دانشگاه اصفهان

دانشکده ریاضی و آمار

گروه ریاضی کاربردی و علوم کامپیوتر

پروژه کارشناسی رشته علوم کامپیوتر

روش‌های خوشه‌بندی اصلاح شده

استاد راهنما:

دکتر محسن علمبردار میبدی

دکتر نجمه حسینی منجزی

دانشجو:

زهرا مشکوتی

بهمن ماه ۱۴۰۳

چکیده

در این تحقیق هدف بهبود عملکرد الگوریتم k-means و معرفی نسخه‌های بهبود یافته این الگوریتم برای حل مسائل خوشه‌بندی با دقت و سرعت بالاتر می‌باشد. به این منظور ابتدا به تعریف کلی الگوریتم k-means معمولی و سپس به تعریف چهار الگوریتم (۱) الگوریتم k-means اصلاح شده، (۲) الگوریتم k-means سراسری، (۳) الگوریتم k-means سریع و (۴) الگوریتم k-means سریع بهبود یافته می‌پردازیم. در انتها تمام الگوریتم‌ها پیاده‌سازی شده و گزارشی از نتایج حاصل ارائه می‌گردد.

واژگان کلیدی: خوشه‌بندی، الگوریتم k-means ، الگوریتم بهبودیافته k-means .

فهرست مطالب

فصل ۱

مقدمه

یکی از قضایای اساسی در مبحث انالیز داده، مبحث خوشه‌بندی است. خوشه‌بندی یا به عبارتی تقسیم داده‌ها و قرار دادن داده‌های مشابه در یک خوشه، برای استخراج الگوها و ساختارهای پنهان موجود در یک مجموعه داده ضروری است. الگوریتم‌های مختلفی برای مسئله‌ی خوشه‌بندی مطرح شده‌اند مانند

- خوشه‌بندی سلسله مراتبی
- DBScan
- k-means
- ....

در بین الگوریتم‌های خوشه‌بندی الگوریتم k-means به سبب اینکه در حین سادگی بسیار مؤثر واقع می‌شود مورد توجه است.

این الگوریتم اولین بار توسط Steinhaus Hugo در سال ۱۹۵۷ معرفی شد و سپس توسط James MacQueen در سال ۱۹۶۷ نام گذاری و به شهرت رسید و تا به این زمان به یکی از روش‌های اساسی

در مبحث یادگیری بدون نظارت تبدیل شده است.

ایده‌ی اصلی این الگوریتم تقسیم مجموعه داده به K خوشه بدون داده‌های مشترک است به طوری که مجموع مربعات فاصله‌ی هر نقطه تا مرکز خوشه‌ای که در آن قرار دارد مینیمم باشد. به این معیار مربع فاصله درون خوشه‌ای (WCSS) گفته می‌شود. در نتیجه خوشه‌بندی نقاط متناظر با حل یک مسئله بهینه‌سازی می‌باشد. تابع هدف این مسئله به صورت زیر می‌باشد.برای حل این مسئله باید مراکز بهینه همراه با تخصیص هر نقطه به شبیه ترین خوشه به آن را بیابیم.
فرمول‌بندی ریاضی این مسئله به صورت زیر بیان می‌شود:

$$\text{minimize} \quad \sum_{i=۱}^{k} \sum_{x \in C_i} \|x - \mu_i\|^۲$$

در این تابع k تعداد خوشه ها، C_i مجموعه نقاط در خوشه ی i ام ، μ_i مرکز خوشه ی i ام و x یک نقطه در مجوعه داده است.

در ادامه در الگوریتم‌های ذکر شده همگی دارای تابع هدف یکسان و برابر با تابع هدف ذکر شده می‌باشند. همچنین معیار شباهت استفاده شده فاصله اقلیدسی می‌باشد و هر نقطه در هر تکرار به خوشه‌ی متناظر با مرکزی که کمترین فاصله از آن را دارد تخصیص داده می‌شود.
امروزه انواع مختلفی از این الگوریتم معرفی شده است مانند Global ،fast k-means Global k-means و modified-kmeans که انواع بهبود یافته‌ی الگوریتم k-means می‌باشند.

به وضوح قابل مشاهده است که k-means و انواع آن در حوزه‌های مختلف مانند بازاریابی و دسته‌بندی مشتری‌ها با رفتار مشابه، حوزه‌های مربوط به تصویر مانند تشخیص چهره و بخش بندی تصویر

و حوزه‌های پزشکی و ... به طور گسترده در حال استفاده هستند.

بنابراین به دلیل کاربر گسترده این الگوریتم خوشه‌بندی، بهبود آن هم از نظر دقت و هم سرعت ضروری و لازم به نظر می‌رسد.

فصل ۲

تعاریف مقدماتی

۱.۲ الگوریتم k-means

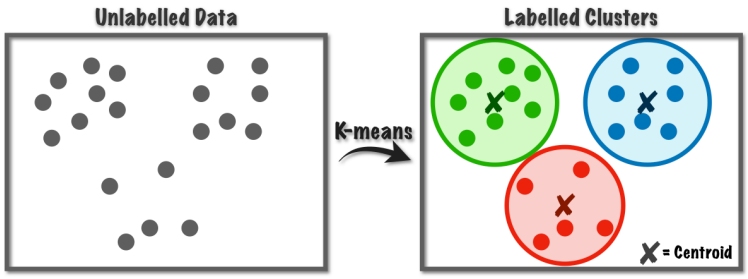
روش‌ها و الگوریتم‌های متعددی برای تبدیل اشیاء به گروه‌های همشکل یا مشابه وجود دارد. الگوریتم k-means یکی از ساده‌ترین و محبوب‌ترین الگوریتم‌هایی است که در داده‌کاوی [۱] بخصوص در حوزه یادگیری نظارت نشده [۲]به کار می‌رود. معمولا در حالت چند متغیره، باید از ویژگی‌های مختلف اشیا به منظور طبقه‌بندی و خوشه کردن آن‌ها استفاده کرد. به این ترتیب با داده‌های چند بعدی سروکار داریم که معمولا به هر بعد از آن، ویژگی یا خصوصیت گفته می‌شود. با توجه به این موضوع، استفاده از توابع فاصله مختلف در این جا مطرح می‌شود.

