خوشه بندی با ثبات با استفاده از منظم سازی داده های پرت-پراکندگی

چکیده :

علی رغم محبوبیت الگوریتم های خوشه بندی سنتی نظیر کامینز و خوشه بندی احتمالی، نتایج خوشه بندی آن ها به حضور داده های پرت در داده ها حساس است. حتی تعداد کمی از داده های پرت قادرند تا توانایی این الگوریتم ها را در شناسایی ساختار های پنهان معنی دار کاهش دهند به طوری که موجب می شوند تا نتایج و برایند آن ها نیز غیر قابل اطمینان شود. این مقاله الگوریتم های خوشه بندی قوی(با ثبات/) را توسعه می دهد، الگوریتم هایی که نه تنها هدف آن ها خوشه بندی داده ها، بلکه شناسایی داده های پرت است. رویکرد های جدید به حضور نادر داده های پرت متکی هستند که به پراکندگی در یک دامنه انتخاب شده به طور منطقی ترجمه می شود. با استفاده از پراکندگی در دامنه داده پرت، رویکرد های خوشه بندی احتمالی و کامینز قوی داده پرت-آگاه پیشنهاد می شوند. تازگی آن ها منوط به شناسایی داده های پرت ضمن تاثیر گذاری بر پراکندگی در دامنه داده های پرت از طریق منظم سازی انتخاب شده به طور منطقی می باشد. یک رویکرد نزول مختصات بلوکی برای دست یابی به الگوریتم های تکراری با تضمین های همگرایی و اندکی پیچیدگی محاسباتی مازاد با توجه به همتاهای غیر قوی آن ها توسعه می یابد. نسخه مرکزی الگوریتم های خوشه بندی قوی نیز برای مدیریت کارامد داده های با بعد زیاد، شناسایی خوشه های قابل تفکیک غیر خطی یا حتی اشیای خوشه که با بردار ها نشان داده نمی شوند توسعه می یابد. تست های عددی در هر دو مجموعه داده های مصنوعی و واقعی برای ارزیابی عملکرد و قابلیت تعمیم و اجرای الگوریتم های جدید استفاده می شوند.

لغات و اصطلاحات کلیدی : نزول مختصات (بلوک)، خوشه بندی، الگوریتم بیشینه سازی امید ریاضی، گروه-لاسو، کامینز، روش های هسته ای، مدل های ترکیبی، استواری، پراکندگی

1- مقدمه

هدف خوشه بندی، تقسیم یک مجموعه داده ها به زیر مجموعه هایی به نام خوشه ها می باشد به طوری که داده های قرار گرفته در یک خوشه از برخی جهات با هم مشابه باشند. کار با داده های بدون برچسب و تحت مفروضات حداقل موجب شده است تا خوشه بندی به یک ابزار چالش بر انگیز و جهانی برای آشکار سازی ساختار های داده

در طیف وسیعی از برنامه ها نظیر تحلیل ریزآرایه DNA و بیو انفورماتیک، تحلیل شبکه های اجتماعی، پردازش تصویر و داده کاوی(18،35) تبدیل شود. به علاوه، خوشه بندی می تواند یک مرحله پیش پردازش برای یادگیری نظارت شده در شرایطی باشد که در آن ها برچسب زدن داده ها به طور جداگانه پر هزینه است. تفسیر های مختلف از خوشه در رشته های مختلف منجر به فراوانی الگوریتم های خاص برنامه شده است(35).

در میان الگوریتم هایی که داده های برداری را خوشه بندی می کنند، خوشه بندی مبتنی بر مدل ترکیبی گوسی(GMM) و کامینز، دو طرح رایج می باشند(26-35). کامینز بستگی به فواصل اقلیدسی به عنوان یک شاخص تشابه دارد که به موجب آن ایجاد بخش ها یا پارتیشن هایی می کند که پراکندگی درون خوشه ای را به حداقل می رساند(18). بر عکس، کامینز نرم(هم چنین موسوم به فازی) نیز برای هم پوشانی خوشه ها از طریق نسبت دادن هر نقطه مبنا به خوشه های چندگانه(2) مناسب است. خوشه بندی مبتنی بر GMM داده های استخراج شده از تابع چگالی احتمالی(pdf) را در نظر می گیرد که در آن هر pdf کلاس مشروط متناظر با یک خوشه(35) متناظر است. سپس خوشه بندی به عنوان یک محصول جانبی چارچوب براورد حداکثر درست نمایی(ML) برای پارامتر های GMM مطرح می شود که معمولا از طریق الگوریتم بیشینه سازی- امید ریاضی(EM) (11) بدست می آید. روش های هسته ای نیز برای خوشه بندی خوشه های قابل تفکیک غیر خطی طراحی شده اند(29-30).

روش های خوشه بندی کامینز و مبتنی بر GMM، علی رغم محبوبیت خود به داده های متناقض موسوم به داده های پرت ناشی از وابستگی کارکردی آن ها به فاصله اقلیدسی(20) حساس می باشند. داده های پرت به ندرت در داده ها به دلیل خطا های خواندن ظاهر می شوند یا دلیل دیگر این است که آن ها متعلق به پدیده های نادر یا بسیار مهم می باشند. با این حال، حتی تعداد کمی از داده های پرت میتوانند موجب شوند تا نتایج خوشه بندی غیر قابل اطمینان شود: براورد های مربوط به مراکز خوشه و پارامتر مدل می توانند به شدت اریبی داشته باشند و از این روی تخصیص داده به خوشه با مشکل مواجه می شود. به این ترتیب باید استوار سازی رویکرد های خوشه بندی در برابر داده های پرت با پیچیدگی محاسباتی کم به منظور آشکار سازی ساختار اصلی در داده ها استفاده شود.

چندین رویکرد خوشه بندی قوی(با ثبات) مورد مطالعه قرار گرفته است(16). آن دسته از رویکرد های مرتبط با چارچوب توسعه یافته در این جا شامل خوشه بندی احتمالی است که بر اساس کامینز فازی با اندازه گیری خصوصیت هر نقطه مبنا با توجه به هر خوشه برای تصمیم گیری در مورد این که آیا یک نقطه مبنا، داده پرت است یا نه(24-28) ایجاد می شود. با این حال، خوشه بندی احتمالی به مقدار دهی حساس است و می تواند یک خروجی را بیش از یک بار تولید کند. مشابه با(19)، روش خوشه بندی نویز ارایه شده توسط(10) یک خوشه اضافی را معرفی می کند که همه داده های پرت را پوشش می دهد و مرکز آن به طور اکتشافی فرض می شود که فاصله مساوی از همه داده های غیرپرت دارد. به منظور کاهش اریبی مرکزی، روش آلفا کات مراحل کامینز را اجرا می کند ولی مراکز خوشه با استفاده از تنها درصد آلفا از داده های تخصیص داده شده به هر خوشه براورد می شوند(36).

