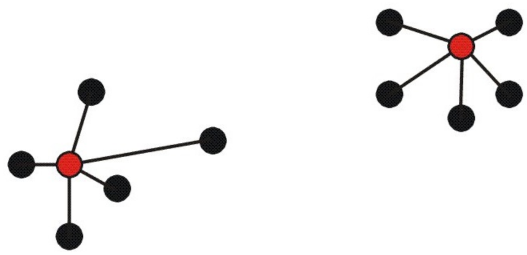
در عمل هیچ کدام از راه‌حل‌های بالا بر دیگری ارجهیت ندارند و باید راه‌حل مد نظر را متناسب با کاربرد مطرح مورد استفاده قرار داد. به طور مثال استفاده از الگوریتم‌های مرکزگرا، برای خوشه‌های غیر محدب به خوبی عمل نمی‌کند. یکی از راه‌حل‌های شناخته شده برای مسئله‌ی خوشه‌بندی، الگوریتم **kـ مرکز** است. در این الگوریتم هدف، پیدا کردن k نقطه به عنوان مرکز دسته‌ها است به طوری‌که شعاع دسته‌ها تا حد ممکن کمینه شود. در نظریه‌ی گراف، مسئله‌ی kـ مرکز متریک[۱۱] یا مسئله‌ی **استقرار تجهیزات متریک**[۱۲] یک مسئله‌ی **بهینه‌سازی ترکیبیاتی**[۱۳] است. فرض کنید که n شهر و فاصله‌ی دوبه‌دوی آن‌ها، داده‌شده است. می‌خواهیم k انبار در شهرهای مختلف بسازیم به‌طوری‌که بیش‌ترین فاصله‌ی هر شهری از نزدیک‌ترین انبار به خود، کمینه گردد. در حالت نظریه‌ی گراف آن، این بدان معناست که مجموعه‌ای شامل k رأس انتخاب کنیم به‌طوری‌که بیش‌ترین فاصله‌ی هر نقطه از نزدیک‌ترین نقطه‌اش داخل مجموعه‌ی k عضوی کمینه گردد. توجه نمایید که فاصله‌ی بین رئوس باید در فضای متریک[۱۴] باشند و یا به زبان دیگر، یک گراف کامل داشته باشیم که فاصله‌ها در آن در رابطه‌ی مثلثی[۱۵] صدق می‌کنند. مثالی از مسئله‌ی ۲ـ مرکز را در شکل ۱ـ۱ نشان داده شده است.

در این پژوهش، مسئله‌ی kـ مرکز با متریک‌های خاص و برای kهای کوچک مورد بررسی قرار گرفته است و از جنبه‌های متفاوتی بهبود یافته است. در زیرفصل بعدی، تعریف رسمی[۱۶] از مسائلی که در این پایان‌نامه مورد بررسی قرار می‌گیرند را تعریف کرده و در مورد هر کدام توضیح مختصری می‌دهیم.

۱ـ۱ تعریف مسئله

تعریف دقیق‌تر مسئله‌ی kـ مرکز در زیر آمده است:

[۹] Distribution-based
[۱۰] Density-based
[۱۱] Metric
[۱۲] Metric facility location
[۱۳] Combinatorial optimization
[۱۴] Metric space
[۱۵] Triangle equation
[۱۶] Formal



شکل ۱ــ۱: نمونه‌ای ازمسئله‌ی ۲ــ مرکز

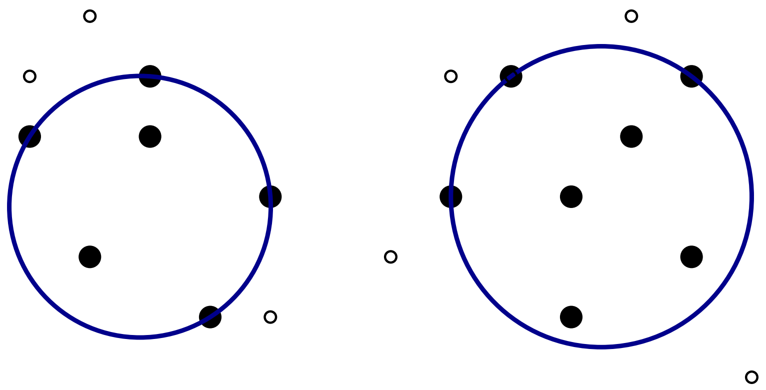
مسئله‌ی ۱ــ۱ *یک گراف کامل بدون جهت $G = (V, E)$ با فاصله‌ی d، که از نامساوی مثلثی پیروی می‌کند داده‌شده است. زیرمجموعه‌ی $S \subseteq V$ با اندازه‌ی k را به گونه‌ای انتخاب کنید به‌طوری‌که عبارت زیر را کمینه کند:*

$$\max_{v \in V}\{\min_{s \in S} d(v, s)\}$$

گونه‌های مختلفی از مسئله‌ی kــ مرکز با محدودیت‌های متفاوتی به‌وسیله‌ی پژوهشگران، مورد مطالعه قرار گرفته است. از جمله‌ی این گونه‌ها، می‌توان به حالتی که در بین داده‌های ورودی، داده‌های پرت وجود دارد، می‌توان اشاره کرد. در واقع در این مسئله، قبل از خوشه‌بندی می‌توانیم تعدادی از نقاط ورودی را حذف نماییم و سپس به خوشه‌بندی نقاط بپردازیم. سختی این مسئله از آنجاست که نه تنها باید مسئله‌ی خوشه‌بندی را حل نمود، بلکه در ابتدا باید تصمیم گرفت که کدام یک از داده‌ها را به‌عنوان داده‌ی پرت در نظر گرفت که بهترین جواب در زمان خوشه‌بندی به دست آید. در واقع اگر تعداد نقاط پرتی که مجاز به حذف است، برابر صفر باشد، مسئله به مسئله‌ی kــ مرکز تبدیل می‌شود. نمونه‌ای از مسئله‌ی ۲ــ مرکز با ۷ داده‌ی پرت را در شکل ۱ــ۲ می‌توانید ببینید. تعریف دقیق‌تر این مسئله در زیر آمده است:

مسئله‌ی ۱ــ۲ *یک گراف کامل بدون جهت $G = (V, E)$ با فاصله‌ی d، که از نامساوی مثلثی پیروی می‌کند داده‌شده است. زیرمجموعه‌ی $Z \subseteq V$ با اندازه‌ی z و $S \subseteq V - Z$ با اندازه‌ی k را انتخاب کنید به‌طوری‌که عبارت زیر را کمینه کند:*

$$\max_{v \in (V-Z)}\{\min_{s \in S} d(v, s)\}$$



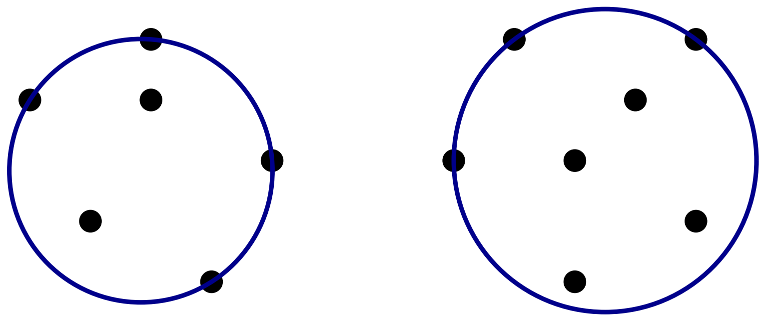
شکل ۱ــ۲: نمونه‌ای ازمسئله‌ی ۲ــ مرکز با داده‌های پرت

گونه‌ی دیگری از مسئله‌ی kــ مرکز که در سال‌های اخیر مورد توجه قرار گرفته است، حالت جویبار داده‌ی آن است. در این گونه از مسئله‌ی kــ مرکز، در ابتدا تمام نقاط در دسترس نیستند، بلکه به مرور زمان نقاط در دسترس قرار می‌گیرند. محدودیت دومی که وجود دارد، محدودیت شدید حافظه است، به طوری که نمی‌توان تمام نقاط را در حافظه نگه داشت و بعضا حتی در حافظه‌ی جانبی نیز نمی‌توان ذخیره نمود و به طور معمول باید مرتبه‌ی حافظه‌ای کم‌تر از مرتبه حافظه‌ی خطی[۱۷] متناسب با تعداد نقاط استفاده نمود. از این به بعد به چنین مرتبه‌ای، مرتبه‌ی **زیرخطی** می‌گوییم. مدلی که ما در این پژوهش بر روی آن تمرکز داریم مدل جویبار داده تک‌گذره [۲] است. یعنی تنها یک بار می‌توان از ابتدا تا انتهای داده‌ها را بررسی کرد و پس از عبور از یک داده، اگر آن را در حافظه ذخیره نکرده باشیم، دیگر به آن دسترسی نداریم.