الگوریتم خوشه‌بندی k-means از گروه روش‌های خوشه‌بندی تفکیکی [۳] محسوب می‌شود و درجه پیچیدگی محاسباتی آن برابر با $O(n^{dk+1})$ است، به شرطی که n تعداد اشیاء، d بعد ویژگی‌ها و k تعداد خوشه‌ها باشد. همچنین پیچیدگی زمانی برای این الگوریتم برابر با $O(ndki)$ است، که البته منظور از i تعداد

[۱] data mining
[۲] unsupervised learning
[۳] Partitioning Clustering

تکرارهای الگوریتم برای رسیدن به جواب بهینه است.

در خوشه‌بندی k-means از بهینه‌سازی یک تابع هدف استفاده می‌شود. پاسخ‌های حاصل از خوشه‌بندی در این روش، ممکن است به کمک کمینه‌سازی یا بیشینه‌سازی تابع هدف صورت گیرد. به این معنی که اگر ملاک «میزان فاصله» بین اشیاء باشد، تابع هدف براساس کمینه‌سازی خواهد بود پاسخ عملیات خوشه‌بندی، پیدا کردن خوشه‌هایی است که فاصله بین اشیاء هر خوشه کمینه باشد. در مقابل، اگر از تابع مشابهت برای اندازه‌گیری مشابهت اشیاء استفاده شود، تابع هدف را طوری انتخاب می‌کنند که پاسخ خوشه‌بندی مقدار آن را در هر خوشه بیشینه کند.



معمولا زمانی که هدف کمینه‌سازی باشد، تابع هدف را تابع هزینه [۴] نیز می‌نامند. این کمینه‌سازی به صورت مسئله بهینه‌سازی زیر تعریف می‌شود.

$$\text{minimize} \quad f_k(x) \quad \text{subject to} \quad x = (x^1, \ldots, x^k) \in \mathbb{R}^{n \times k},$$

where

$$f_k(x^1, \ldots, x^k) = \frac{1}{m} \sum_{i=1}^{m} \min_{j=1,\ldots,k} \|x^j - a^i\|^2.$$

[۴] cost function

در این تابع $x = (x^1, \ldots, x^k) \in \mathbb{R}^{n \times k}$ بیانگر مراکز خوشه‌ها و m تعداد نقاط می باشد.

۱.۱.۲ مراحل پیاده‌سازی الگوریتم

گام مقداردهی اولیه:

در الگوریتم k-means معمولی در ابتدا k نقطه اولیه به عنوان مرکز خوشه‌ها به صورت تصادفی انتخاب می‌شوند. روش‌های دیگری مانند k-means++ وجود دارند که نقاط اولیه را با معیاری مشخص و غیر تصادفی انتخاب می‌کنند. k-means به شدت به انتخاب نقاط اولیه وابسته است و ممکن است در بهینه‌ی محلی باقی بماند.

گام تکرار :

هر نقطه را به نزدیک‌ترین خوشه‌اش نسبت می‌دهیم. اپدیت مراکز به این صورت که میانگین نقاط درون هر خوشه به عنوان مراکز جدید محاسبه می‌شوند و دوباره هر نقطه به نزدیک‌ترین خوشه‌اش نسبت داده می‌شود.

شرط توقف :

مرحله‌ی قبل تا جایی ادامه پیدا می‌کند که یا به ماکزیمم مقدار تکرار که از قبل معلوم شده برسیم و یا مراکز همگرا شوند. همگرا شدن مراکز به این معناست که دیگر تغییر و بهبود محسوسی در مراکز محاسبه شده دیده نشود.

فصل ۳

انواع بهبودیافته‌ی الگوریتم k-means

۱.۳ الگوریتم k-means سراسری

همانطور که گفته شد الگوریتم k-means معمولی بسیار به انتخاب نقاط اولیه مراکز وابسته است و انتخاب نقاط اولیه‌ی نامناسب می‌تواند در نهایت منجر به گرفتار شدن در بهینه‌ی محلی گردد. بنابراین انواع بهبودیافته‌ای از این الگوریتم معرفی شد که تمرکز بر روی انتخاب مراکز اولیه‌ی مناسب باشد. در این الگوریتم‌ها در هر تکرار یک نقطه به عنوان نقطه‌ی اولیه برای q امین مرکز ($k \geqslant q$) مسئله با k خوشه انتخاب می‌شود و به این نوع از انتخاب مراکز، افزایشی [۱] گفته می‌شود.

الگوریتم k-means سراسری یا به اختصار GKM، یک الگوریتم از نوع افزایشی می‌باشد. در این الگوریتم در هر تکرار تمام نقاط مجموعه داده به غیر از مراکز فعلی به عنوان نقطه‌ی اولیه احتمالی برای مرکز بعدی در نظر گرفته می‌شود. بنابراین در هر مرحله دارای چندین راه‌حل اولیه هستیم که برای هر کدام الگوریتم

[۱] incremental

k-means معمولی را به کار می‌بریم تا بتوانیم بهترین راه‌حل را میان تمام راه‌حل های اولیه بیابیم. این روند تا جایی که هر k خوشه مشخص شود ادامه میابد و همانطور که به نظر می‌رسد بار محاسباتی این الگوریتم بسیار بالاست بنابراین برای مجموعه داده‌های متوسط و بزرگ کاربردی نمی‌باشد.

۱.۱.۳ مراحل پیاده‌سازی الگوریتم

- گام اول: مقداردهی اولیه

  برای محاسبه‌ی x^1 میانگین تمام نقاط مجموعه داده A به عنوان اولین مرکز در نظر می‌گیریم.

  قرار دهید q = ۱

- گام دوم : بررسی شرط توقف

  قرار دهید $q = q + 1$ اگر q>k متوقف شود.

- گام سوم : مراکز $x^1, \ldots, x^{q-1}$ از مرحله‌ی قبل را در نظر بگیرید. سپس هر نقطه‌ی a_i داخل مجموعه داده که عضو مراکز نباشد را به عنوان مرکز اولیه برای امینq مرکز خوشه در نظر می‌گیریم. به این معنا که در این مرحله دارای چندین راه حل اولیه با q نقطه هستیم : $(x^1, \ldots, x^{k-1}, a_i)$ الگوریتم k-means را بر روی هر کدام از گزینه‌ها اجرا می‌کنیم و بهترین حالت را انتخاب می‌کنیم. بهترین گزینه حالتی است که تابع هدف با در نظر گرفتن آن در مقایسه با بقیه کمترین مقدار را دارا باشد.