سایر جایگزین های قوی شامل رویکرد های خوشه بندی متوالی می باشند که یک تک خوشه را در یک زمان شناسایی کرده و نقاط آن را از مجموعه داده ها حذف می کنند(22-38). یک بیضی حداقل حجم حاوی یک کسر از پیش تعیین شده از داده ها در هر مرحله در(22) شناسایی می شود، در حالی که(38) مدل آلوده-ٔ هابر را با GMM ترکیب می کند. با این حال، حذف متوالی نقاط موجب تاخیر در ساختار داده اصلی می شود.

روش های خوشه بندی مبتنی بر فاصله l_1 (K-مدین) با الهام از آماره های باثبات، تابع دو وزنی توکی و میانگین پیراسته نیز پیشنهاد شده اند(4، 15، 23)، با این حال آن ها همگی محدود به خوشه های تفکیک پذیر خطی می باشند. یک رویکرد خوشه بندی برای شناسایی خوشه های با شکل دلخواه با استفاده از توابع هسته ای در (1) توسعه یافت.اگرچه این روش در برابر داده های پرت مقاوم و انعطاف پذیر است، با این حال هدف این روش براورد تراکم است در حالی که تعداد خوشه های شناسایی شده به شدت بستگی به جست و جوی شبکه نسبت به یک پارامتر هسته ای دارند. رویکرد های مبتنی بر GMM با ثبات، Pdf های داده پرت-آگاه را معرفی کرده و مسئله ML معمولا از طریق الگوریتم های شبه-EM حل می شوند(27-31).

اولین هدف این مطالعه، معرفی یک مدل داده برای خوشه بندی می باشد که صریحا داده های پرت را از طریق یک بردار داده پرت قطعی به ازای هر نقطه مبنا پوشش دهد(بخش دوم). یک نقطه مبنا در صورتی به عنوان داده پرت در نظر گرفته می شود که بردار داده پرت متناظر آن، غیر صفر باشد. بررسی این موضوع که داده های پرت

پراکندگی کم تری در دامنه بردار داده های پرت دارد، منجر به ایجاد یک ارتباط مبرهن بین خوشه بندی و پارادایم سنجش فشرده(CS)(7) می شود. بر اساس این مدل، یک روش خوشه بندی داده پرت-آگاه برای خوشه بندی هر دو از دیدگاه های قطعی(کامینز) و احتمالی(GMM) توسعه می یابد.

دومین هدف این مطالعه بررسی الگوریتم های خوشه بندی تکراری مختلف توسعه یافته برای خوشه بندی کامینز سخت، کامینز نرم و خوشه بندی مبتنی بر GMM(بخش سوم) می باشد. الگوریتم ها بر مبنای تکرار نزول مختصات بلوک(BCD) بوده و تولید آپدیت های شکل بسته برای هر مجموعه از متغیر های بهینه سازی می کند. به ویژه، براورد داده های پرت برای حل مسئله گروه- لاسو استفاده می شود(37) که راه حل آن در شکل بسته محاسبه می شود. الگوریتم های خوشه بندی با ثبات جدید با پیچیدگی محاسباتی کم هزینه با رتبه مشابه با الگوریتم های خوشه بندی غیر با ثبات عمل می کنند.

چندین کاربرد معاصر در زمینه بیو انفورماتیک، تحلیل شبکه های اجتماعی، پردازش تصویر و یادگیری ماشینی مستلزم خوشه بندی داده پرت-آگاه داده های با بعد بالا یا مستلزم خوشه های تفکیک پذیر غیر خطی می باشند. برای رفع این نیاز های خوشه بندی، الگوریتم های خوشه بندی با ثبات جدید در بخش چهارم بررسی شده و این سومین هدف این مطالعه است. مدل مفروض نه تنها امکان این هسته سازی را برای هر دو خوشه بندی کامینز و و احتمالی را می دهد، بلکه منجر به الگوریتم های تکراری با آپدیت های شکل بسته می شود. در بخش 5، الگوریتم های توسعه یافته با استفاده از مجموعه داده های مصنوعی و نیز واقعی از سیستم های تشخیص رقم دست نوشته و شبکه های اجتماعی تست می شوند. نتایج موید کارایی و اثر بخشی روش ها است. در بخش شش نتیجه گیری ارایه می شود.

یادداشت:حروف برجسته پایین نویس(بالا نویس) بیانگر بردار های ستونی(ماتریس ها) و حروف کالیگرافی نشان دهنده مجموعه ها می باشند: $(\cdot)^T$ نشاندهنده ترانسپوز، $\mathbb{N}_N$ مجموعه مقادیر طبیعی $\{1,\ldots,N\}$، $\mathbf{0}_p$ $(\mathbf{1}_p)$ بردار $p\times 1$ از همه مقادیر صفر، $\mathbf{I}_p$ ماتریس همانی $p\times p$ $\mathrm{diag}(x_1,\ldots,x_p)$ یک ماتریس متعامد $p\times p$ با درایه های قطری $x_1,\ldots,x_p$، $\mathrm{range}(\mathbf{X})$ فضای دامنه ماتریس $\mathbf{X}$، $\mathbb{E}[\cdot]$ عملگر امید ریاضی، $\mathcal{N}(\mathbf{x};\mathbf{m},\mathbf{\Sigma})$ pdf گوسی چند متغیره با میانگین m و ماتریس کواریانس $\mathbf{\Sigma}$ ارزیابی شده در

$\|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\mathbf{x}^T\mathbf{A}\mathbf{x}}$ برای یک ماتریس نیمه متناهی مثبت $\mathbf{A}$؛ $\|\mathbf{x}\|_p := (\sum_{i-1}^{n} |x_i|^p)^{\frac{1}{p}}$ با $p \geq 1$ برای قاعده ℓ_p در $\mathbb{R}^n$ است.

2- خوشه بندی پراکندگی-آگاه: شرایط و معیار ها

بعد از مرور فرایند خوشه بندی، یک مدل مربوط به داده های آلوده به داده پرت معرفی می شود. بر اساس این مدل، رویکرد های با ثبات و قوی برای خوشه بندی کامینز(بخش 2 الف) و احتمالی(بخش 2 ب) توسعه می یابند.

الف: **خوشه بندی کامینز**

با توجه به یک مجموعه از بردار های p- بعدی $\mathcal{X} := \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$، فرض کنید که $\{\mathcal{X}_1, \ldots, \mathcal{X}_C\}$ تقسیم $\mathcal{X}$ به زیر C زیر مجموعه(خوشه ها) $\mathcal{X}_c \subset \mathcal{X}$ for $c \in \mathbb{N}_C$ باشد که مجتمع، دو به دوناسازگار و غیر تهی می باشند. هدف خوشه بندی تقسیمی، تقسیم $\mathcal{X}$ می باشد به طوری که دو بردار تخصیص داده شده به یک خوشه از برخی جهات مشخص نظیر فاصله اقلیدسی نسبت به بردار های تخصیص داده شده به خوشه های دیگر می باشند.