یکی از دغدغه‌هایی که در مسائل جویبار داده وجود دارد، عدم امکان دسترسی به تمام نقاط است. در واقع هم این مشکل وجود دارد که به تمام داده‌های قبلی دسترسی نداریم و هم این مشکل وجود دارد که هیچ اطلاعی از داده‌های آتی نداریم. در نتیجه یکی از تبعات این گونه از مسئله‌ی kــ مرکز، امکان انتخاب نقطه‌ای به عنوان مرکز برای یک دسته به‌طوری‌که در بین نقاط ورودی نیست. این گونه از مسئله‌ی kــ مرکز، معمولا تنها برای L_pــ متریک مطرح می‌شود یا حالتی که ما مجموعه‌ای از تمام نقاط فضا به انضمام فاصله‌هایشان را داشته باشیم. زیرا مرکز دسته‌ها ممکن است در هر نقطه از فضا قرار بگیرد و ما نیاز داریم که فاصله‌ی آن را از تمام نقاط بدانیم. نمونه‌ای از مسئله‌ی ۲ــ مرکز در حالت پیوسته، در شکل ۱ــ۳ نشان داده شده است. تعریف دقیق گونه‌ی جویبار داده‌ی مسئله‌ی kــ مرکز، در

[۱۷] Sublinear

زیر آمده است:



شکل ۱ـ۳: نمونه‌ای ازمسئله‌ی ۲ـ مرکز در حالت پیوسته

مسئله‌ی ۱ـ۳ *مجموعه‌ی U از نقاط فضای d بعدی داده شده است. زیرمجموعه $S \subseteq U$ با اندازه‌ی k را انتخاب کنید به‌طوری‌که عبارت زیر را کمینه کند:*

$$\max_{u \in U}\{\min_{s \in S} L_p(u, s)\}$$

از آنجایی که گونه‌ی جویبار داده و داده پرت مسئله‌ی kـ مرکز به علت داغ شدن مبحث داده‌های بزرگ[18]، به تازگی مورد توجه قرار گرفته است و نتایج به‌دست‌آمده قابل بهبود است، در این تحقیق سعی شده است که تمرکز بر روی این‌گونه‌ی خاص از مسئله باشد. همچنین در این پژوهش سعی می‌شود گونه‌های مسئله را برای انواع متریک‌ها و برای kهای کوچک نیز مورد بررسی قرار داد.

۱ـ۲ اهمیت موضوع

مسئله‌ی kـ مرکز و گونه‌های آن کاربردهای زیادی در داده‌کاوی دارند. این الگوریتم یکی از رایج‌ترین الگوریتم‌های مورد استفاده برای خوشه‌بندی محسوب می‌شود. به علت افزایش حجم داده‌ها و تولید داده‌ها در طول زمان، گونه‌ی جویبار داده‌ی مسئله در سال‌های اخیر مورد توجه قرار گرفته است. از طرفی مسئله‌ی kـ مرکز در بعضی مسائل مانند ارسال نامه از تعدادی مراکز پستی به گیرنده‌ها، نیاز دارد که تمام نامه‌ها را به دست گیرنده‌ها برساند و در نتیجه باید همه‌ی نقاط را پوشش دهد، ولی برای بعضی از مسائل تجاری، نیازی به پوشش تمام نقاط هدف نیست و از لحاظ اقتصادی به‌صرفه است که نقاط پرت را در نظر نگرفت. به طور مثال، Kmarts در سال ۲۰۱۰ اعلام کرده است که به ۸۸ درصد جمعیت

[18] Big data

آمریکا با فاصله‌ای حداکثر ۶ مایل، می‌تواند سرویس دهد. درصورتی‌که اگر این شرکت قصد داشت تمام جمعیت آمریکا را پوشش دهد، نیاز داشت تعداد شعب یا شعاع پوشش خود را به میزان قابل توجهی افزایش دهد که از لحاظ اقتصادی به‌صرفه نیست [۳]. علاوه بر دو گونه‌ی مطرح‌شده در این قسمت، گونه‌های دیگر از مسئله‌ی kـ مرکز در مرجع [۳] آمده است.

۱ـ۳ ادبیات موضوع

همان‌طور که ذکر شد مسئله‌ی kـ مرکز در حالت داده‌های پرت و جویبار داده، گونه‌های تعمیم‌یافته از مسئله‌ی kـ مرکز هستند و در حالت‌های خاص به مسئله‌ی kـ مرکز کاهش پیدا می‌کنند. مسئله‌ی kـ مرکز در حوزه‌ی مسائل ان‌پی‌ـ سخت[۱۹] قرار می‌گیرد و با فرض $P \neq NP$ الگوریتم دقیق با زمان چندجمله‌ای برای آن وجود ندارد. بنابراین برای حل کارای[۲۰] این مسائل از الگوریتم‌های تقریبی[۲۱] استفاده می‌شود.

برای مسئله‌ی kـ مرکز، دو الگوریتم تقریبی معروف وجود دارد. در الگوریتم اول، که به روش حریصانه[۲۲] عمل می‌کند، در هر مرحله بهترین مرکز ممکن را انتخاب می‌کند به طوری تا حد ممکن از مراکز قبلی دور باشد. این الگوریتم، الگوریتم تقریبی با ضریب تقریب ۲ ارائه می‌دهد. در الگوریتم دوم، با استفاده از مسئله‌ی مجموعه‌ی غالب کمینه[۲۳]، الگوریتمی با ضریب تقریب ۲ ارائه می‌گردد. همچنین ثابت شده است، که بهتر از این ضریب تقریب، الگوریتمی نمی‌توان ارائه داد مگر آن‌که $P = NP$ باشد.

۱ـ۴ اهداف تحقیق

در این پایان‌نامه مسئله‌ی kـ مرکز در حالت جویبار داده با داده‌های پرت در حالت‌های مختلف مورد بررسی قرار می‌گیرد و سعی خواهد شد که نتایج قبلی در این مسائل را از جنبه‌های مختلفی مورد بهبود دهد.

[۱۹] NP-hard
[۲۰] Efficient
[۲۱] Approximation Algorithm
[۲۲] Greedy
[۲۳] Dominating set

اولین مسئله‌ای که مورد بررسی قرار گرفته است، ارائه الگوریتمی تقریبی، برای مسئله‌ی ۱ ــ مرکز در حالت جویبار داده است به طوری که، نه تنها تمام نقاط ورودی را می‌پوشاند بلکه تضمین می‌کند که دایره‌ی بهینه‌ی جواب ۱ ــ مرکز برای نقاط ورودی را نیز بپوشاند. در تلاش اول، الگوریتم جدیدی با حافظه و زمان به‌روزرسانی $\mathcal{O}(\frac{d}{\epsilon})$ و با ضریب تقریب $\epsilon+$ ۱/۸ ، برای این مسئله ارائه گردید. با بررسی‌های بیش‌تر، الگوریتم دیگری با ضریب تقریب ۱/۶۹ ، با الگوریتمی کاملا متفاوت، با حافظه‌ی $\mathcal{O}(d)$ برای این مسئله ارائه گردید.

مسئله‌ی دومی که مورد بررسی قرار گرفت، مسئله‌ی ۱ ــ مرکز در حالت جویبار داده با داده‌های پرت است. در ابتدا الگوریتمی ساده با حافظه‌ی $\mathcal{O}(zd)$ و زمان به‌روزرسانی $\mathcal{O}(zd\log(z))$ با ضریب تقریب ۲ برای این مسئله ارائه گردید. در بررسی‌های بعدی، با استفاده از نتایج بدست آمده در قسمت قبل، برای z های کوچک، الگوریتمی با حافظه‌ی $\mathcal{O}(dz^d)$ با ضریب تقریب ۱/۶۹ برای این مسئله ارائه شد که ضریب تقریب بهترین الگوریتم موجود را، از ۱/۷۹ به ۱/۶۹ کاهش می‌دهد.

مسئله‌ی سومی که مورد بررسی قرار گرفت، مسئله‌ی ۲ ــ مرکز در حالت جویبار داده با داده‌های پرت است. بهترین الگوریتمی که در حال حاضر برای این مسئله وجود دارد، یک الگوریتم با ضریب تقریب $\epsilon +$ ۴ است [۱۹]. ما با ارائه الگوریتمی جدید، الگوریتمی با ضریب تقریب $\epsilon +$ ۱/۸ برای این مسئله ارائه دادیم که بهبودی قابل توجه محسوب می‌شود.

۱ ــ ۵ ساختار پایان‌نامه

این پایان‌نامه در پنج فصل به شرح زیر ارائه خواهد شد. در فصل دوم به بیان مفاهیم و تعاریف مرتبط با موضوعات مورد بررسی در بخش‌های دیگر خواهیم پرداخت. سعی شده، تا حد امکان با زبان ساده و ارجاع‌های مناسب، پایه‌ی لازم را برای فصول بعدی در این فصل فراهم آوریم. فصل سوم این پایان‌نامه شامل مطالعه و بررسی کارهای پیشین انجام شده مرتبط با موضوع این پایان‌نامه خواهد بود. این فصل در سه بخش تنظیم گردیده است. در بخش اول، مسئله‌ی kــ مرکز در حالت ایستا مورد بررسی قرار می‌گیرد. در بخش دوم، حالت جویبار داده‌ی مسئله و مجموعه هسته‌های[۲۴] مطرح برای این مسئله مورد بررسی قرار می‌گیرد. در نهایت، در بخش سوم، مسئله‌ی kــ مرکز با داده‌های پرت مورد بررسی قرار می‌گیرد.

[۲۴] coreset

در فصل چهارم، نتایج جدیدی که در این پایان‌نامه به دست آمده است، ارائه می‌شود. این نتایج شامل الگوریتم‌های جدید برای مسئله‌ی ۱ـ مرکز در حالت جویبار داده، مسئله‌ی ۱ـ مرکز با داده‌های پرت در حالت جویبار داده و مسئله‌ی دو مرکز در حالت جویبار داده با داده‌های پرت می‌شود.

در فصل پنجم، به جمع‌بندی کارهای انجام شده در این پژوهش و ارائه‌ی پیشنهادهایی برای انجام کارهای آتی و تعمیم‌هایی که از راه‌حل ارائه شده وجود دارد، خواهیم پرداخت.

فصل ۲

مفاهیم اولیه

در این فصل به تعریف و بیان مفاهیم پایه‌ای مورد استفاده در فصل‌های بعد می‌پردازیم. با توجه به مطالب مورد نیاز در فصل‌های آتی، مطالب این فصل به سه بخش، مسائل ان‌پی-سخت، الگوریتم‌های تقریبی و الگوریتم‌های جویبارداده تقسیم می‌شود.