  سپس مراکز $(y^1, \ldots, y^q)$ بدست آمده را ذخیره می‌کنیم.

- گام چهارم: قرار دهید $x^i = y^i,\ i = 1, \ldots, q$ و به مرحله‌ی دوم بروید.

۲.۳ الگوریتم k-means سراسری سریع

همانطور که اشاره کردیم الگوریتم GKM به دلیل استفاده‌ی متعدد از k-means در هر مرحله دارای زمان پیاده‌سازی بسیار طولانی و بار محاسباتی بالا می‌باشد. بنابراین به دنبال رفع این مشکلات الگوریتم k-means سراسری سریع [۲]معرفی گردید.

این الگوریتم نیز افزایشی می‌باشد و در هر مرحله برای یافتن نقطه‌ی اولیه برای امینq مرکز ($k \geqslant q$) با در نظر گرفتن هر نقطه‌ی a_j که عضو مراکز نباشد به جای پیاده کردن k-means بر روی هر گزینه یک مقدار با نام b_j محاسبه می‌شود و بر اساس آن پاسخ را میابیم.

- گام اول: مقداردهی اولیه

  محاسبه‌ی x^1:میانگین تمام نقاط مجموعه داده A به عنوان اولین مرکز در نظر می‌گیریم.

  قرار دهید q = ۱

- گام دوم:بررسی شرط توقف

  قرار دهید $q = q + 1$ اگر q>k متوقف شود.

- گام سوم: مراکز $x^1, \ldots, x^{q-1}$ از مرحله ی قبل را در نظر بگیرید. برای انتخاب نقطه‌ی اولیه برای امینq مرکز خوشه، برای هر نقطه‌ی a_j داخل مجموعه داده که عضو مراکز نباشد یک مجموعه شامل تمام نقاطی که به a_j نزدیکتر از مرکز خوشه‌ای که در آن قرار دارند هستند، تشکلیل می‌دهیم. که در آن d^i_{q-1} بیانگر مینیمم فاصله‌ی نقطه‌ی i از q-۱ مرکز خوشه قبلی است.

[۲] fast global k-means

$$I = \{i \in \{۱, \dots, m\} : \|a^i - a^j\|^۲ < d^i_{q-۱}\}$$

در این مرحله b_j را به صورت زیر تعریف می‌کنیم :

$$f_{q-۱}(x^۱, \dots, x^{q-۱}) - f_k(x^۱, \dots, x^{q-۱}, a^j) = \sum_{i \in I}(d^i_{q-۱} - \|a^j - a^i\|^۲) = b_j$$

مقدار b_j نشان می‌دهد با انتخاب a_j به انواع نقطه‌ی اولیه برای q امین مرکز به چه اندازه مقدار فعلی تابع هدف کاهش میابد. بنابراین نقطه‌ای که دارای ماکزیمم مقدار b باشد انتخاب ما برای نقطه‌ی اولیه برای q امین مرکز خوشه می‌باشد. مراکز بدست آمده ذخیره شده و الگوریتم kmeans روی آنها اجرا می‌شود و پاسخ پیاده‌سازی را به صورت زیر داریم.

$$(y^۱, \dots, y^q)$$

- گام چهارم: قرار دهید $x^i = y^i,\ i = ۱, \dots, q$ و به مرحله‌ی دوم بروید.

بار محاسباتی این الگوریتم بسیار نسبت به GKM پایین‌تر است اما واضح است که دقت این الگوریتم نسبت به آن پایین‌تر می‌باشد زیرا در GKM برای تمام گزینه‌ها k-means اعمال می‌شود و بر اساس مقدار تابع هدف نهایی، جواب مورد نظرمان را می‌یابیم ولی در GKM سریع تنها پس از انتخاب مرکز اولیه الگوریتم k-means اعمال شده و به مرحله‌ی بعدی می‌رود. مقدار b_j مقداری است که در یک تکرار انتظار داریم با انتخاب a_j به عنوان نقطه‌ی اولیه برای مرکز از تابع هدف کاسته شود اما ممکن است در نهایت این مقدار کاسته شده از تابع هدف نباشد.

پس چگونه می‌توانیم با در نظر گرفتن بار محاسباتی دقت الگوریتم را افزایش دهیم؟

.

۳.۳ الگوریتم بهبودیافته‌ی k-means سراسری سریع

[۳]

در الگوریتم بهبودیافته ی GKM سریع تمام مراحل مانند GKM سریع انجام می‌شوند با این تفاوت که به هنگام انتخاب نقطه‌ی اولیه برای امینq مرکز ($k \geqslant q$) n درصد از نقاطی که دارای b ماکزیمم بودند را انتخاب کرده و روی پاسخ‌های اولیه‌ی بدست آمده الگوریتم k-means اعمال می‌شود.

در الگوریتم GKM برای تمام نقاط به غیر از مراکز این عملیات انجام می‌شود و در GKM سریع تنها برای نقطه با b ماکزیمم k-means را اعمال می‌کردیم، درنهایت در الگوریتم gkm fast modified برای درصدی از نقاط k-means را اعمال می‌کنیم تا در نهایت پاسخی که مقدار تابع هدف بهتری نسبت به باقی گزینه‌ها داشت را انتخاب کنیم. اینکار سبب بدست آوردن مراکز بهینه‌تر و در نتیجه کاهش احتمال گرفتار شدن در بهینه‌ی محلی می‌شود.

هرچه درصد n بزرگ‌تری بگیریم تعداد بیشتری از پاسخ‌های اولیه بررسی می‌شود و دقت بالاتر رفته ولی از طرفی زمان اجرا بیشتر خواهد شد.

۱.۳.۳ مراحل پیاده‌سازی الگوریتم

- گام اول : مقداردهی اولیه

برای محاسبه‌ی x^1میانگین تمام نقاط مجموعه داده A به عنوان اولین مرکز در نظر می‌گیریم.

قرار دهید $q = ۱$

[۳] modified fast global k-means algorithm

- گام دوم : بررسی شرط توقف

قرار دهید $q = q + ۱$ اگر q>k متوقف شود.