در میان روش های های خوشه بندی تقسیمی، کامینز یکی از رایج ترین روش ها با قابلیت های شناخته شده و تاریخچه طولانی است(5). در شرایط کامینز، یک $\mathbf{m}_c \in \mathbb{R}^p$ مرکزی در هر خوشه $\mathcal{X}_c$ معرفی می شود. سپس به جای مقایسه فواصل بین جفت نقاط در $\mathcal{X}$، فواصل مرکزی نقطه $\|\mathbf{x}_n - \mathbf{m}_c\|_2$ در نظر گرفته می شوند. به علاوه، برای هر بردار ورودی $\mathbf{x}_n$، کامینز عضویت های مجهول u_{nc} برای $c \in \mathbb{N}_C$ را معرفی می کند به طور یکه وقتی $\mathbf{x}_n \in \mathcal{X}_c$ است برابر با 1 و در غیر این صورت برابر 0 است. برای تضمین یک تقسیم بندی معتبر، ضرایب عضویت نبایستی دو دویی باشند$(\mathbf{c1})\text{:}\ u_{nc} \in \{0,1\}$، آن ها هم چنین باید قادر رفع محدودیت ها(c2) باشند: $\sum_{n=1}^{N} u_{nc} > 0$ به ازای همه C برای حذف خوشه های خالی و (c3) $\sum_{c=1}^{C} u_{nc} = 1$ به ازای همه n می باشد به طوری که هر بردار به خوشه نسبت داده می شود.

فرایند خوشه بندی کامینز همانند یافتن مراکز $\{\mathbf{m}_c\}_{c=1}^{C}$ و تخصیص خوشه u_{nc} با حل مسئله بهینه سازی می باشد

$$\min_{\{\mathbf{m}_c\},\{u_{nc}\}} \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}\,\|\mathbf{x}_n - \mathbf{m}_c\|_2^2 \text{ subject to (c1)–(c3).} \quad (1)$$

با این حال، مسئله(1) حتی برای C=2 گفته می شود که ان پی هارد است. عملا، یک راه حل نیمه بهینه با استفاده از الگوریتم کامینز معروف دنبال می شود. این الگوریتم محدودیت(c2) را ندارد به طوری که در عوض در مرحله پس از پردازش چک می شود. سپس، به طور متناوب هزینه را در(1) با توجه به یک مجموعه از متغیر های $\{\mathbf{m}_c\}$ یا $\{u_{nc}\}$ کمینه سازی کرده ضمن این که متغیر دیگر را ثابت در نظر می گیرد و مرحله را تکرار می کند. تکرار های کامینز به نقطه ساکن (1)(32) تبدیل می شوند.

به منظور دست یابی به اطلاعات بیشتر در خصوص خوشه بندی کامینز، یک مدل داده مرتبط $\mathbf{x}_n = \sum_{c=1}^{C} u_{nc}\mathbf{m}_c + \mathbf{v}_n$ مفروض است که در آن $\mathbf{v}_n$ یک بردار میانگین صفر می باشد که انحراف $\mathbf{x}_n$ را از نقطه مرکزی $\mathbf{m}_c$ پوشش می دهد. می توان به آسانی دید که تحت C1-C3، کمینه ساز های(1)، یک برازش داده حداقل مربعات را از داده های $\{\mathbf{x}_n\}_{n=1}^{N}$ در خصوص محدودیت های تخصیص خوشه در اختیار می گذارند. با این حال این مدل ساده ولی هنوز پر کاربرد، داده های پرت را در نظر نمی گیرد یعنی نقاط $\mathbf{x}_n$ که مدل مفروض را نقض می کنند. این واقعیت به همراه حساسیت هزینه LS به باقیمانده های بزرگ، آسیب پذیری کامینز را به داده های پرت(10) توجیه می کنند.

به منظور استوار سازی و با ثبات سازی کامینز، مدل داده های زیر را در نظر بگیرید که صریحا داده های پرت را پوشش می دهد:

$$\mathbf{x}_n = \sum_{c=1}^{C} u_{nc}\mathbf{m}_c + \mathbf{o}_n + \mathbf{v}_n,\ n \in \mathbb{N}_N \qquad (2)$$

که در آن بردار داده پرت $\mathbf{o}_n$ در صورتی که $\mathbf{x}_n$ متناظر با داده پرت باشد به صورت غیر صفر قطعی تعریف شده و در غیر این صورت به صورت $\mathbf{0}_p$ تعریف می شود. مجهولات $\{u_{nc}, \mathbf{m}_c, \mathbf{o}_n\}$ در (2) را می توان

اکنون با استفاده از رویکرد LS به عنوان کمینه ساز های $\sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}\|\mathbf{x}_n-\mathbf{m}_c-\mathbf{o}_n\|_2^2$ براورد کرد و یا به دلیل (c1)(c3 به عنوان کمینه ساز های $\sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}\|\mathbf{x}_n-\mathbf{m}_c-\mathbf{o}_n\|_2^2$ در نظر گرفت که اگر $\mathbf{v}_n \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_p)$. باشد، به صورت براورد های حداکثر درست نمایی(ML) مطرح می شوند.

حتی گر u_{nc} معلوم باشد، براورد $\{\mathbf{m}_c\}$ و $\{\mathbf{o}_n\}$ بر اساس تنها $\{\mathbf{x}_n\}$ یک مسئله از پیش تعیین شده خواهد بود. یک یافته کلیدی این است که اکثر $\mathbf{o}_n$ ها صفر می باشند. این موجب می شود تا معیار های زیر برای شناسایی و خوشه بندی داده های پرت $s \leq N$ استفاده شوند

$$\begin{aligned}\min_{\mathbf{M},\mathbf{O},\mathbf{U}\in\mathcal{U}_1} \quad & \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}\|\mathbf{x}_n-\mathbf{m}_c-\mathbf{o}_n\|_2^2 \\ \text{s. to} \quad & \sum_{n=1}^{N}\mathrm{I}(\|\mathbf{o}_n\|_2>0)\leq s \qquad (3)\end{aligned}$$