۲-۱ مسائل ان‌پی-سخت

یکی از اولین سوال‌های بنیادی مطرح در علم کامپیوتر، اثبات **عدم حل پذیری** بعضی از مسائل است. به عنوان نمونه، می‌توان از دهمین مسئله‌ی هیلبرت[۱] در کنگره‌ی ریاضی یاد کرد. هیلبرت این مسئله را این‌گونه بیان کرد: "فرآیندی طراحی کنید که در تعداد متناهی گام بررسی کند که آیا یک چندجمله‌ای، ریشه‌ی صحیح[۲] دارد یا خیر". با مدل محاسباتی که به‌وسیله‌ی تورینگ[۳] ارائه شد، این مسئله معادل پیدا کردن الگوریتمی برای این مسئله است که اثبات می‌شود امکان‌پذیر نیست [۴]. برخلاف مثال بالا، عمده‌ی مسائل علوم کامپیوتر از نوع بالا نیستند و برای طیف وسیعی از آن‌ها، الگوریتم‌های پایان‌پذیر وجود دارد. بیش‌تر تمرکز علوم کامپیوتر هم بر روی چنین مسائلی است.

اگر چه برای اکثر مسائل، الگوریتمی پایان‌پذیر وجود دارد، اما وجود چنین الگوریتمی لزومی بر حل

[۱] Hilbert
[۲] Integral root
[۳] Touring

شدن چنین مسائلی نیست. در عمل، علاوه بر وجود الگوریتم، میزان کارامدی[۴] الگوریتم نیز مطرح می‌گردد. به طور مثال، اگر الگوریتم حل یک مسئله مرتبه‌ی بالا یا نمایی داشته باشد، الگوریتم ارائه شده برای آن مسئله برای ورودی‌های نسبتا بزرگ قابل اجرا نیست و نمی‌توان از آن‌ها برای حل مسئله استفاده کرد. برای تشخیص و تمیز کارآمدی الگوریتم‌های مختلف و همچنین میزان سختی مسائل در امکان ارائه‌ی الگوریتم‌های کارآمد یا غیرکارآمد، نظریه‌ی پیچیدگی[۵]، دسته‌بندی‌های مختلفی برای سختی مسائل و حل‌پذیری آن‌ها ارائه داده است تا بتوان به‌طور رسمی[۶] در مورد این معیارها صحبت کرد. برای دسته‌بندی مسائل در نظریه‌ی پیچیدگی، ابتدا آن‌ها را به صورت تصمیم‌پذیر بیان می‌کنند.

مسئله‌ی ۲ــ۱ (مسائل تصمیم‌گیری)[۷] *به دسته‌ای از مسائل گفته می‌شود که پاسخ آن‌ها تنها بله یا خیر است.*

به عنوان مثال، اگر بخواهیم مسئله‌ی ۱ ــ مرکز در فضای $\mathbb{R}^d$ را به صورت تصمیم پذیر بیان کنیم، به مسئله‌ی زیر می‌رسیم:

مسئله‌ی ۲ــ۲ (نسخه‌ی تصمیم پذیر ۱ ــ مرکز) *مجموعه‌ی نقاط در فضا $\mathbb{R}^d$ و شعاع r داده شده است، آیا دایره‌ای به شعاع r وجود دارد که تمام نقاط را بپوشاند؟*

در نظریه‌ی پیچیدگی، می‌توان گفت عمده‌ترین دسته‌بندی موجود، دسته‌بندی مسائل تصمیم‌گیری به مسائل پی(P) و ان‌پی(NP) است. رده‌ی مسائل P، شامل تمامی مسائل تصمیم‌گیری است که راه‌حل چندجمله‌ای برای آن‌ها وجود دارد. از طرفی رده‌ی مسائل NP، شامل تمامی مسائل تصمیم‌گیری است که در زمان چندجمله‌ای قابل صحت‌سنجی[۸] اند. تعریف صحت‌سنجی در نظریه پیچیدگی، یعنی اگر جواب مسئله‌ی تصمیم‌گیری بله باشد، می‌توان اطلاعات اضافی با طول چندجمله‌ای ارائه داد، که در زمان چندجمله‌ای از روی آن، می‌توان جواب بله الگوریتم را تصدیق نمود. به طور مثال، برای مسئله‌ی ۱ ــ مرکز، کافی است به عنوان تصدیق جواب بله، مرکز دایره‌ی پوشاننده، ارائه داده شود. در این صورت، می‌توان با مرتبه‌ی خطی بررسی نمود که تمام نقاط داخل این دایره قرار می‌گیرند یا نه. برای مطالعه‌ی بیش‌تر و تعاریف دقیق‌تر می‌توان به مرجع [۴] مراجعه نمود.

[۴] Efficiency
[۵] Complexity theory
[۶] Formal
[۷] Decision problems
[۸] Verifiable

همان‌طور که می‌دانید درستی یا عدم درستی $P \subset NP$ از جمله معروف‌ترین مسائل حل نشده[۹] در نظریه پیچیدگی است. حدس بسیار قوی وجود دارد که $P \neq NP$ و بسیاری از مسائل، با این فرض حل می‌شوند و در صورتی که زمانی، خلاف این فرض اثبات گردد، آن‌گاه قسمت عمده‌ای از علوم کامپیوتر زیر سوال می‌رود.

در نظریه‌ی پیچیدگی، برای دسته‌بندی مسائل، یکی از روش‌های دسته‌بندی کاهش چندجمله‌ای[۱۰] مسائل به یک‌دیگر است.

تعریف ۲ـ۱ *می‌گوییم مسئله‌ی A در زمان چندجمله‌ای به مسئله‌ی B کاهش می‌یابد، اگر وجود داشته باشد الگوریتم چندجمله‌ای C که به ازای هر ورودی α برای مسئله‌ی A، یک ورودی β در زمان چندجمله‌ای برای مسئله‌ی B بسازد، به طوری که A، α را می‌پذیرد اگر و تنها اگر B، β را بپذیرد. در این‌جا منظور از پذیرفتن جواب بله به ورودی است.*

از این به بعد برای سادگی به جای **کاهش چندجمله‌ای** از واژه‌ی **کاهش** استفاده می‌کنیم. در پی جست‌جوهایی که برای برابری دسته‌ی پی و ان‌پی صورت گرفت، مجموعه‌ای از مسائل که عمدتا داخل ان‌پی هستند استخراج گردید که اگر ثابت شود یکی از آن‌ها متعلق به پی است، آن‌گاه تمام مسائل دسته‌ی ان‌پی متعلق به پی خواهند بود و در نتیجه $P = NP$ خواهد بود. به این مجموعه مسائل ان‌پی‌ـ سخت می‌گویند. در واقع مسائل این دسته، مسائلی هستند که تمام مسائل داخل دسته‌ی ان‌پی، به آن‌ها کاهش می‌یابند.

کوک و لوین در قضیه‌ای به نام **قضیه‌ی کوک‌ـ لوین** ثابت کردند مسئله‌ی صدق‌پذیری[۱۱] یک مسئله‌ی ان‌پی‌ـ سخت است [۴]. با پایه قرار دادن این اثبات و استفاده از تکنیک کاهش، اثبات ان‌پی‌ـ سخت بودن سایر مسائل، بسیار ساده‌تر گردید. در ادامه مسئله‌ی پوشش رأسی[۱۲] را تعریف می‌کنیم.

۲ـ۱ـ۱ پوشش رأسی

در این پایان‌نامه، از این مسئله به عنوان مسئله‌ی پایه برای اثبات ان‌پی‌ـ سخت بودن مسئله‌ی kـ مرکز استفاده می‌شود. تعریف این مسئله مطابق زیر است:

---

[۹] Open problem

[۱۰] Polynomial Reduction

[۱۱] Satisfiability problem

[۱۲] Vertex Coverage

تعریف ۲ــ۲ *گراف بدون جهت $G(V,E)$ داده شده است. هدف مسئله پیدا کردن مجموعه‌ی $S \subset V$ با کم‌ترین تعداد اعضا است به طوری که هر رأس $v \in V$ در یکی از شرایط زیر صدق کند:*

- $v \in S$
- *وجود دارد رأسی $u \in S$ به طوری $(v,u) \in E$*

به عبارت ساده‌تر هر رأسی یا خودش یا یکی از همسایگانش داخل مجموعه‌ی S قرار دارد.

نسخه‌ی تصمیم‌گیری این مسئله به این گونه تعریف می‌شود که آیا گراف داده‌شده دارای پوشش رأسی با اندازه‌ی k است یا نه.