- گام سوم : مراکز $x^۱, \ldots, x^{q-۱}$ از مرحله‌ی قبل را در نظر بگیرید. برای انتخاب نقطه‌ی اولیه برای امینq مرکز خوشه، برای هر نقطه‌ی a_j داخل مجموعه داده که عضو مراکز نباشد یک مجموعه شامل تمام نقاطی که به a_j نزدیکتر از مرکز خوشه‌ای که در آن قرار دارند هستند، تشکلیل می‌دهیم. که در آن $d^i_{q-۱}$ بیانگر مینیمم فاصله‌ی نقطه‌ی i از q-۱ مرکز خوشه قبلی است.

$$I = \{i \in \{۱, \ldots, m\} : \|a^i - a^j\|^۲ < d^i_{q-۱}\}$$

در این مرحله b_j را به صورت زیر تعریف می‌کنیم:

$$f_{q-۱}(x^۱, \ldots, x^{q-۱}) - f_k(x^۱, \ldots, x^{q-۱}, a^j) = \sum_{i \in I}(d^i_{q-۱} - \|a^j - a^i\|^۲) = b_j$$

برای تمام نقاط به غیر از مراکز این مقدار محاسبه شده و به صورت نزولی مرتب می‌کنیم. سپس $n < m$ نقطه اول در این لیست که دارای بیشترین مقادیر b هستند را دخیره می‌کنیم. بسته به کاربرد می‌توان مقدار n را درصدی مجموعه داده گرفت (m تعداد کل نقاط مجموعه داده است.).

مراکز بدست آمده ذخیره شده به عنوان نقاط اولیه برای q امین مرکز درنظر گرفته می‌شوند بنابراین دارای n راه حل اولیه با q نقطه هستیم.

$$(x^۱, \ldots, x^{k-۱}, a_i)$$

الگوریتم kmeans را روی هر راه‌حل اعمال کرده و گزینه‌ای که تابع هدف نسبت به بقیه حالت‌ها

کمتر می‌شود انتخاب می‌شود.پاسخ پیاده‌سازی به صورت زیر در نظر می‌گیریم.

$$(y^۱, \ldots, y^q)$$

• گام چهارم: قرار دهید $x^i = y^i,\ i = ۱, \ldots, q$ و به مرحله‌ی دوم بروید.

انتظار می‌رود دقت این الگوریتم از GKM کمتر ولی از GKM سریع بیشتر .همچنین از GKM سریع، کندتر ولی از GKM سریع‌تر باشد که نتایج پیاده‌سازی همین موضوع را تصدیق می‌کنند.

۴.۳ الگوریتم k-means بهبود یافته

[۴]

الگوریتم k-means بهبود یافته نیز یک الگوریتم بهبود یافته‌ی k-means معمولی از نوع افزایشی می‌باشد اما بر خلاف الگوریتم‌های معرفی شده در این الگوریتم مقدار k معلوم نیست و روند خوشه‌بندی هنگامی که شرط توقف برقرار شود، متوقف می‌شود.

همچنین در این الگوریتم برای یافتن نقطه‌ی اولیه برای k امین مرکز بعدی یک الگوریتم جداگانه به شرح زیر تعریف می‌شود.

۱.۴.۳ الگوریتم یافتن نقطه‌ی اولیه k-th مرکز

فرض کنید مراکز $(x^۱, \ldots, x^{k-۱})$ برای مسئله‌ی با k-۱ خوشه معلوم باشد. برای یافتن k امین مرکز در این الگوریتم برای هر نقطه a_j که عضو مراکز نباشد مجموعه ی I که شامل تمام نقاطی است که به این نقطه نسبت به مرکز خوشه فعلیشان نزدیک‌تر هستند را تشکیل می‌دهیم.

[۴] modified k-menas algorithm

$$I = \{i \in \{۱, \ldots, m\} : \|a^i - a^j\|^۲ < d^i_{q-۱}\}$$

سپس مرکز c_jکه مرکز این مجموعه است را محاسبه می‌کنیم. در این مرحله یک تابع کمکی [۵] به صورت زیر تعریف می‌کنیم :

$$\bar{f}_k(Y) = \frac{۱}{m} \sum_{i=۱}^{m} \min\{d^i_{k-۱}, \|y - a^i\|^۲\}$$

در این تابع $d^i_{k-۱}$ بیانگر مینیمم فاصله‌ی نقطه‌ی a_i از k-۱ مرکز خوشه قبلی است.

متناظر با هر c_j مقدار $\bar{f}_k(c_j)$ را محاسبه می‌کنیم. هر پارامتر c_j که مقدار تابع را مینیمم سازد به عنوان نقطه‌ی اولیه برای مرکز بعدی انتخاب می‌شود.

۲.۴.۳ مراحل پیاده‌سازی الگوریتم

پارامتر ε یک استانه [۶] در شرط توقف می‌باشد. که در این پیاده‌سازی برابر با ۰٫۰۰۱ است. انتخاب ε بزرگ باعث به وجود آمدن خوشه‌های بزرگتر و ε کوچکتر باعث به وجود آمدن خوشه‌های مجازی [۷] می‌شود .

- گام اول:

  انتخاب استانه مناسب. $\varepsilon > ۰$

  محاسبه‌ی $x^۱$: میانگین تمام نقاط مجموعه داده A به عنوان اولین مرکز در نظر می‌گیریم.

  مقدار $f^۱$که مقدار متناظر با تابع هدف می‌باشد را محاسبه می‌کنیم.

[۵] auxiliary cluster function
[۶] tolerance
[۷] artificial clusters

قرار می‌دهیم $k = ۱$

- گام دوم : محاسبه‌ی مرکز خوشه بعدی

قرار دهید : k+۱ = k

فرض کنید $x^۱, \ldots, x^(k-۱)$ مراکز خوشه‌ها برای مسئله‌ی با k-۱ خوشه باشد .

با استفاده از الگوریتمی که در ادامه به آن می‌پردازیم، نقطه شروع $\bar{y}$ رابرای k امین مرکز می‌یابیم.