که در آن $\mathbf{M}:=[\mathbf{m}_1\ \cdots\ \mathbf{m}_C], \mathbf{O}:=[\mathbf{o}_1\ \cdots\ \mathbf{o}_N], \mathbf{U}\in\mathbb{R}^{N\times C}$ نشان دهنده ماتریس عضویت با درایه های $[\mathbf{U}]_{n,c}:=u_{nc}$، $\mathcal{U}_1$ مجموعه همه ماتریس های U مطابق با C1-C3 و $\mathrm{I}(\cdot)$ تابع شاخص است. چون مسئله(3) به مسئله کامینز در(1) به ازای $s=0$ کاهش می یابد، مسئله سه، سختی ان پی یا ان پی هاردی را به مسئله 1 تعمیم می دهد. اکنون شکل لاگرانژی(3) را در نظر بگیرید

$$\min_{\substack{\mathbf{M},\mathbf{O},\\ \mathbf{U}\in\mathcal{U}_1}} \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}\|\mathbf{x}_n-\mathbf{m}_c-\mathbf{o}_n\|_2^2+\lambda\sum_{n=1}^{N}\mathrm{I}(\|\mathbf{o}_n\|_2>0) \qquad (4)$$

که در آن $\lambda\geq 0$ یک پارامتر کنترل کننده داده پرت است. به ازای $\lambda=0$، با قرار دادن $\mathbf{o}_n=\mathbf{x}_n-\mathbf{m}_c$ برای مقداری از C، یک هزینه بهینه صفر را در اختیار می گذارد که در آن همه Xm ها به صورت داده های پرت در نظر گرفته می شوند. برای λ با مقدار بزرگ تر، $\mathbf{o}_n$ بهینه برابر باصفر بوده، $\mathcal{X}$ به صورت عاری از داده پرت در نظر گرفته شده و مسئله(4) به کامینز در (1) کاهش می یابد.

در امتداد خطوط کامینز، تکرار های مشابه را می توان برای حل نیمه بهینه(4) دنبال کرد. با این حال این تکرار ها نمی توانند ضمین همگرایی را به دلیل عدم پیوستگی تابع شاخص در صفر ارایه کنند. با فرض روش حل عملی و ممکن(4)، ابتدا $\mathbf{U} \in \mathcal{U}_1$ را در نظر بگیرید. بهینه سازی با توجه به $\{\mathbf{M}, \mathbf{O}\}$ ، به دلیل $\sum_{n=1}^{N} \mathbb{I}(\|\mathbf{o}_n\|_2 > 0)$ غیر محدب است. بر اساس پاردایم موفق CS، که درآن قاعده شبه ℓ_0 یک بردار $\mathbf{x} \in \mathbb{R}^N$ به صورت $\|\mathbf{x}\|_0 := \sum_{n=1}^{N} \mathbb{I}(|x_n| > 0)$ تعریف شد، با قاعده ℓ_1محدب $\|\mathbf{x}\|_1$ جایگزین شد، مسئله ای که در(4) با معادله زیر جایگزین می شود

$$\min_{\mathbf{M},\mathbf{O},\mathbf{U} \in \mathcal{U}_1} \sum_{n=1}^{N} \sum_{c=1}^{C} u_{nc} \|\mathbf{x}_n - \mathbf{m}_c - \mathbf{o}_n\|_2^2 - \lambda \sum_{n=1}^{N} \|\mathbf{o}_n\|_2 . \quad (5)$$

هدف رویکرد کامینز با ثبات پیشنهادی، کمینه سازی(5) است که در $\{\mathbf{M}, \mathbf{O}\}$ محدب است ولی به صورت غیر محدب مشترک باقی مانده است. الگوریتم برای حل نیمه بهینه مسئله غیر محدب در(5) برای بخش 3-الف موکول می شود. توجه کنید که کمینه سازی در(5) مشابه با معیار گروه لاسو مورد استفاده برای بازیابی بردار پراکندگی بلوک در رگرسیون خطی است(37). این موجب ایجاد یک رابطه جالب بین خوشه بندی با ثبات و CS می شود. دو تبصره اکنون وجود دارد.

تبصره 1(نویز رنگی): در صورتی که محدب کواریانس $\mathbf{v}_n$ در (2) معلوم باشد، یعنی $\mathbf{\Sigma}$، قواعد ℓ_2 در (5) را به ترتیب می توان با قواعد وزنی $\|\mathbf{x}_n - \mathbf{m}_c - \mathbf{o}_n\|_{\mathbf{\Sigma}^{-1}}^2$ و $\|\mathbf{o}_n\|_{\mathbf{\Sigma}^{-1}}$ جایگزین کرد.

تبصره 2(پنالتی-ℓ_1 برای داده های پرت ورودی): جمله منظم سازی $\sum_{n=1}^{N} \|\mathbf{o}_n\|_2$ در(5)، امکان شناسایی کل بردارهای داده را به صورت داده های پرت می دهد. جایگزینی آن با $\sum_{n=1}^{N} \|\mathbf{o}_n\|_1$ امکان بازیابی ورودی داده های پرت را به جای بردار کل می دهد. راه حل های تکراری برای این مورد را می توان با استفاده از روش ارایه شده در بحش 3 توسعه داد: به دلیل محدودیت فضا، این مورد در این مقاله بررسی نمی شود.

محدودیت های(c1)(c2) مستلزم تخصیص عضویت سخت(هارد) می باشند به این معنی که هر بردار به یک خوشه تخصیص داده می شود. با این حال خوشه بندی نرم که به هر بردار امکان می دهد تا به طور جزیی متعلق

به خوشه های مختلف شود، می تواند به طور بهتر خوشه های دارای هم پوشانی را شناسایی کند(2). یک شیوه دست یابی به عضویت های کسری، استفاده از کامینز نرم است. تفاوت کامینز نرم از کامینز سخت در موارد زیر است: 1- کاهش محدودیت الفبای دو دویی(c1) به محدویت جعبه ای(c4): $u_{nc} \in [0,1]$ و 2- افزایش u_{nc} در(1) به q امین توان که در آن $q > 1$ یک پارامتر تعدیل کننده است(2). طرح کامینز نرم با ثبات پیشنهادی $\mathbf{x}_n$ را با نسخه بدون داده پرت $(\mathbf{x}_n - \mathbf{o}_n)$ جایگزین کرده و پراکندگی $\mathbf{o}_n$ را اهرم بندی کنند. این مراحل منجر به معیار زیر می شوند:

$$\min_{\mathbf{M},\mathbf{O},\mathbf{U}\in\mathcal{U}_2} \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}^{q}\left(\|\mathbf{x}_n - \mathbf{m}_c - \mathbf{o}_n\|_2^2 + \lambda\|\mathbf{o}_n\|_2\right) \quad (6)$$

که در آن $\mathcal{U}_2$ مجموعه ای از همه شاخص های u مطابق با محدودیت های c3-c4 می باشد. یک الگوریتم برای حل تقریبی(6) در بخش 3-الف ارایه می شود. به یاد داشته باشید که پارتیشن سخت یا تقسیم بندی سخت x را می توان از u_{nc} نرم با تخصیص $\mathbf{x}_n$ به $\hat{c} := \arg\max_c u_{nc}$ امین خوشه بدست آورد که در آن $\hat{c}$ می باشد.