قضیه‌ی ۲ــ۱ *مسئله‌ی پوشش رأسی، یک مسئله‌ی ان‌پی‌ــ سخت است.*

اثبات. *برای مشاهده‌ی اثبات ان‌پی‌ــ سخت بودن مسئله‌ی پوشش رأسی، نیاز به زنجیره‌ای از مسائل که از مسئله‌ی صدق‌پذیری شروع می‌شود است، به طوری هر عضو از این زنجیره، به عضو بعدی در زمان چندجمله‌ای کاهش می‌یابد و در نهایت نتیجه می‌شود که مسئله‌ی صدق‌پذیری در زمان چندجمله‌ای به مسئله‌ی پوشش رأسی کاهش می‌یابد و در نتیجه چون مسئله‌ی صدق‌پذیری یک مسئله‌ی ان‌پی‌ــ سخت است، بنابراین مسئله‌ی پوشش رأسی نیز ان‌پی‌ــ سخت خواهد بود. برای مطالعه‌ی روند اثبات به مرجع [۴] مراجعه کنید.*

□

۲ــ۲ الگوریتم‌های تقریبی

تا این‌جا با رده‌بندی مسائل به دو دسته‌ی پی و ان‌پی آشنا شدیم. نه تنها مسائل ان‌پی، بلکه بعضی از مسائل پی نیز دارای الگوریتم کارامدی نیستند. در عمل، یک الگوریتم چندجمله‌ای با مرتبه‌ی بیش از ۳، یک الگوریتم ناکارآمد محسوب می‌شود. به طور مثال هنوز الگوریتم کارامدی برای فهمیدن این‌که یک عدد اول است یا نه پیدا نشده است، با اینکه الگوریتم ارائه شده یک الگوریتم چندجمله‌ای است. عمده‌ی مسائل کاربردی مطرح در دنیای واقع، یا متعلق به دسته‌ی ان‌پی هستند و در نتیجه راه‌حل چندجمله‌ای ندارند، یا اگر راه‌حل چند جمله‌ای داشته باشند، مرتبه‌ی چندجمله‌ای بالاست و در نتیجه

راه‌حل کارآمدی محسوب نمی‌گردد. یکی از رویکردهای رایج در برابر چنین مسائلی، صرف نظر کردن از دقت راه‌حل‌هاست. به طور مثال راه‌حل‌های مکاشفه‌ای[۱۳] گوناگونی برای مسائل مختلف به خصوص مسائل ان‌پی بیان شده است. این راه‌حل‌ها بدون این‌که تضمین کنند راه‌حل خوبی ارائه می‌دهند یا حتی جوابشان به جواب بهینه نزدیک است، اما با معیارهایی سعی در بهینه عمل کردن دارند و تا حد ممکن سعی در ارائه‌ی جواب بهینه یا نزدیک بهینه دارند. اما در عمل، تنها برای دسته‌ای از کاربردها پاسخ قابل قبولی ارائه می‌دهند.

مشکل عمده‌ی راه‌حل‌های مکاشفه‌ای، عدم امکان استفاده از آن‌ها برای تمام کاربردها است. بنابراین در رویکرد دوم که اخیرا نیز مطرح شده است و عمر کمی دارد، سعی در ارائه‌ی الگوریتم‌های مکاشفه‌ای شده است که تضمین می‌کنند اختلاف زیادی با الگوریتمی که جواب بهینه می‌دهد، نداشته باشند. در واقع این الگوریتم‌ها همواره و در هر شرایطی، تقریبی از جواب بهینه را ارائه می‌دهند. به چنین الگوریتم‌هایی، ***الگوریتم‌های تقریبی***[۱۴] *می‌گویند. علت اصلی این نام‌گذاری، تقریب زدن جواب الگوریتم بهینه است. ضریب تقریب یک الگوریتم تقریبی، به حداکثر نسبت جواب الگوریتم تقریبی به جواب بهینه گفته می‌شود.*

الگوریتم‌های تقریبی تنها به علت محدودیت کارایی الگوریتم‌هایی که جواب بهینه می‌دهند، مورد استفاده قرار نمی‌گیرند. هر نوع محدودیتی ممکن است، استفاده از الگوریتم تقریبی را نسبت به الگوریتمی که جواب بهینه می‌دهد، مقرون به صرفه کند. به طور مثال از جمله عوامل دیگری که ممکن است باعث این انتخاب شود، کاهش میزان حافظه‌ی مصرفی باشد. برای طیف وسیعی از مسائل، کمبود حافظه، باعث می‌شود الگوریتم‌هایی با حافظه‌ی مصرفی کم‌تر طراحی شود که به دقت الگوریتم‌های بهینه عمل نمی‌کند اما می‌تواند با مصرف کم‌تر به تقریبی از جواب بهینه دست یابد. معمولا چنین الگوریتم‌هایی حافظه‌ی مصرفی از مرتبه‌ی زیرخطی[۱۵] دارند و به همین دلیل برای داده‌های حجیم بسیار کاربرد دارند.

[۱۳] Heuristic
[۱۴] Approximation Algorithm
[۱۵] Sublinear

مسئله	کران پایین تقریب‌پذیری
پوشش رأسی	۱٫۳۶۰۶ [۵]
kــ مرکز	۲[۶]
kــ مرکز در فضای اقلیدسی	۱٫۸۲۲[۱۰]
۱ ــ مرکز در حالت جویبار داده	$\frac{۱+\sqrt{۲}}{۲}$ [۷]
kــ مرکز با نقاط پرت و نقاط اجباری	۳[۳]

جدول ۲ ــ ۱: نمونه‌هایی از ضرایب تقریب برای مسائل بهینه‌سازی

۲ ــ ۲ ــ ۱ میزان تقریب‌پذیری مسائل

همان‌طور که تا این‌جا دیدیم، یکی ار راه‌کارهایی که برای کارآمد کردن راه‌حل ارائه شده برای یک مسئله وجود دارد، استفاده از الگوریتم‌های تقریبی برای حل آن مسئله است. یکی از عمده‌ترین دغدغه‌های مطرح در الگوریتم‌های تقریبی کاهش ضریب تقریب است. حتی در بعضی از موارد حتی امکان ارائه‌ی الگوریتم تقریبی با ضریبی ثابت نیز وجود ندارد. به طور مثال، همان‌طور که در فصل کارهای پیشین بیان خواهد شد، الگوریتم تقریبی با ضریب تقریب کم‌تر از ۲ ، برای مسئله‌ی kــ مرکز وجود ندارد مگر اینکه $P = NP$ باشد. برای مسائل مختلف، معمولا می‌توان کران پایینی برای میزان تقریب‌پذیری آن‌ها ارائه داد. در واقع برای مسائل ان‌پی، علاوه بر این الگوریتم کارآمدی وجود ندارد، بعضا الگوریتم تقریبی برای حل آن با ضریبی تقریب کم و نزدیک به یک نیز وجود ندارد. در جدول ۲ ــ ۱ از میزان تقریبی مسائل مختلفی که در این پایان‌نامه مورد استفاده قرار می‌گیرد ببینید.

۲ ــ ۳ الگوریتم‌های جویبارداده

فصل ۳

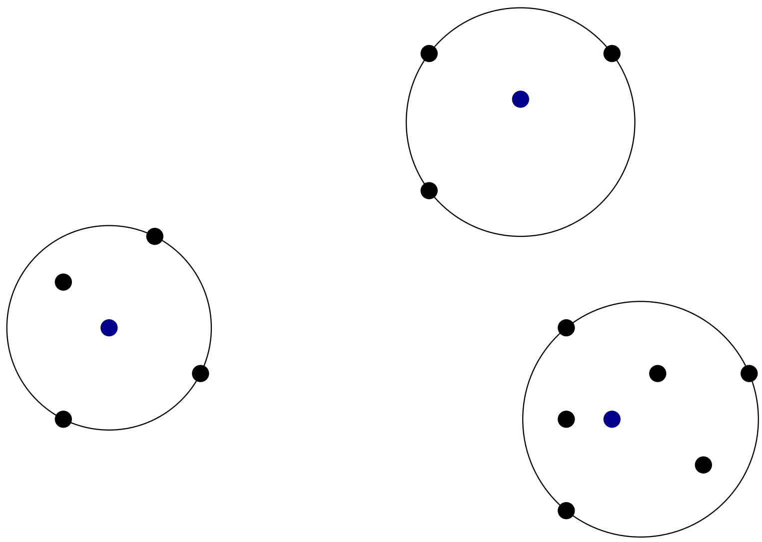
کارهای پیشین

در این فصل کارهای پیشین انجام‌شده روی مسئله‌ی kـ مرکز، در سه بخش مورد بررسی قرار می‌گیرد. در بخش اول، مسئله‌ی kـ مرکز مورد بررسی قرار می‌گیرد. در بخش دوم، حالت جویبار داده‌ی مسئله و مجموعه هسته‌های مطرح برای این مسئله مورد بررسی قرار می‌گیرد. در نهایت، در بخش سوم، مسئله‌ی kـ مرکز با داده‌های پرت مورد بررسی قرار می‌گیرد.

۳ـ۱ kـ مرکز در حالت ایستا

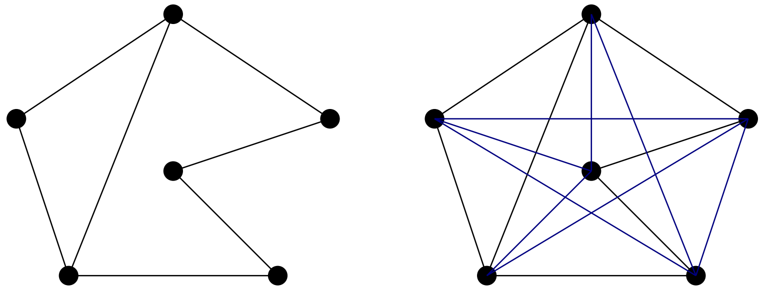
مسئله‌ی kـ مرکز به عنوان مسئله‌ی شناخته شده در علوم کامپیوتر مطرح است. این مسئله، در واقع یک مسئله‌ی بهینه‌سازی است که سعی در کاهش بیش‌ترین فاصله نقاط از مرکز دسته‌ها را دارد. سختی اصلی این مسئله در انتخاب مرکز دسته‌هاست. زیرا اگر بتوانیم مرکز دسته‌ها را به درستی تشخیص دهیم، کافی است هر نقطه را به دسته‌ای که نزدیک‌ترین مرکز را دارد، تخصیص دهیم. به وضوح چنین تخصیصی بهینه‌ترین تخصیص ممکن است. نمونه‌ای از این تخصیص را در شکل ۳ـ۱ نشان داده شده است.

در سال ۱۹۷۹ ، نه تنها اثبات گردید که این مسئله در حالت کلی یک مسئله‌ی ان‌پی‌ـ سخت است [۸]، بلکه ثابت شده است که این مسئله در صفحه‌ی دو بعدی و با متریک اقلیدسی نیز ان‌پی‌ـ سخت است [۹]. فراتر از این، ثابت شده است که برای مسئله‌ی kـ مرکز با متریک دلخواه هیچ الگوریتم تقریبی با ضریب تقریب بهتر از ۲ وجود ندارد. ایده‌ی اصلی این کران پایین، کاهش مسئله‌ی پوشش رأسی، به مسئله‌ی kـ مرکز است. برای چنین کاهشی کافی است، از روی گراف اصلی که می‌خواهیم



شکل ۳ــ۱: نمونه‌ای از تخصیص نقاط به ازای مراکز آبی رنگ.