- گام سوم: اصلاح مراکز

نقاط $(x^۱, \ldots, x^{k-۱}, \bar{y})$ به عنوان مراکز اولیه در الگویتم kmeans در نظر می‌گیریم و مسئله‌ی k-partition را حل می‌کنیم.

فرض کنید $(y^۱, \ldots, y^k)$ مراکز به دست امده‌ی مسئله‌ی k-partition باشند. مقدار تابع هدف f^k را محاسبه می‌کنیم.

- گام چهارم : شرط توقف، اگر

$$\frac{f^{k-۱} - f^k}{f^۱} < \epsilon$$

متوقف شود و در غیر ای صورت قرار دهید:

$$x^i = y^i,\ i = ۱, \ldots, k$$

و به گام ۲ برو.

۵.۳ نتایج پیاده‌سازی و مشاهدات

در این جداول مقادیر f ، α و t برای پیاده‌سازی الگوریتم‌های معرفی شده بر روی مجموعه داده‌های مختلف نمایش داده شده‌اند.

به ترتیب مقدار f بیانگر مقدار تابع هدف، α بیانگر تعداد نرم‌های محاسبه شده و t بیانگر زمان پیاده‌سازی می‌باشد.

توجه کنید در پیاده‌سازی الگوریتم k-means بهبود یافته مقدار k توسط کاربر معلوم نمی‌شود بنابراین ترتیب نتایج بدست آمده از الگوی باقی پیاده‌سازی‌ها پیروی نمی‌کند.

در سه جدول اول نتایج نمایش داده شده بر روی دو مجموعه داده به نام‌های breast و concrete می‌باشد که به ترتیب دارای ۶۸۳ و ۱۰۳۰ تعداد داده و هر دو دارای ۹ بعد می‌باشند.

در سه جدول دوم نتایج نمایش داده شده بر روی دو مجموعه داده به نام‌های TSPLIB۱۰۶۰ و TSPLIB۳۰۳۸ می باشد که به ترتیب دارای ۱۰۶۰ و ۳۰۳۶ تعداد داده و هر دو دارای ۲ بعد می‌باشند.

تصاویری که به عنوان نمونه مشاهده می‌کنید نتیجه‌ی خوشه‌بندی مجموعه داده ی TSPLIB۱۰۶۰ با الگوریتم‌های تعریف شده می‌باشد .

$$k = ۵$$

Table :۳٬۱ Breast and Concrete data sets

		k-means	gkm	fast gkm	modified fast gkm	modif
No.	f_opt	f	f	f	f	f
۲	۰	۱۹۳۲۳٫۱۷۴	۱۹۳۲۳٫۱۷۴	$۱٫۹۳E+۰۴$	۱۹۳۲۳٫۲۰۵	$۱٫۹۳E$
۵	۰	۱۳۷۷۰٫۵۵۷	۱۳۷۰۶٫۳۸۶	$۱٫۳۸E+۰۴$	۱۳۷۰۶٫۷۹۶	$۱٫۴۲E$
۱۰	۰	۱۱۷۱۴٫۵۰۲	۱۰۲۰۲٫۲۶۰	$۱٫۲۴E+۰۴$	۱۰۲۵۲٫۲۴۸	$۱٫۰۶E$
۱۵	۰	۱۰۳۸۱٫۶۱۴	۸۶۹۹٫۷۹۹	$۹٫۹۳E+۰۳$	۹۰۱۳٫۶۱۳	$۹٫۰۰E$
۲۰	۰	۸۱۸۶٫۲۹۱	۷۶۵۶٫۵۴۰	$۸٫۴۷E+۰۳$	۸۰۰۴٫۹۲۱	$۸٫۱۹E$
۲۳	–	–	–	$-$	–	$۸٫۰۰E$
۲	۰	۳۱۴۷۳۱۵۷٫۳۷۲	۳۱۴۷۳۱۵۷٫۳۷۲	$۳٫۲۲E+۰۷$	۳۱۴۷۳۱۵۷٫۳۷۲	$۳٫۱۵E$
۵	۰	۲۱۱۰۷۴۴۳٫۴۴۹	۱۹۵۳۵۷۳۱٫۹۸۷	$۲٫۰۱E+۰۷$	۱۹۵۳۵۷۳۱٫۹۸۶	$۲٫۰۰E$
۱۰	۰	۱۳۱۳۲۱۹۱٫۳۶۶	۱۲۲۸۳۹۷۸٫۹۱۵	$۱٫۴۳E+۰۷$	۱۲۳۰۳۷۳۷٫۵۳۷	$۱٫۲۴E$
۱۵	۰	۹۹۴۸۴۴۴٫۷۱۵	۹۲۲۱۸۵۸٫۱۴۹	$۹٫۶۴E+۰۶$	۹۲۸۴۸۰۴٫۰۰۲	$۹٫۵۲E$
۲۰	۰	۸۳۶۷۰۷۶٫۵۴۳	۷۶۲۰۲۸۵٫۹۲۴	$۸٫۶۱E+۰۶$	۷۶۸۳۶۸۸٫۶۹۸	$۸٫۰۸E$
۲۶	–	–	–	$-$	–	$۷٫۰۶E$