ب: خوشه بندی احتمالی

یک شیوه جایگزین برای انجام خوشه بندی نرم، استفاده از رویکرد احتمالی است(35). برای این منظور، یک مدل توزیع ترکیبی برای $\mathbf{x}_n$ فرض می شود در حالی که $\{u_{nc}\}_{c=1}^{C}$ اکنون به صورت متغیر های تصادفی مشاهده نشده (پنهان) تفسیر می شود. مراکز $\{\mathbf{m}_c\}_{c=1}^{C}$ به صورت پارامتر های قطعی توزیع ترکیبی در نظرگرفته می شوند و براورد های ML متعاقبا از طریق الگوریتم EM بدست می آیند.

برای پوشش دادن داده های پرت، خوشه بندی احتمالی به مدل(2) تعمیم داده می شود. فرض کنید که $\{\mathbf{x}_n\}$ در (2) به صورت متغیرهای تصادفی، مستقل با توزیع یکسان باشند که از مدل ترکیبی گرفته شده اند که در آن $\{\mathbf{o}_n\}$ پارامتر های قطعی هستند. عضویت های $\mathbf{u}_n := [u_{n1}\cdots u_{nC}]^T$ بردار های تصادفی پنهان متناظر با ردیف های U بوده و مقادیر در دامنه $\{\mathbf{e}_1,\ldots,\mathbf{e}_C\}$ را اختیار می کند که در آن $\mathbf{e}_c$، c امین

ستون از lc می باشد. در صورتی که $\mathbf{x}_n$ ازc امین مولفه ترکیبی مشتق شود، آنگاه $\mathbf{u}_n - \mathbf{e}_c$ است. هم چنین فرض کنید که pdf های مشروط کلاسی به صورت گاوسی باشند و به صورت $p(\mathbf{x}_n|\mathbf{u}_n - \mathbf{e}_c) - \mathcal{N}(\mathbf{x}_n; \mathbf{m}_c + \mathbf{o}_n, \boldsymbol{\Sigma})$ به ازای همه مقادیر n و c مدل سازی شوند. این نشان میدهد که $p(\mathbf{x}_n) = \sum_{c} \pi_c \mathcal{N}(\mathbf{x}_n; \mathbf{m}_c + \mathbf{o}_n, \boldsymbol{\Sigma})$ با $\pi_c := \Pr(\mathbf{u}_n = \mathbf{e}_c)$ است. در صورتی که $\mathbf{x}_n$ مستقل باشند، انگاه احتمال لگاریتمی داده های ورودی به صورت زیر خواهد بود:

$$L(\mathbf{X}; \boldsymbol{\pi}, \mathbf{M}, \mathbf{O}, \boldsymbol{\Sigma}) := \sum_{n=1}^{N} \log\left(\sum_{c=1}^{C} \pi_c \mathcal{N}(\mathbf{x}_n; \mathbf{m}_c + \mathbf{o}_n, \boldsymbol{\Sigma})\right) \quad (7)$$

که $\mathbf{X} := [\mathbf{x}_1 \cdots \mathbf{x}_N]$ و $\boldsymbol{\pi} := [\pi_1 \cdots \pi_C]^T$ میباشد. کنترل تعداد داده های پرت(تعداد صفر بردار $\mathbf{o}_n$)، احتمال لگاریتمی منفی منظم را به صورت زیر کمینه سازی می کند

$$\min_{\boldsymbol{\Theta}} -L(\mathbf{X}; \boldsymbol{\Theta}) + \lambda \sum_{n=1}^{N} \|\mathbf{o}_n\|_{\boldsymbol{\Sigma}^{-1}} \quad (8)$$

که $\boldsymbol{\Theta} := \{\boldsymbol{\pi} \in \mathcal{P}, \mathbf{M}, \mathbf{O}, \boldsymbol{\Sigma} \succ \mathbf{0}\}$، مجموعه ای از همه پارامتر های مدل، $\mathcal{P}$ سیمپلکس احتمال $\mathcal{P} := \{\boldsymbol{\pi} : \boldsymbol{\pi}^T \mathbf{1} = 1, \boldsymbol{\pi} \geq \mathbf{0}\}$ و $\boldsymbol{\Sigma} \succ \mathbf{0}$ به این معنی است که $\boldsymbol{\Sigma}$ماتریس معین مثبت است. یک راه حل مبتنی بر EM(8) در بخش 3 ب مشتق شده است. بعد از براورد پارامتر های احتمالی، احتمالات پسین $\gamma_{nc} := \Pr(\mathbf{u}_n = \mathbf{e}_c|\mathbf{x}_n)$ را به آسانی می توان به دست آورده و به صورت عضویت های نرم آن را تفسیر کرد.

اگرچه داشتن یک کواریانس مشترک $\boldsymbol{\Sigma}$ $\forall c$ به نظر محدود کننده است، با این حال تضمین کننده این است که GMM یک روش شناخته می باشد و از این روی از مقادیر احتمال نامحدود کاذب اجتناب می کند(3، صفحه 433). به طور ویژه، می توان دید که حتی اگر همه $\mathbf{o}_n$ برابر با صفر قرار داده شوند، احتمال لگاریتمی یک GMM با $\boldsymbol{\Sigma}_c$ متفاوت به ازای هر ترکیب به صورت نامحدود رشد می کند برای مثال با برابر قرار دادن یکی از $\mathbf{m}_c$ با $\mathbf{x}_n$ و فرض این که $\boldsymbol{\Sigma}_c \to \mathbf{0}$ به ازای هر مقدار c ویژه باشد. این احتمال برای بیکران بودن

در(8) نیز مطرح شده است و توجیه کننده استفاده از Σ مشترک است. با این حال حتی با کواریانس مشترک، بردار های $\mathbf{o}_n$ قادر به تغییر احتمال لگاریتمی در (7) به سمت بی نهایت می باشند: برای مثال در نظر بگیرید که هرجفت $(\mathbf{m}_c, \mathbf{o}_n)$ مطابق با $\mathbf{x}_n = \mathbf{m}_c + \mathbf{o}_n$ و نیز فرض کنید که $\boldsymbol{\Sigma} \longrightarrow \mathbf{0}$ باشد. برای شناخته شدن مسئله کمینه سازی $L(\mathbf{X}; \boldsymbol{\Theta})$ و به منظور مقایسه آن با شرایط قطغی(تبصره 1)، تنظیم کننده $\|\mathbf{o}_n\|_{\Sigma^{-1}}$ نیز معرفی می شود. هم چنین توجه کنید که به ازای $\lambda \longrightarrow \infty$، $\mathbf{O}$ بهینه به صورت صفر است و (8) به براورد MLE از GMM کاهش می یابد: در حالی که به ازای $\lambda \longrightarrow 0$، هزینه در(8) از معادله زیر به صورت محدود می شود

3- الگوریتم های خوشه بندی با ثبات

الگوریتم ها برای حل مسائل ارایه شده در بخش 2 در این جا توسعه می یابند. بخش 3 الف بر کمینه سازی(6) تاکید دارد در حالی که کمینه سازی در (5) از(6) به ازای $q = 1$ بدست می آید. در بخش 3 ب، الگوریتم برای کمینه سازی(8) بر اساس رویکرد EM بدست می اید. در نهایت نسخه های اصلاح شده الگوریتم های جدید با انعطاف پذیری زیاد به داده های پرت در بخش 3 پ ارایه شده است.