کوچک‌ترین مجموعه‌ی پوشش رأسی را در آن پیدا کنیم، یک گراف کامل بسازیم به طوری که به ازای هر یال در گراف اصلی، یک یال با وزن یک و به ازای هر یال که در گراف اصلی وجود ندارد، یک یال با وزن ۲ قرار می‌دهیم. نمونه‌ای از چنین تبدیلی را در شکل ۳ــ۲ می‌توانید مشاهده کنید. حال اگر الگوریتمی بتواند مسئله‌ی kــ مرکز را با ضریب تقریب بهتر از ۲ حل نماید، آن‌گاه گراف جدید دارای یک kــ مرکز با شعاع کم‌تر از ۲ است، اگر و تنها اگر گراف اصلی دارای یک پوشش رأسی با اندازه‌ی k باشد. برای متریک L_2 یا فضای اقلیدسی[1] نیز ثابت شده است برای مسئله‌ی kــ مرکز، الگوریتم تقریبی با ضریب تقریب بهتر از ۱/۸۲۲ وجود ندارد [۱۰].



شکل ۳ــ۲: نمونه‌ای از تبدیل یک گراف ورودی مسئله‌ی پوشش رأسی به یک ورودی مسئله‌ی kــ مرکز(در گراف سمت چپ، یال‌های سیاه وزن ۱ و یال‌های آبی، وزن ۲ دارند)

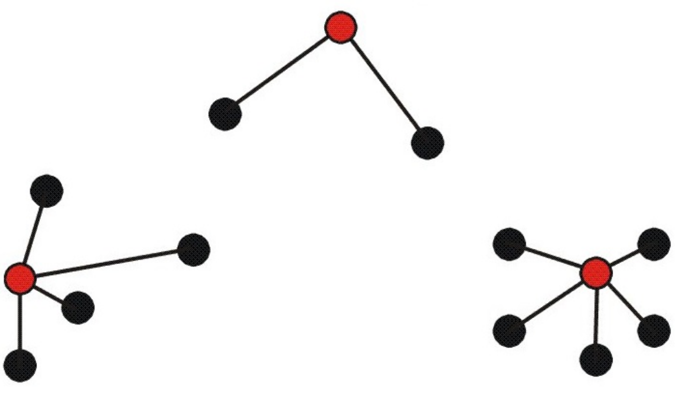
[1] Euclidean space

یکی از اولین الگوریتم‌های تقریبی برای مسئله‌ی kـ مرکز به وسیله‌ی گنزالز[۲] *ارائه شده است* [۱۱]. *این الگوریتم یک الگوریتم تقریبی با ضریب ۲ است و در زمان $\mathcal{O}(kn)$ قابل اجراست. الگوریتم گنزالز، از نظر روش برخورد با مسئله، یک الگوریتم حریصانه*[۳] *محسوب می‌شود. برای این‌که عملکرد الگوریتم گنزالز را درک کنید، نیاز به تعریف فاصله‌ی یک نقطه از یک مجموعه نقطه داریم.*

تعریف ۳ـ۱ *فاصله‌ی نقطه‌ی v از مجموعه‌ای ناتهی از نقاط S را برابر فاصله‌ی نقطه‌ای درون S از v تعریف می‌کنیم، به گونه‌ای که از تمام نقاط S به v نزدیک‌تر باشد. در واقع داریم:*

$$d(v, S) = \min_{u \in S}\{d(u, v)\}$$

روش اجرای این الگوریتم به این گونه است که در ابتدا یک نقطه‌ی دلخواه را به عنوان مرکز دسته‌ی اول در نظر می‌گیرد. سپس دورترین نقطه از آن را به عنوان مرکز دسته‌ی دوم در نظر می‌گیرد. در هر مرحله، دورترین نقطه از مرکز مجموعه دسته‌های موجود را به عنوان مرکز دسته‌ی جدید به مجموعه مراکز دسته‌ها اضافه می‌کند. با اجرای الگوریتم تا k مرحله، مراکز دسته‌ها انتخاب می‌شود. حال اگر هر نقطه را به نزدیک‌ترین مرکز انتخابی تخصیص دهیم، می‌توان نشان داد که شعاع بزرگ‌ترین دسته، حداکثر دو برابر شعاع بهینه برای مسئله‌ی kـ مرکز است. فدر[۴] *و سایرین، زمان اجرای الگوریتم گنزالز را برای هر L_pـ متریک به مرتبه‌ی $\mathcal{O}(n \log k)$ بهبود بخشیدند. نمونه‌ای از اجرای الگوریتم گنزالز، در شکل ۳ـ۳ نشان داده شده است.*



شکل ۳ـ۳: نمونه‌ای از حل مسئله‌ی ۳ـ مرکز با الگوریتم گنزالز

[۲] Gonzalez
[۳] Greedy
[۴] Feder

تا به اینجا، تنها بر روی حالت کلی مسئله‌ی kـ مرکز صحبت شد و تنها محدودیتی که مورد توجه قرار گرفت، متریک مطرح برای فاصله‌ی نقاط بوده است. در ادامه به بررسی، حالاتی از مسئله‌ی k ـ مرکز، که k تعداد دسته‌ها یا d ابعاد فضا ثابت باشند می‌پردازیم. آگاروال[۵] و سایرین الگوریتمی دقیق با زمان اجرای $n^{O(k^{1-\frac{1}{d}})}$ برای مسئله kـ مرکز در فضای L_pـ متریک با ابعاد ثابت d ارائه داده‌اند [۱۲]. قابل توجه است که اگر d ثابت نباشد، مسئله‌ی kـ مرکز حتی برای متریک اقلیدسی(L_2ـ متریک) با تعداد دسته‌ی ثابت $k \geqslant 2$، ان‌پی‌ـ سخت است [۱۳].

علاوه بر حالاتی که ابعاد فضا یا تعداد دسته‌ها ثابت‌اند، مسئله‌ی kـ مرکز برای حالتی که مقادیر k و d کوچک هستند، مورد بررسی قرار گرفته‌اند و الگوریتم‌های بهینه‌تری از الگوریتم‌های کلی برای این حالت‌های خاص ارائه شده است. به طور مثال، برای مسئله‌ی ۱ ـ مرکز در فضای اقلیدسی با ابعاد ثابت، الگوریتم خطی با زمان اجرای $O((d+1)!n)$ وجود دارد [۱۴]. الگوریتم ارائه شده بر پایه‌ی دو نکته‌ی اساسی بنا شده است. اول اینکه کره‌ی بهینه را می‌توان با حداکثر $d+1$ نقطه‌ی واقع در پوسته‌ی کره‌ی بهینه مشخص نمود و دوم اینکه اگر نقاط ورودی را با ترتیبی تصادفی پیمایش کنیم احتمال اینکه نقطه‌ی پیمایش شده جزء نقاط مرزی باشد از مرتبه‌ی $O(\frac{d}{n})$ است که با توجه به ثابت بودن d این احتمال برای nهای بزرگ کوچک محسوب می‌شود و زمان اجرای الگوریتم، به طور میانگین خطی خواهد بود.

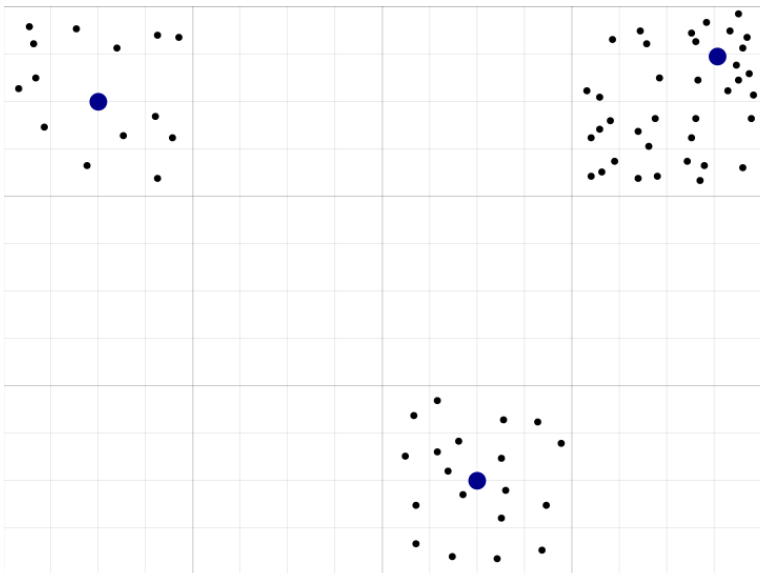
در صفحه‌ی اقلیدسی برای مسئله‌ی ۲ ـ مرکز، بهترین الگوریتم را چن[۶] با زمان اجرای $O(c\log^2 n \log^2 \log n)$ و حافظه‌ی $O(n)$ ارائه داده است [۱۵]. برای فضای سه‌بعدی اقلیدسی نیز آگاروال و سایرین، الگوریتمی با متوسط زمان اجرای $O(n^3 \log^8 n)$ ارائه داده است [۱۶].