Table :۳٬۲ Breast and Concrete data sets

		k-means	gkm	fast gkm	modified fast gkm	modified k-me
No.	f_opt	α	α	α	α	α
۲	۰	۶۸۳۰	۵۰۱۸۰۰۱	۲۰۷۵۷۳	۱۱۳۶۴۵۳	۱۴۰۸۳۴۶
۵	۰	۱۱۹۵۲۵	۵۶۴۰۰۷۷۴	۳۵۱۶۹۰۷	۱۸۳۵۴۰۰۱	۱۱۲۸۹۳۰۷
۱۰	۰	۶۸۳۰۰	۲۱۲۴۰۴۸۰۴	۱۸۰۵۶۰۲۷	۶۳۷۷۱۴۳۰	۴۶۶۱۷۴۸۲
۱۵	۰	۱۷۴۱۶۵	۴۰۲۱۱۶۹۳۳	۴۴۳۳۸۰۲۸	۱۲۵۲۹۲۲۹۹	۱۰۵۳۸۸۲۶۶
۲۰	۰	۲۱۸۵۶۰	۶۳۶۲۰۶۳۰۴	۸۲۶۸۷۳۱۹	۲۰۴۴۵۹۶۲۷	۱۸۷۴۰۱۵۴۰
۲۳	–	–	–	–	–	۲۴۸۰۴۱۰۱۲
۲	۰	۱۴۴۲۰	۲۳۱۸۶۳۳۰	۲۰۸۷۰۷	۵۱۷۳۳۰۷	۳۲۰۴۳۳۰
۵	۰	۸۲۴۰۰	۱۹۶۰۹۱۴۰۰	۵۷۵۱۲۴۰	۴۱۶۲۳۷۶۱	۲۵۵۶۴۶۰۰
۱۰	۰	۳۷۰۸۰۰	۶۶۳۶۰۸۴۰۰	۳۴۳۰۶۴۰۰	۱۷۶۵۰۹۷۰۷	۱۰۵۸۱۳۹۶۰
۱۵	۰	۲۰۰۸۵۰	۱۱۹۹۱۳۷۳۳۰	۹۱۲۴۰۸۸۸	۳۵۰۵۱۲۲۵۳	۲۳۹۲۵۶۶۴۰
۲۰	۰	۳۲۹۶۰۰	۱۷۹۲۵۵۶۳۸۰	$۱٫۷۶E+۰۸$	۵۷۱۸۵۴۵۷۱	۴۲۵۶۲۳۸۱۰
۲۶	–	–	–	–	–	۷۱۹۶۷۰۲۷۰

Table :۳٬۳ Breast and Concrete data sets

		k-means	gkm	fast gkm	modified fast gkm	modified k-means
No.	f_opt	t	t	t	t	t
۲	۰	۰٫۰۴	۰٫۴۳	۵٫۴۹	۶٫۲۷	۸٫۱۷
۵	۰	۰٫۰۴	۲٫۱۰	۳۸٫۷۲	۴۰٫۵۵	۱۵٫۰۱
۱۰	۰	۰٫۰۴	۵٫۳۰	۱۲۸٫۸۹	۱۷۷٫۵۴	۲۲٫۲۱
۱۵	۰	۰٫۰۴	۷٫۰۱	۲۵۴٫۸۲	۲۴۸٫۷۷	۳۳٫۹۲
۲۰	۰	۰٫۰۴	۱۰٫۵۷	۴۱۷٫۰۳	۴۱۰٫۴۶	۴۰٫۷۴
۲۳	–	–	–	—	–	۵۶۶٫۷۳
۲	۰	۰٫۰۳	۰٫۹۶	۹٫۰۰	۷٫۷۲	۱۹٫۴۴
۵	۰	۰٫۰۴	۴٫۷۶	۷۷٫۰۱	۷۰٫۲۳	۳۷٫۳۳
۱۰	۰	۰٫۰۴	۹٫۹۲	۲۷۶٫۰۰	۲۵۱٫۸۶	۵۱٫۹۱
۱۵	۰	۰٫۰۴	۱۵٫۰۴	۵۵۵٫۱۴	۴۷۸٫۶۴	۶۷٫۴۳
۲۰	۰	۰٫۰۴	۲۲٫۳۷	۹۳۶٫۱۰	۹۱۰٫۹۸	۸۱٫۸۹
۲۶	–	–	–	—	–	۱۵۷۹٫۵۸

Table :۳٬۴ TSPLIB۱۰۶۰ and TSPLIB۳۰۳۸ data sets

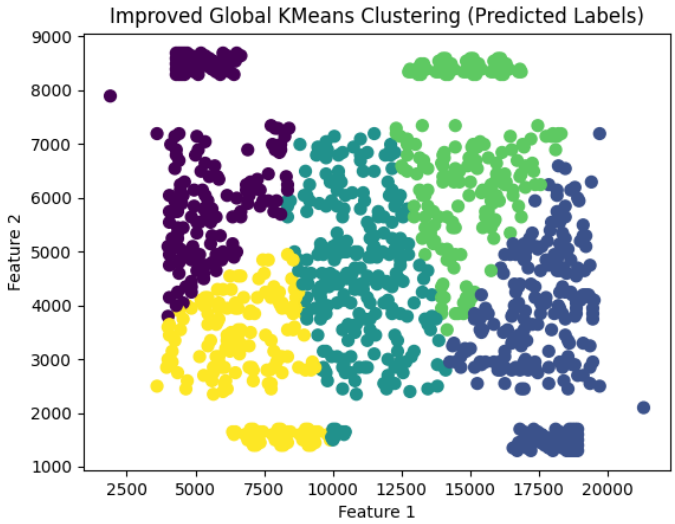
		k-means	gkm	fast gkm	modified fast gkm
No.	f_opt	f	f	f	f
۲	۰	۹۸۳۱۹۴۹۸۸۴٫۲۰۱	$۹٫۸۳۱۹۴۹۸۸e+۰۹$	$۹٫۸۳e+۰۹$	$۹٫۸۳۱۹۴۹۸۸e+۰۹$
۵	۰	۳۸۰۰۹۶۸۳۲۹٫۷۱	$۳٫۷۹۱۵۲۹۸۳e+۰۹$	$۳٫۸۰e+۰۹$	$۳٫۸۰۰۳۳۷۷۲e+۰۹$
۱۰	۰	۱۷۶۱۷۸۳۰۳۴٫۱۲۴	$۱٫۷۵۴۸۵۸۰۶e+۰۹$	$۱٫۷۸e+۰۹$	$۱٫۷۵۴۸۵۸۰۶e+۰۹$
۱۵	۰	۱۲۵۵۱۹۰۲۴۳٫۲۷۸	$۱٫۱۲۱۶۲۰۱۲e+۰۹$	$۱٫۱۴e+۰۹$	$۱٫۱۲۵۸۳۱۶۷e+۰۹$
۲۰	۰	۸۸۰۹۰۶۱۴۱٫۹۶۶	$۷٫۹۲۰۹۶۹۸۵e+۰۸$	$۸٫۸۵e+۰۸$	$۸٫۰۳۱۱۴۵۶۱e+۰۸$
۲	۰	۳۱۶۸۸۲۶۲۹۸٫۴۲۳	$۳٫۱۶۸۸۰۴۶۸e+۰۹$	$۳٫۱۷E+۰۹$	$۳٫۱۶۸۸۱۳۲۱e+۰۹$
۵	۰	۱۲۰۱۱۲۳۱۲۶٫۳۶۱	$۱٫۱۹۸۲۰۲۶۹e+۰۹$	$۱٫۲۰E+۰۹$	$۱٫۱۹۸۲۰۸۶۰e+۰۹$
۱۰	۰	۵۷۴۲۶۸۷۰۳٫۳۳۷	$۵٫۶۰۴۱۱۰۷۸e+۰۸$	$۵٫۸۳E+۰۸$	$۵٫۶۰۴۱۱۰۷۸e+۰۸$
۱۱	–	–	–	$-$	–
۱۵	۰	۳۵۶۱۰۴۵۴۸٫۰۵۷	$۳٫۵۶۱۳۲۹۸۸e+۰۸$	$۳٫۶۴E+۰۸$	$۳٫۵۶۳۰۲۵۰۶e+۰۸$
۲۰	۰	۲۷۹۸۰۳۲۹۳٫۹۰۱	$۲٫۶۷۰۲۱۸۳۵e+۰۸$	$۲٫۸۳E+۰۸$	$۲٫۶۷۳۷۸۲۲۷e+۰۸$