الف: الگوریتم های کامینز(نرم) با ثبات

ابتدا حل معادله 6 را به ازای $q > 1$ در نظر بگیرید. اگرچه هزینه، غیر محدب مشترک است، با توجه به هر یک از M، O و U به صورت محدب در نظر گرفته می شود. به منظور توسعه یک راه حل عملی نیمه بهینه، این تحدب به ازای هر متغیر موجب ایجاد یک الگوریتم BCD می شود که موجب کمینه سازی تکراری هزینه با توجه به هر متغیر بهینه سازی ضمن ثابت نگه داشتن دو متغیر دیگر می شود. فرض کنید که $\mathbf{M}^{(t)}, \mathbf{O}^{(t)}$ و $\mathbf{U}^{(t)}$ به معنی راه حل های یافت شده در t امین تکرار است. هم چنین، $\mathbf{U}^{(0)}$ را به طور تصادفی در $\mathcal{U}_2$ مقدار دهی کرده و $\mathbf{O}^{(0)}$ را برابر با صفر قرار می دهد.

در اولین مرحله ازt امین تکرار،(6) در M برای $\mathbf{U} = \mathbf{U}^{(t-1)}$و $\mathbf{O} = \mathbf{O}^{(t-1)}$ بهینه سازی می شود. بهینه سازی در $\mathbf{m}_c$ تجزیه شده و هر $\mathbf{m}_c^{(t)}$ یک راه حل شکل بسته از مسئله LS می باشد به صورت زیر:

$$\mathbf{m}_c^{(t)} = \frac{\sum_{n=1}^{N} (u_{nc}^{(t-1)})^q \left(\mathbf{x}_n - \mathbf{o}_n^{(t-1)}\right)}{\sum_{n=1}^{N} (u_{nc}^{(t-1)})^q}. \qquad (9)$$

در دومین مرحله، هدف اصلی کمینه سازی(6) با توجه به $\mathbf{O}$ به ازای $\mathbf{M} = \mathbf{M}^{(t)}$ و $\mathbf{U} = \mathbf{U}^{(t-1)}$ است. مسئله بهینه سازی به ازای هر شاخص n تجزیه می شود به طوری که هر $\mathbf{o}_n$ را می توان به صورت کمینه ساز در نظر گرفت

$$\phi^{(t)}(\mathbf{o}_n) := \sum_{c=1}^{C} (u_{nc}^{(t-1)})^q \left(\|\mathbf{x}_n - \mathbf{m}_c^{(t)} - \mathbf{o}_n\|_2^2 + \lambda \|\mathbf{o}_n\|_2 \right). \qquad (10)$$

هزینه $\phi^{(t)}(\mathbf{o}_n)$ به صورت محدب ولی مشتق ناپذیر می باشد. با این حال، کمینه ساز آن را می توان در شکل بسته به شکلی که در فرض زیر نشان داده شده است بیان کرد.

فرض 1: مسئله بهینه سازی در (10) به طور منحصر به فردی با معادله زیر کمینه سازی می شود

$$\mathbf{o}_n^{(t)} = \mathbf{r}_n^{(t)} \left[1 - \frac{\lambda}{2\|\mathbf{r}_n^{(t)}\|_2} \right]_+ \qquad (11)$$

که $[x]_+ := \max\{x, 0\}$ و $\mathbf{r}_n^{(t)}$ به صورت زیر تعریف میشود

$$\mathbf{r}_n^{(t)} := \frac{\sum_{c=1}^{C} (u_{nc}^{(t-1)})^q \left(\mathbf{x}_n \quad \mathbf{m}_c^{(t)}\right)}{\sum_{c=1}^{C} (u_{nc}^{(t-1)})^q}. \qquad (12)$$

اثبات: به پیوست A مراجعه شود

آپدیت برای $\mathbf{o}_n^{(t)}$ در (11) دو نقطه متقاطع را نشان می دهد:1- هزینه $\phi^{(t)}(\mathbf{o}_n)$ در واقع به سمت کمینه ساز های صفر است و 2- تعداد داده های پرت با λ کنترل می شود. بعد از به روز رسانی و آپدیت $\mathbf{r}_n^{(t)}$، قواعد آن با آستانه $\frac{\lambda}{2}$ مقایسه می شود. اگر $\|\mathbf{r}_n^{(t)}\|_2$ مازاد بر $\frac{\lambda}{2}$ باشد، بردار $\mathbf{x}_n$ به صورت داده پرت در نظر گرفته می

شود و با $\mathbf{o}_n^{(t)}$ غیر صفر جبران می شود. در غیر این صورت، $\mathbf{o}_n^{(t)}$ برابر با صفر در نظر گرفته شده و $\mathbf{x}_n$ به صورت نقطه منظم خوشه بندی می شوند.

در طی مرحله نهایی t امین تکرار، (6) در $\mathbf{U} \in \mathcal{U}_2$ به ازای $\mathbf{O} = \mathbf{O}^{(t)}$ و $\mathbf{M} = \mathbf{M}^{(t)}$ کمینه سازی می شود. مشابه با کامینز نرم مرسوم، کمینه ساز در شکل بسته(2) قابل دسترس است.

$$u_{nc}^{(t)} = \left[\sum_{c'=1}^{C} \left(\frac{\|\mathbf{x}_n - \mathbf{m}_c^{(t)} - \mathbf{o}_n^{(t)}\|_2^2 + \lambda \|\mathbf{o}_n^{(t)}\|_2}{\|\mathbf{x}_n - \mathbf{m}_{c'}^{(t)} - \mathbf{o}_n^{(t)}\|_2^2 + \lambda \|\mathbf{o}_n^{(t)}\|_2} \right)^{\frac{1}{q-1}} \right]^{-1}. \tag{13}$$

در رابطه با کامینز هارد با ثبات، یک رویکرد مشابه BCD برای حل (5) منجر به آپدیت $\mathbf{M}^{(t)}$ و $\mathbf{O}^{(t)}$ از طریق(9) و (11-12) به ازای $q = 1$ می شود. به روز رسانی $\mathbf{U}^{(t)}$ به قاعده فاصله حداقل تبدیل می شود

$$u_{nc}^{(t)} = \begin{cases} 1, & c = \arg\min_{c'} \|\mathbf{x}_n - \mathbf{m}_{c'}^{(t)} - \mathbf{o}_n^{(t)}\|_2 \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

به یاد داشته باشید که(14) یک مورد حد(13) به ازای $q \longrightarrow 1^+$ می باشد.