۳ـ۲ kـ مرکز در حالت جویبار داده

در مدل جویبار داده، مشکل اصلی عدم امکان نگه‌داری تمام داده‌ها در حافظه است و باید سعی شود تنها داده‌هایی که ممکن است در ادامه مورد نیاز باشد را، نگه‌داریم. یکی از راه‌های رایج برای این کار نگه‌داری مجموعه‌ای از نقاط (نه لزوما زیرمجموعه‌ای از نقاط ورودی) به عنوان نماینده‌ی نقاط به طوری که جواب مسئله‌ی kـ مرکز برای آن‌ها منطبق با جواب مسئله‌ی kـ مرکز برای نقاط اصلی باشد (با تقریب قابل قبولی). به چنین مجموعه‌ای مجموعه‌ی هسته‌ی نقاط گفته می‌شود.

[۵] Agarwal
[۶] Chan

بهترین مجموعه هسته‌ای که برای مسئله‌ی kـ مرکز ارائه شده است، روش ارائه‌شده به وسیله‌ی ضرابی‌زاده برای نگه‌داری یک ϵـ هسته با حافظه‌ی $O(\frac{k}{\epsilon^d})$ برای L_pـ متریک‌ها است [۱۷]. در روش ارائه شده، از چند ایده‌ی ترکیبی استفاده شده است. در ابتدا، الگوریتم با استفاده از یک الگوریتم تقریبی با ضریب ثابت، یک تقریب از جواب بهینه به دست می‌آورد. به طور مثال با استفاده از الگوریتم گنزالز، یک دو تقریب از شعاع بهینه به علاوه‌ی مرکز دسته‌های پیدا شده را به دست می‌آورد. حال کافی است که با طول شعاع الگوریتم تقریبی، حول هر مرکز به دست آمده، یک توری با $O(\frac{1}{\epsilon})$ شبکه‌بندی در هر بعد تشکیل داد و چون هر نقطه در حداقل یکی از توری‌ها قرار می‌گیرد، می‌توان با حداکثر ϵ تقریب در جواب نهایی نقاط را به نقاط شبکه‌بندی توری گرد نمود. با این کار، دیگر ما نیازی به نگهداری تمام نقاط ورودی نداشته و تنها نیاز به نگهداری نقاط شبکه‌بندی توری داریم. با این روش می‌توان به یک ϵـ هسته برای مسئله‌ی kـ مرکز رسید. نکته‌ی اساسی برای سازگار سازی روش ارائه‌شده با مدل جویبار داده‌ی تک‌گذره استفاده از روش دوبرارسازی[۷] رایج در الگوریتم‌های جویبار داده است. نمونه‌ای از اجرای الگوریتم ضرابی‌زاده را در شکل ۳ـ۴ نشان داده شده است. برای دیدن اثبات‌ها و توضیح بیش‌تر در مورد روش ارائه شده می‌توانید به مرجع [۱۷] مراجعه کنید.



شکل ۳ـ۴: نمونه‌ای از توری‌بندی الگوریتم ضرابی‌زاده (نقاط آبی، مراکز به دست آمده از الگوریتم تقریبی است). پس از توری‌بندی کافی است برای هر کدام از خانه‌های شبکه‌بندی، تنها یکی را به عنوان نماینده در نظر بگیریم.

[۷] Doubling

از جمله مشکلات وارده به الگوریتم ضرابی زاده، وابستگی اندازه‌ی مجموعه‌ی هسته به ابعاد فضا است. بنابراین هسته‌ی ارائه شده به وسیله‌ی ضرابی‌زاده را نمی‌توان برای ابعاد بالا مورد استفاده قرار داد. از طرفی، حساب کردن جواب از روی هسته در زمان چندجمله‌ای بر اساس k و d قابل انجام نیست. با توجه با موارد گفته شده، برای قابل استفاده شدن الگوریتم‌ها برای ابعاد بالا، الگوریتم‌هایی ارائه می‌شود که ضریب تقریب بدتری دارند، اما میزان حافظه‌ی مصرفی یا اندازه‌ی مجموعه هسته‌ی آن‌ها چندجمله‌ای بر اساس d و k و $\log n$ باشد. به چنین الگوریتم‌هایی الگوریتم‌های جویبار داده برای ابعاد بالا گفته می‌شود.

اولین الگوریتم ارائه شده برای ابعاد بالا، الگوریتمی با ضریب تقریب ۸ و حافظه‌ی مصرفی $\mathcal{O}(dk)$ است [۱۸]. پس از آن، گوها[۸]، به طور موازی با مک‌کاتن[۹] و سایرین، الگوریتمی با ضریب تقریب $(۲+\epsilon)$ با حافظه‌ی $\mathcal{O}(\frac{dk}{\epsilon}\log\frac{۱}{\epsilon})$ برای مسئله‌ی kـ مرکز در هر فضای متریکی ارائه دادند [۱۹، ۲۰]. در سال ۲۰۱۴، اهن[۱۰] و سایرین، الگوریتمی با همین ضریب تقریب و حافظه‌ی $\mathcal{O}((k+۳)!۲^k\frac{d}{\epsilon})$ ارائه داده‌اند که برای kهای ثابت، حافظه را از مرتبه‌ی $\mathcal{O}(\log\frac{۱}{\epsilon})$ کاهش می‌دهد [۲۱].

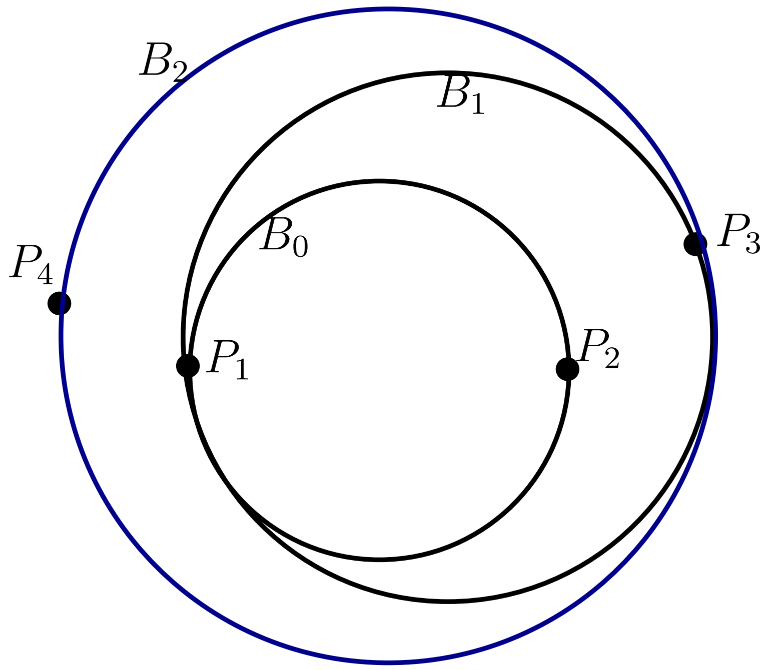
تا به اینجا ما به بررسی مسئله‌ی kـ مرکز در حالت جویبار داده بدون محدودیت خاصی پرداختیم. برای حالت‌های خاص k، به خصوص ۱ و ۲، مسئله‌ی kـ مرکز با متریک اقلیدسی مورد توجه زیادی قرار گرفته است و راه‌حل‌های بهینه‌تری نسبت به حالت کلی برای آن‌ها پیشنهاد شده است. به طور مثال، می‌توان یک هسته با اندازه‌ی $\mathcal{O}(\frac{۱}{\epsilon^{\frac{d-۱}{۲}}})$، با استفاده از نقاط حدی[۱۱] در تعداد مناسبی جهت، به صورت جویبار داده ساخت.

همان‌طور که برای الگوریتم ضرابی زاده مطرح شد، مشکل عمده‌ی مجموعه هسته‌ی ارائه‌شده، وابستگی حافظه‌ی مصرفی آن به d است. در راستای حل این مشکل، ضرابی‌زاده و سایرین [۲۲] برای ابعاد بالا و متریک اقلیدسی، الگوریتمی با ضریب تقریب ۱/۵ و حافظه‌ی مصرفی $\mathcal{O}(d)$ ارائه دادند. در واقع در الگوریتم آن‌ها، در هر لحظه تنها یک مرکز و یک شعاع را نگه داشته می‌شود، که کم‌ترین حافظه‌ی ممکن برای مسئله‌ی ۱ـ مرکز است. الگوریتم ارائه شده نقطه‌ی اول را به عنوان مرکز با شعاع صفر در نظر می‌گیرد. با فرا رسیدن هر نقطه‌ی جدید، اگر نقطه‌ی مد نظر، داخل کره‌ی فعلی بیفتد که بدون هیچ تغییری ادامه داده و در صورتی که بیرون کره‌ی فعلی بیفتد، آن را با کوچک‌ترین کره‌ای که

[۸] Guha
[۹] McCutchen
[۱۰] Ahn
[۱۱] Extreme points

نقطه‌ی جدید به علاوه‌ی کره‌ی قبلی را به طور کامل می‌پوشاند، جایگزین می‌کند.

به وضوح در هر لحظه کره‌ی ساخته شده تمام نقاطی که تا کنون در جویبار داده آمده است را می‌پوشاند. از طرفی ثابت می‌شود شعاع کره در هر لحظه، حداکثر ۱/۵ ـ برابر شعاع کره‌ی بهینه است. نکته‌ی قابل توجه در مورد این الگوریتم، نگه‌داری کم‌ترین حافظه‌ی ممکن برای مسئله‌ی ۱ ـ مرکز است. زیرا تنها یک مرکز و یک شعاع نگه می‌دارد. نمونه‌ای از اجرای الگوریتم را برروی چهار نقطه می‌توان در شکل ۳ـ۵ دید. برای اثبات کامل‌تر می‌توانید به مرجع [۲۲] مراجعه کنید.



شکل ۳ـ۵: نمونه‌ای از اجرای الگوریتم ضرابی‌زاده بر روی چهار نقطه $P_1 \cdots P_4$ که به ترتیب اندیس در جویبار داده فرا می‌رسند و دایره‌های $B_0 \cdots B_2$ دایره‌هایی که الگوریتم به ترتیب نگه می‌دارد.