Table :۳.۵ TSPLIB۱۰۶۰ and TSPLIB۳۰۳۸ data sets

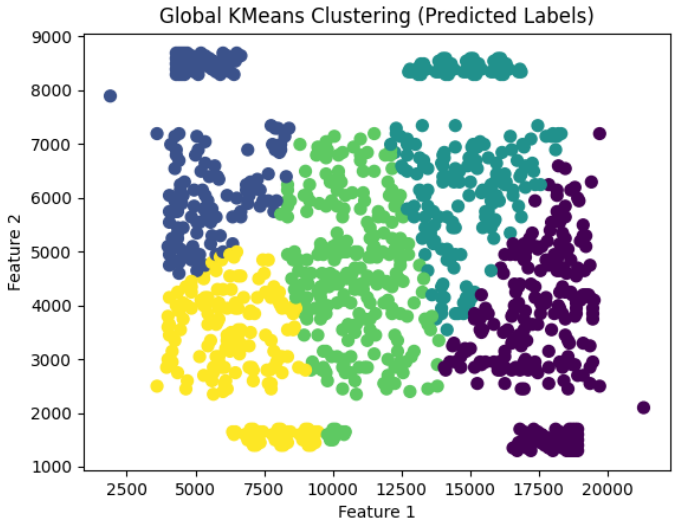
		k-means	gkm	fast gkm	modified fast gkm	modified k-
No.	f_opt	α	α	α	α	α
۲	۰	۱۴۸۴۰	۱۳۱۴۷۱۸۰	۳۹۲۵۳۶	۲۶۸۶۳۷۶	۳۳۸۶۷۰۰
۵	۰	۵۳۰۰۰	۲۵۱۶۲۴۹۲۰	۸۵۲۹۸۹۱	۴۷۸۸۸۳۵۸	۲۷۱۴۹۷۸۰
۱۰	۰	۲۳۳۲۰۰	۹۲۶۹۳۵۰۲۰	۴۳۸۹۱۸۴۳	۲۲۳۷۴۰۷۳۴	۱۱۱۸۱۷۲۸۰
۱۵	۰	۴۲۹۳۰۰	۱۶۹۴۸۶۸۹۸۰	۱۰۷۲۷۰۱۵۹	۴۵۹۳۲۴۶۹۸	–
۲۰	۰	۴۸۷۶۰۰	۲۶۱۷۵۸۰۹۶۰	۱۹۷۶۸۱۶۶۲	۷۸۳۴۸۳۹۹۰	–
۲	۰	۷۲۹۱۲	۱۶۰۹۵۰۲۰۲	۲۷۶۸۵۴۴	۲۹۱۳۸۳۸۴	۲۷۷۳۹۹۷۸
۵	۰	۳۳۴۱۸۰	۱۷۷۳۴۳۵۵۳۸	۶۲۷۳۸۱۲۳	۳۵۳۷۰۸۰۰۹	۲۲۲۰۸۶۹۱۴
۱۰	۰	۱۳۰۶۳۴۰	۷۸۸۳۴۸۵۴۴۲	۳۴۶۱۸۱۵۱۶	۱۸۴۷۰۳۸۲۱۵	۹۱۶۱۴۸۳۹۴
۱۱	–	–	–	–	–	۱۱۱۰۳۰۰۸۹
۱۵	۰	۱۴۱۲۶۷۰	۱۸۳۱۹۱۵۸۲۲۸	۸۶۴۳۳۷۰۶۱	۴۲۷۶۲۹۴۷۵۸	–
۲۰	۰	۱۵۱۹۰۰۰	۳۱۱۸۸۱۷۱۷۹۸	۱۶۱۵۱۶۶۲۸۷	۷۹۶۵۸۴۰۸۹۰	–