الگوریتم کامینز با ثبات(RKM) به صورت الگوریتم 1 جدول بندی می شود. RKM زمانی به پایان می رسد که $\frac{\|\mathbf{M}^{(t)} - \mathbf{M}^{(t-1)}\|_F}{|\mathbf{M}^{(t)}|_F} \leq \epsilon_s$ باشد که در آن $\|\cdot\|_F$ نشان دهنده قاعده فوربنیوس ماتریس است و ϵ_s یک آستانه مثبت کوچک است برای مثال $\epsilon_s = 10^{-6}$ است. منابع محاسباتی مورد نیاز توسط RKM بعدا خلاصه می شود.

الگوریتم 1: کامینز با ثبات

مطلوب است: ماتریس داده ورودی X، تعدا خوشه های C، $q > 1$ و $\lambda > 0$،

1- $\mathbf{O}^{(0)}$ را برابر با صفر و $\mathbf{U}^{(0)}$ به طور تصادفی در $\mathcal{U}_2$ مقدار دهی کنید

2- به ازای $t = 1, 2, \ldots$ موارد زیر را انجام دهید

3- به روز رسانی $\mathbf{M}^{(t)}$ از طریق(9)

4- به روز رسانی $\mathbf{O}^{(t)}$ از طریق(11-12)

5- به روز رسانی $\mathbf{U}^{(t)}$ از طریق $(q > 1)$ (13) یا $(q = 1)$ (14)

6- پایان

تبصره 3(پیچیدگی محاسباتی RKM): فرض کنید که:(as1) تعداد خوشه های کوچک باشد برای مثال $C < p$ و (as2) تعداد نقاطی است که بزرگ تر از بعد ورودی باشد یعنی $N \gg p$. وقتی که(as2) صادق نباشد، اصلاح RKM در بخش 4 توسعه داده شده است. بر اساس محدودیت های (as2)-(as1)، الگوریتم کامینز مرسوم عملیات اسکالر $\mathcal{O}(NCp)$ به ازای هر تکرار را اجرا کرده و از این روی مستلزم ذخیره سازی متغیر های اسکالر $\mathcal{O}(Np)$ باشد. برای RKM، شمارش دقیق نشان می دهد که پیچیدگی زمانی به ازای هر تکرار در $\mathcal{O}(NCp)$ حفظ می شود:(13) مستلزم محاسبه فواصل اقلیدسیNC $\|\mathbf{x}_n - \mathbf{m}_c^{(t-1)} - \mathbf{o}_n^{(t-1)}\|_2^2$ می باشد و قواعد N که $\|\mathbf{o}_n^{(t-1)}\|_2$ در $\mathcal{O}(NCp)$، $\mathbf{m}_c^{(t)}$ در $\mathcal{O}(NCp)$به روز رسانی می شود در حالی که 11-12 مستلزم عملیات $\mathcal{O}(NCp)$ است. به علاوه، ملزومات حافظه RKM دارای مرتبه یکسان با ملزومات کامینز است. هم چنین توجه داشته باشید که ماتریس $N \times p$ اضافی از $\mathbf{O}$ را می توان با استفاده از ساختار های پراکنده ذخیره کرد.

تکرار RKM تحت شرایط متوسط همگرا است. دلیل این است که توالی مقادیر تابع هزینه به صورت غیر افزایشی است. چون هزینه به صورت محدود است، توالی مقدار تابع به صورت همگرا می شود. همگرایی تکرار های RKM در ادامه بررسی شده است.

فرض 2: الگوریتم RKM به ازای $q > 1$ به یک حداقل مختصات محور از(6) همگرا می شود. به علاوه، الگوریتم RKM سخت$(q = 1)$ به یک کمینه محلی از(5) همگرا می شود.

اثبات: به پیوست B مراجعه شود.

ب: الگوریتم های خوشه بندی احتمالی با ثبات

یک رویکرد EM در این زیر بخش برای انجام کمینه سازی در (8) توسعه می یابد. در صورتی که $\mathbf{U}$ معلوم باشد، پارامتر های مدل $\boldsymbol{\Theta}$ را می توان با کمینه سازی احتمال لگاریتمی منفی منظم از داده های کامل($\mathbf{X}, \mathbf{U}$) براورد کرد یعنی

$$\min_{\boldsymbol{\Theta}} -L(\mathbf{X}, \mathbf{U}; \boldsymbol{\Theta}) + \lambda \sum_{n-1}^{N} \|\mathbf{o}_n\|_{\boldsymbol{\Sigma}^{-1}} \qquad (15)$$

با این حال چون U مشاهده نمی شود، هزینه در(15) به طور نیمه بهینه با تکرار دو مرحله روش EM کمینه سازی می شود. فرض کنید که $\boldsymbol{\Theta}^{(t}$ نشان دهنده مقادیر پارامتر مدل در t امین تکرار باشد. در طی مرحله-E از t امین تکرار، امید ریاضی $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t-1)}) := \mathrm{E}_{\mathbf{U}|\mathbf{X};\boldsymbol{\Theta}^{(t-1)}}[L(\mathbf{X}, \mathbf{U}; \boldsymbol{\Theta})]$ ارزیابی می شود. چون $L(\mathbf{X}, \mathbf{U}; \boldsymbol{\Theta})$ تابع خطی از U می باشد و u_{nc} متغیر های تصادفی دو دویی است که به صورت زیر است

$$Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t-1)}) = \sum_{n=1}^{N} \sum_{c=1}^{C} \gamma_{nc}^{(t)} \left(\log \pi_c + \log \mathcal{N}(\mathbf{x}_n; \mathbf{m}_c + \mathbf{o}_n, \boldsymbol{\Sigma})\right) \qquad (17)$$

که در آن $\gamma_{nc}^{(t)} := \Pr\left(\mathbf{u}_n = \mathbf{e}_c | \mathbf{x}_n; \boldsymbol{\Theta}^{(t-1)}\right)$ می باشد. با استفاده از قاعده بیزی، احتمالات پسین $\gamma_{nc}^{(t)}$ در شکل بسته به صورت زیر ارزیابی می شود.

$$\gamma_{nc}^{(t)} = \frac{\pi_c^{(t-1)} \mathcal{N}(\mathbf{x}_n; \mathbf{m}_c^{(t-1)} + \mathbf{o}_n^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)})}{\sum_{c'=1}^{C} \pi_{c'}^{(t-1)} \mathcal{N}(\mathbf{x}_n; \mathbf{m}_{c'}^{(t-1)} + \mathbf{o}_n^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)})}. \qquad (18)$$