در ادامه، آگاروال و سایرین [۷] الگوریتمی تقریبی با حافظه‌ی مصرفی $\mathcal{O}(d)$ ارائه دادند. در الگوریتم ارائه شده، ضریب تقریب، برابر $\frac{1+\sqrt{3}}{2}$ تخمین زده شد، اما با تحلیل دقیق‌تری که چن و سایرین [۲۳] بر روی همان الگوریتم انجام دادند، مشخص شد همان الگوریتم دارای ضریب تقریب ۱/۲۲ است.

الگوریتم آگاروال از الگوریتم کلارکسون و سایرین [۲۴] به عنوان یک زیرالگوریتم استفاده می‌کند. الگوریتم کلارکسون، الگوریتمی کاملا مشابه الگوریتم گنزالز است و به این گونه عمل می‌کند که در ابتدا یک نقطه دلخواه را به عنوان نقطه‌ی اول انتخاب می‌کند، سپس دورترین نقطه از نقطه‌ی اول را به عنوان دومین نقطه انتخاب می‌کند. ازین به بعد در هر مرحله، نقطه‌ای که از نقاط انتخاب شده‌ی قبلی

بیش‌ترین فاصله را دارد به عنوان نقطه‌ی جدید انتخاب می‌کند. اگر این الگوریتم را تا $O(\frac{1}{\epsilon})$ مرحله ادامه بدهیم، به مجموعه‌ای با اندازه‌ی $O(\frac{1}{\epsilon})$ خواهیم رسید که کلارکسون و سایرین اثبات کرده‌اند که یک ϵـ هسته برای مسئله‌ی ۱ ـ مرکز است. در واقع آن‌ها نشان‌داده‌اند یک $O(\frac{1}{\epsilon})$ـ مرکز به دست آمده برای مجموعه‌ای از نقاط، یک ϵـ هسته برای مسئله‌ی ۱ ـ مرکز برای همان مجموعه نقاط است.

الگوریتم آگاروال به این گونه عمل می‌کند که اولین نقطه‌ی جویبار داده را به عنوان تنها نقطه‌ی مجموعه‌ی K_1 در نظر می‌گیرد. حال تا وقتی که نقاطی که فرا می‌رسند داخل $(1+\epsilon)Meb(K_1)$ قرار بگیرند، ادامه می‌دهد. اولین نقطه‌ای که در شرایط ذکر شده صدق نمی‌کند را p_2 بنامید. حال الگوریتم کلارکسون را بر روی $K_1 \cup \{p_2\}$ اجرا کرده و مجموع هسته‌ی به دست آمده را K_2 بنامید. با ادامه‌ی این روند، الگوریتم، دنباله‌ای از مجموعه هسته $\kappa = \{K_1, \cdots, K_u\}$ نگه می‌دارد و زمانی که نقطه‌ی p_{u+1} پیدا شود که در هیچ کدام از $(1+\epsilon)Meb(K_j)$ به ازای $1 \leqslant j \leqslant u$ نباشد، الگوریتم کلارکسون را برای $\cup_{j=1}^{u} K_j \cup \{p_{u+1}\}$ اجرا نموده و مجموعه هسته‌ی به دست آمده را K_{u+1} می‌نامد. با توجه به نحوه‌ی ساخته شدن K_iها، به راحتی می‌توان نشان داد رابطه‌ی زیر برقرار است:

$$P \subset \cup_{i=1}^{u} (1+\epsilon)Meb(K_i)$$

حال در نهایت برای به دست آوردن جواب نهایی کافی است کوچک‌ترین کره‌ای که

$$\cup_{i=1}^{u} (1+\epsilon)Meb(K_i)$$

را می‌پوشاند را به عنوان جواب محاسبه کنیم. چن و سایرین ثابت کرده‌اند که کره‌ی نهایی دارای شعاع حداکثر ۱/۲۲ برابر شعاع بهینه است. برای مشاهده جزئیات بیش‌تر، به [۲۳، ۷] مراجع کنید.

آگاروال نه تنها الگوریتمی ارائه داد که در نهایت، ثابت شد حداکثر جوابی با ضریب تقریب ۱/۲۲ برابر جواب بهینه می‌دهد، بلکه نشان داد، که با حافظه‌ی چندجمله‌ای بر اساس $\log n$ و d نمی‌توان الگوریتمی ارائه داد که ضریب تقریب بهتر از $\frac{1+\sqrt{2}}{2}$ داشته باشد.

قضیه‌ی ۳ـ ۱ هر الگوریتم تحت مدل جویبار داده که یک αـ تقریب برای مسئله‌ی ۱ ـ مرکز برای مجموعه‌ی S شامل n نقطه در فضای $\mathbb{R}^d$ نگه می‌دارد، برای $\alpha \leqslant \frac{1+\sqrt{2}}{2}(1-\frac{2}{d^{\frac{1}{3}}})$ با احتمال حداقل $\frac{2}{3}$ نیاز به $\Omega(\min\{n, e^{d^{\frac{1}{3}}}\})$ حافظه مصرف می‌کند.

اثبات. ایده‌ی اصلی اثبات بر اساس قضیه‌ی معروف آلیس و باب[۱۲] در نظریه انتقال اطلاعات بنا شده

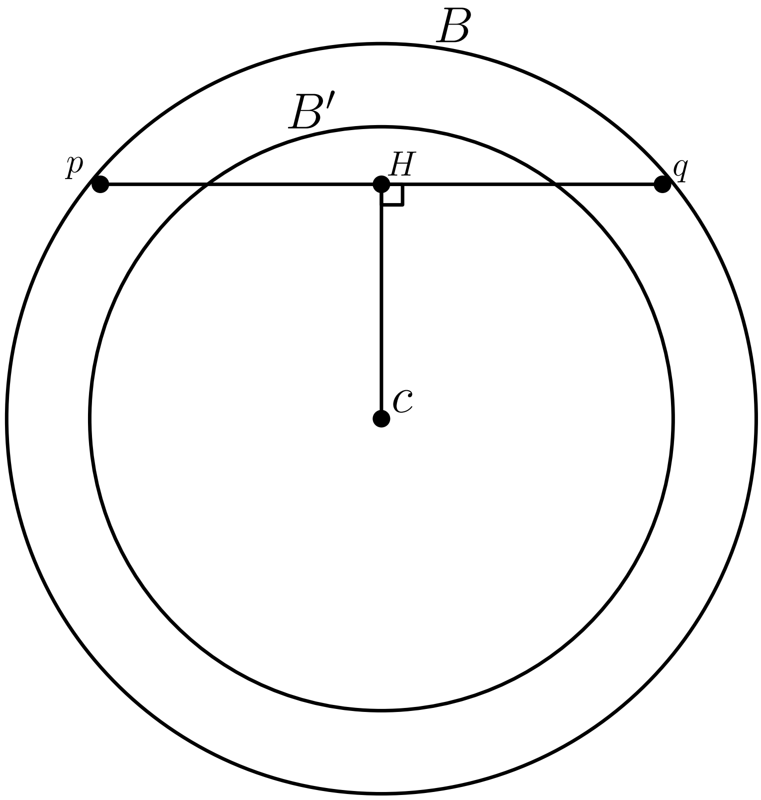
[۱۲] alice and bob

است. برای خواندن اثبات این قضیه می‌توانید به مرجع [۷] مراجعه کنید.

□

علاوه بر مسئله‌ی ۱ ــ مرکز، مسئله‌ی دو مرکز نیز در سال‌های اخیر مورد توجه قرار گرفته است و بهبودهایی نیز برای این مسئله ارائه شده است. آهن و سایرین [۲۵] در سال ۲۰۱۴، اولین الگوریم با ضریب تقریب کم‌تر از دو را برای مسئله‌ی ۲ ــ مرکز در فضای اقلیدسی ارائه دادند. این الگوریتم تقریبا پایه‌ی کار این پایان‌نامه برای حالت‌های مختلف است. به همین منظور، تعدادی از لم‌های داخل این الگوریتم که در آینده استفاده می‌شود این‌جا توضیح داده می‌شود.

لم۳ــ۲ *فرض کنید B یک کره‌ی واحد با مرکز c در فضای اقلیدسی $\mathbb{R}^d$ باشد. هر پاره‌خط pq به طول حداقل ۱/۲ که به طور کامل داخل B قرار دارد، کره‌ی $B'(c, ۰/۸)$ را قطع می‌کند.*



شکل ۳ــ۶: اثبات لم ۳ــ۲

اثبات. *صفحه‌ی گذرنده از پاره‌خط و مرکز کره را نظر بگیرید. ادامه اثبات تنها به همین صفحه محدود می‌شود، بنابراین نیاز به در نظر گرفتن ابعاد بزرگ‌تر از ۲ نیست. همان‌طور که در شکل ۳ــ۶ مشخص*

شده است، پای عمود از مرکز کره بر پاره‌خط pq را h بنامید. بدون کم شدن از کلیت مسئله فرض می‌کنیم $\|hq\| \leqslant \|hp\|$. بنابراین داریم:

$$۰/۶ = \frac{۱/۲}{۲} \leqslant \|hq\|$$

از طرفی چون پاره‌خط pq به طور کامل داخل کره‌ی واحد قرار گرفته است، بنابراین تمام نقاط آن، شامل دو سر آن، از مرکز کره، فاصله‌ی حداکثر ۱ دارند. بنابراین طبق رابطه‌ی فیثاغورث، داریم:

$$\|hc\| = \sqrt{\|qc\|^۲ - \|qh\|} \leqslant \sqrt{۱ - ۰/۶^۲} = ۰/۸$$

بنابراین نقطه‌ی h داخل کره‌ی B' قرار می‌گیرد. از طرفی چون، $\|pq\| \leqslant ۱$ بنابراین، h داخل پاره‌خط قرار دارد و در نتیجه کره‌ی B' با پاره‌خط pq تقاطع دارد.