Table :۳٬۶ TSPLIB۱۰۶۰ and TSPLIB۳۰۳۸ data sets

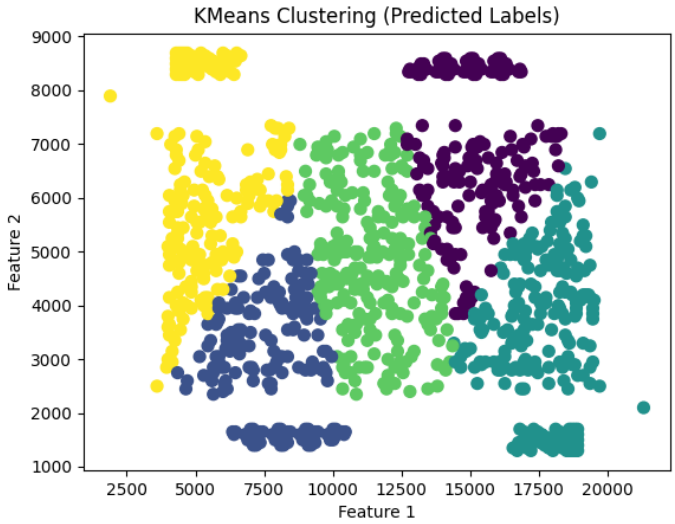
		k-means	gkm	fast gkm	modified fast gkm	modified k-means
No.	f_opt	t	t	t	t	t
۲	۰	۰٫۰۳	۰٫۷۱	۱۲٫۷۶	۱۰٫۶۳	۱۵٫۳۷
۵	۰	۰٫۰۴	۴٫۳۹	۹۶٫۳۹	۷۷٫۲۳	۲۶٫۰۰
۱۰	۰	۰٫۰۴	۹٫۹۹	۳۱۲٫۴۹	۲۳۹٫۱۸	۲۵۸٫۴۵
۱۵	۰	۰٫۰۴	۱۵٫۲۰	۵۹۸٫۱۶	۴۸۴٫۴۱	–
۲۰	۰	۰٫۰۴	۲۰٫۷۰	۱۰۰۱٫۷۱	۷۴۲٫۵۳	–
۲	۰	۰٫۰۴	۳٫۱۳	۹۶٫۳۲	۹۳٫۵۴	۱۲۲٫۹۷
۵	۰	۰٫۰۴	۱۶٫۲۴	۷۶۶٫۶۴	۷۸۵٫۳۴	۲۳۳٫۹۹
۱۰	۰	۰٫۰۴	۴۳٫۵۶	۲۵۹۳٫۲۴	۲۵۸۲٫۴۹	۳۳۶٫۴۰
۱۱	–	–	–	—	–	۳۲۱۱٫۹۵
۱۵	۰	۰٫۰۴	۸۲٫۷۱	۵۶۵۸٫۳۶	۴۱۰۱٫۵۲	–
۲۰	۰	۰٫۰۴	۱۱۸٫۱۴	۸۲۵۰٫۰۵	۶۵۵۸٫۲۸	–


Improved Global KMeans Clustering (Predicted Labels)
Feature 2
Feature 1
9000
8000
7000
6000
5000
4000
3000
2000
1000
2500
5000
7500
10000
12500
15000
17500
20000

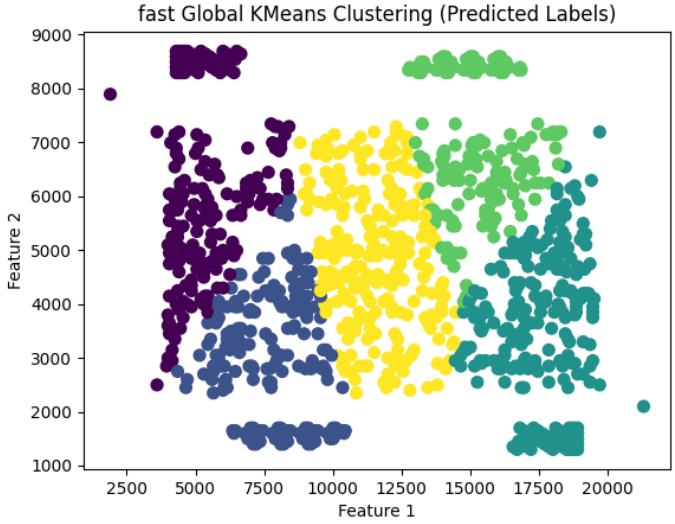
$k = ۵$


Global KMeans Clustering (Predicted Labels)
Feature 2
Feature 1
9000
8000
7000
6000
5000
4000
3000
2000
1000
2500
5000
7500
10000
12500
15000
17500
20000

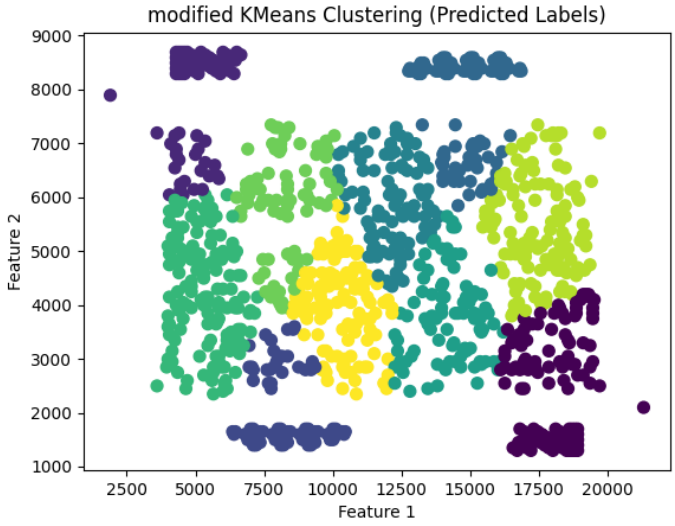
$k = ۵$


KMeans Clustering (Predicted Labels)
Feature 2
Feature 1
9000
8000
7000
6000
5000
4000
3000
2000
1000
2500
5000
7500
10000
12500
15000
17500
20000

$k = ۵$


fast Global KMeans Clustering (Predicted Labels)
Feature 2
Feature 1
9000
8000
7000
6000
5000
4000
3000
2000
1000
2500
5000
7500
10000
12500
15000
17500
20000

$$k = ۱۰$$


modified KMeans Clustering (Predicted Labels)
Feature 2
Feature 1
9000
8000
7000
6000
5000
4000
3000
2000
1000
2500
5000
7500
10000
12500
15000
17500
20000

مراجع

[1] Bagirov, A., Hoseini-Monjezi, N., Taheri, S. (2023). A novel optimization approach towards improving separability of clusters. Computer & Operations Research, 152, 106135.

[2] Ye, J., Zhao, Z., Liu, H. (2007, June). Adaptive distance metric learning for clustering. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-7). IEEE.

[3] Xing, E., Jordan, M., Russell, S. J., Ng, A. (2002). Distance metric learning with application to clustering with side-information. Advances in neural information processing systems, 15.

[4] Bagirov, A. M. (2008). Modified global k-means algorithm for minimum sum-of-squares clustering problems. Pattern Recognition, 41(10), 3192-3199.

Abstract.

In this project ....

Key words: Clustring, k-means algorithm, modified k-means algorithm

University of Isfahan

Facility of Mathematics and Statistics

Departemant of Applied Mathematic and Computer Science

Bachelor Project

Modified Clustring Methods

Supervisor:

Dr. Mohsen Alambardar Meybodi

Dr. Najmeh Hoseini Monjezi

By:

Zahra Meshkati

Feburay 2025