□

لم بالا در واقع نشان می‌دهد اگر در طول الگوریتم بتوانیم دو نقطه‌ی دور نسبت به هم (حداقل ۱/۲ برابر شعاع بهینه) از یکی از دو کره‌ی بهینه را بیابیم، این پاره‌خط از مرکز کره‌ی بهینه فاصله‌ی کمی (حداکثر ۰/۸ شعاع بهینه) دارد.

لم۳ـ۳ *فرض کنید B کره‌ای به مرکز c و شعاع واحد در $\mathbb{R}^d$ باشد. پاره‌خط دلخواه pq با طول حداقل ۱/۲ که به طور کامل داخل B قرار دارد را در نظر بگیرید. هر نقطه‌ی x از پاره‌خط pq که از دو سر آن حداقل ۰/۶ فاصله داشته باشد، داخل کره‌ی $B'(c, ۰/۸)$ قرار می‌گیرد.*

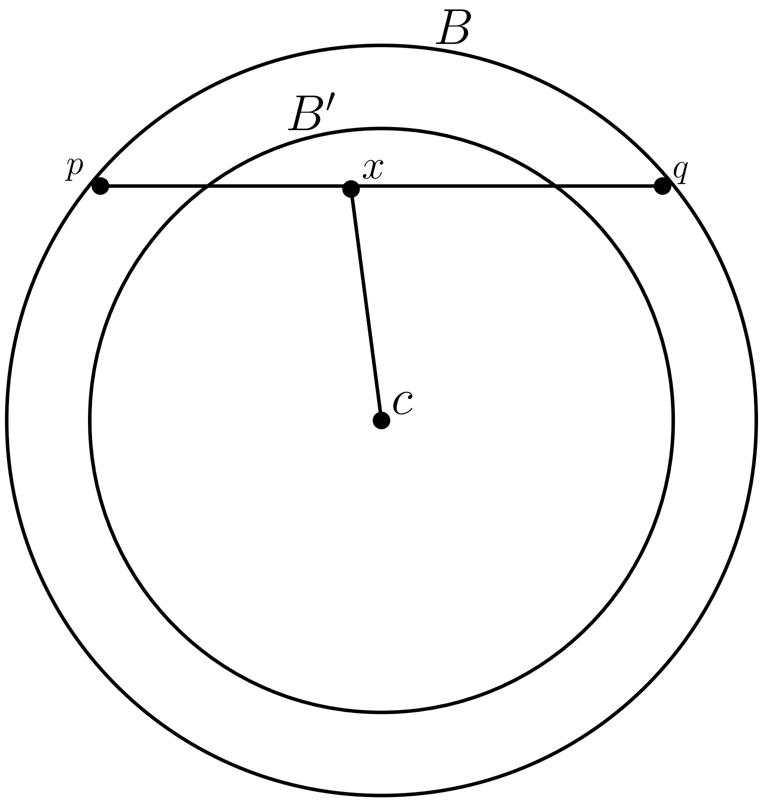
اثبات. *اثبات این لم نیز کاملا مشابه لم ۳ـ۲ است. بدون کم شدن از کلیت مسئله همان‌طور که در شکل ۳ـ۷ مشخص شده است، فرض کنید زاویه‌ی $\angle pxc$ بزرگ‌تر مساوی ۹۰ درجه است. در نتیجه داریم:*

$$\sqrt{\|px\|^۲ + \|xc\|^۲} \leqslant \|pc\| \leqslant ۱$$

از طرفی طبق فرض مسئله داریم:

$$۰/۶ \leqslant \|px\| \implies \|xc\| \leqslant \sqrt{۱ - ۰/۶^۲} = ۰/۸$$

□



شکل ۳ـ۷: اثبات لم ۳ـ۳

الگوریتم آهن، با استفاده از دو لم بالا و تقسیم مسئله به دو حالتی که دو کره‌ی بهینه بیش از دو برابر شعاع بهینه یا کم‌تر از برابر شعاع بهینه فاصله داشته باشند، دو الگوریتم کاملا جداگانه ارائه می‌دهد که به طور موازی اجرا می‌گردند. اجرای موازی این دو الگوریتم، در هر لحظه دو جواب درست ارائه می‌دهد و کافی است برای جواب نهایی بین دو شعاعی که به عنوان جواب خود می‌دهند، شعاع کم‌تر را به عنوان جواب نهایی الگوریتم بدهیم.

۳ـ۳ kـ مرکز با داده‌های پرت

فصل ۴

نتایج جدید

در این فصل نتایج جدید به‌دست‌آمده در پایان‌نامه توضیح داده می‌شود. در صورت نیاز می‌توان نتایج جدید را در قالب چند فصل ارائه نمود. همچنین در صورت وجود پیاده‌سازی، بهتر است نتایج پیاده‌سازی را در فصل مستقلی پس از این فصل قرار داد.

فصل ۵

نتیجه‌گیری

در این فصل، ضمن جمع‌بندی نتایج جدید ارائه‌شده در پایان‌نامه، مسائل باز باقی‌مانده و همچنین پیش‌نهادهایی برای ادامه‌ی کار ارائه می‌شوند.

پیوست آ

مطالب تکمیلی

پیوست‌های خود را در صورت وجود می‌توانید در این قسمت قرار دهید.

کتاب‌نامه

[1] *J. Han and M. Kamber.* Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, 2006.*

[2] *C. C. Aggarwal.* Data streams: models and algorithms*, volume 31. Springer Science & Business Media, 2007.*

[3] *M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. In* Proceedings of the 12th ACM-SIAM Symposium on Discrete Algorithms*, pages 642–651, 2001.*

[4] *M. Sipser.* Introduction to the Theory of Computation*. Cengage Learning, 2012.*

[5] *I. Dinur and S. Safra. On the hardness of approximating minimum vertex cover.* Annals of mathematics*, pages 439–485, 2005.*

[6] *V. V. Vazirani.* Approximation algorithms*. Springer Science & Business Media, 2013.*

[7] *P. K. Agarwal and R. Sharathkumar. Streaming algorithms for extent problems in high dimensions. In* Proceedings of the 21st ACM-SIAM Symposium on Discrete Algorithms*, pages 1481–1489, 2010.*

[8] *R. G. Michael and S. J. David. Computers and intractability: a guide to the theory of np-completeness.* WH Freeman & Co., San Francisco*, 1979.*

[9] *N. Megiddo and K. J. Supowit. On the complexity of some common geometric location problems.* SIAM Journal on Computing*, 13(1):182–196, 1984.*

[10] *M. Bern and D. Eppstein. Approximation algorithms for geometric problems.* Approximation algorithms for NP-hard problems*, pages 296–345, 1996.*

[11] *J. Han, M. Kamber, and J. Pei.* Data mining: concepts and techniques: concepts and techniques. *Elsevier, 2011.*

[12] *P. K. Agarwal and C. M. Procopiuc. Exact and approximation algorithms for clustering.* Algorithmica*, 33(2):201–226, 2002.*

[13] *N. Megiddo. On the complexity of some geometric problems in unbounded dimension.* Journal of Symbolic Computation*, 10(3):327–334, 1990.*

[14] *B. Chazelle and J. Matoušek. On linear-time deterministic algorithms for optimization problems in fixed dimension.* Journal of Algorithms*, 21(3):579–597, 1996.*

[15] *T. M. Chan. More planar two-center algorithms.* Computational Geometry: Theory and Applications*, 13(3):189–198, 1999.*

[16] *P. K. Agarwal, R. B. Avraham, and M. Sharir. The 2-center problem in three dimensions.* Computational Geometry*, 46(6):734–746, 2013.*

[17] *H. Zarrabi-Zadeh. Core-preserving algorithms. In* Proceedings of the 20th Canadian Conference on Computational Geometry*, pages 159–162, 2008.*

[18] *M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental clustering and dynamic information retrieval. In* Proceedings of the 29th ACM Symposium on Theory of Computing*, pages 626–635. ACM, 1997.*

[19] *R. M. McCutchen and S. Khuller. Streaming algorithms for k-center clustering with outliers and with anonymity. In* Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 165–178. 2008.*

[20] *S. Guha. Tight results for clustering and summarizing data streams. In* Proceedings of the 12th International Conference on Database Theory*, pages 268–275, 2009.*

[21] *H.-K. Ahn, H.-S. Kim, S.-S. Kim, and W. Son. Computing k centers over streaming data for small k.* International Journal of Computational Geometry and Applications*, 24(02):107–123, 2014.*

[22] *H. Zarrabi-Zadeh and T. M. Chan. A simple streaming algorithm for minimum enclosing balls. In* Proceedings of the 18th Canadian Conference on Computational Geometry*, pages 139–142, 2006.*

[23] *T. M. Chan and V. Pathak. Streaming and dynamic algorithms for minimum enclosing balls in high dimensions.* Computational Geometry: Theory and Applications*, 47(2):240–247, 2014.*

[24] *M. Bâdoiu and K. L. Clarkson. Smaller core-sets for balls. In* Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms*, pages 801–802. Society for Industrial and Applied Mathematics, 2003.*

[25] *S.-S. Kim and H.-K. Ahn. An improved data stream algorithm for clustering. In* Proceedings of the 11th*, pages 273–284. 2014.*

واژه‌نامه

الف

ابتکاری .... heuristic

ارزش .... worth

Abstract

This part should be completed.

Keywords: *Clustering, K-Center, Streaming Data, Approximation Algorithm*

Sharif University of Technology

Department of Computer Engineering

M.Sc. Thesis

Approximation Algorithms for Clustering Points in the Data Stream Model

By:

Behnam Hatami-Varzaneh

Supervisor:

Dr. Hamid Zarrabi-zade

September 2015