در طی مرحله M، $\boldsymbol{\Theta}^{(t)}$ به صورت زیر به روز رسانی می شود

$$\boldsymbol{\Theta}^{(t)} = \arg\min_{\boldsymbol{\Theta}} -Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t-1)}) + \lambda \sum_{n=1}^{N} \|\mathbf{o}_n\|_{\boldsymbol{\Sigma}^{-1}}. \qquad (19)$$

یک راهبرد BCD که هر مجموعه از پارامتر ها را در $\boldsymbol{\Theta}$ به صورت یک باره با فرض ثابت بودن همه موارد دیگر به روز رسانی می کند در ادامه توصیف شده است. ابتدا هزینه در(19) با توجه به $\boldsymbol{\pi}$ کمینه سازی می شود. با توجه به این که $\sum_{c=1}^{C} \gamma_{nc}^{(t)} = 1$ به ازای همه مقادیر n است، کمینه ساز $\sum_{n=1}^{N} \sum_{c=1}^{C} \gamma_{nc}^{(t)} \overline{\log \pi_c}$ نسبت به $\mathcal{P}$ در شکل بسته به صورت زیر نشان داده می شود.

$$\pi_c^{(t)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nc}^{(t)} \text{ for all } c \in \mathbb{N}_C. \qquad (20)$$

متعاقبا، (19) با توجه به M کمینه سازی کرد در حالی که $\boldsymbol{\pi}, \mathbf{O}, \boldsymbol{\Sigma}$ به ترتیب برابر با $\boldsymbol{\pi}^{(t)}, \mathbf{O}^{(t-1)}$ و $\boldsymbol{\Sigma}^{(t-1)}$ در نظر گرفته می شود. مراکز به صورت کمینه ساز های یک هزینه LS وزنی به روز رسانی کرد که می دهد:

$$\mathbf{m}_c^{(t)} = \frac{\sum_{n=1}^{N} \gamma_{nc}^{(t)} \left(\mathbf{x}_n - \mathbf{o}_n^{(t-1)} \right)}{\sum_{n=1}^{N} \gamma_{nc}^{(t)}} \text{ for all } c \in \mathbb{N}_C. \qquad (21)$$

سپس (19) با توجه به O با فرض ثابت بودن سایر پارامتر های مدل با مقدار به روز رسانی شده قبلی آن ها کمینه سازی می شود. این بهینه سازی نسبت به n تجزیه شده و می توان معادله زیر

$$\min_{\mathbf{o}_n} \sum_{c=1}^{C} \frac{\gamma_{nc}^{(t)}}{2} \|\mathbf{x}_n - \mathbf{m}_c^{(t)} - \mathbf{o}_n\|^2_{(\boldsymbol{\Sigma}^{(t-1)})^{-1}} + \lambda \|\mathbf{o}_n\|_{(\boldsymbol{\Sigma}^{(t-1)})^{-1}} \qquad (22)$$

به ازای همه مقادیر $n \in \mathbb{N}_N$ حل کرد. به یک کواریانس کامل $\boldsymbol{\Sigma}$، (22) را می توان به صورت یک برنامه مخروطی درجه دوم حل کرد. برای مورد خوشه های کروی، یعنی $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_p$، حل(22) به طور قابل توجهی ساده می شود. به طور اخص، هزینه را می توان به صورت $\sum_{c=1}^{C} \gamma_{nc}^{(t)} \|\mathbf{x}_n - \mathbf{m}_c^{(t)} -$

$\|\mathbf{o}_n\|_2^2 - 2\lambda\sigma^{(t-1)}\|\mathbf{o}_n\|_2$ نوشت که مشابه با هزینه در(10) به ازای $q=1$ و به ازای یک λ با مقیاس مناسب است. بر اساس راه حل(10)، $\mathbf{o}_n$ به صورت زیر

$$\mathbf{o}_n^{(t)} = \mathbf{r}_n^{(t)}\left[1 - \frac{\lambda\sigma^{(t-1)}}{\|\mathbf{r}_n^{(t)}\|_2}\right]_+ \qquad (23)$$

بعد از باز تعریف بردار باقی مانده به صورت $\mathbf{r}_n^{(t)} := \sum_{c=1}^{C}\gamma_{nc}^{(t)}(\mathbf{x}_n - \mathbf{m}_c^{(t)})$ مطابق با (12) آپدیت می شود. جالب این که، قاعده تعیین آستانه(23) نشان می دهد که $\sigma^{(t-1)}$ بر تشخیص داده های پرت اثر دارد. در حقیقت، در این شرایط احتمالی، آستانه برای شناسایی داده پرت متناسب با مقدار براورد انحراف معیار بدون داده پرت بوده و از این روی با توزیع تجربی داده ها سازگار است.

مرحله M با کمینه سازی(19) با توجه به $\boldsymbol{\Sigma}$ به ازای $\mathbf{O} - \mathbf{O}^{(t)}$ و $\boldsymbol{\pi} - \boldsymbol{\pi}^{(t)}, \mathbf{M} - \mathbf{M}^{(t)}$ پایان می یابد یعنی

$$\min_{\boldsymbol{\Sigma}\succ\mathbf{0}} \sum_{n=1}^{N}\sum_{c=1}^{C}\frac{\gamma_{nc}^{(t)}}{2}\|\mathbf{x}_n - \mathbf{m}_c^{(t)} - \mathbf{o}_n^{(t)}\|_{\boldsymbol{\Sigma}^{-1}}^2 + \frac{N}{2}\log\det\boldsymbol{\Sigma} + \lambda\sum_{n=1}^{N}\|\mathbf{o}_n^{(t)}\|_{\boldsymbol{\Sigma}^{-1}} \qquad (24)$$

برای یک $\boldsymbol{\Sigma}$ عمومی،(24) باید به طور عددی برای مثال از طریق نزول گرادیان یا روش های نقطه ای حل شود. با در نظر گرفتن خوشه های کروی به دلیل سهولت، شرط بهینگی مرتبه اول برای(24) مستلزم حل یک معادله کوادراتی در $\sigma^{(t)}$ است. با نادیده گرفتن ریشه منفی این معادله، $\sigma^{(t)}$ به صورت زیر محاسبه می شود

$$\sigma^{(t)} = \frac{\lambda}{2Np}\sum_{n=1}^{N}\|\mathbf{o}_n^{(t)}\|_2 + \sqrt{\sum_{n=1}^{N}\sum_{c=1}^{C}\frac{\gamma_{nc}^{(t)}}{Np}\|\mathbf{x}_n - \mathbf{m}_c^{(t)} - \mathbf{o}_n^{(t)}\|_2^2 + \left(\lambda\sum_{n=1}^{N}\frac{\|\mathbf{o}_n^{(t)}\|_2}{2Np}\right)^2}. \qquad (25)$$

طرح خوشه بندی احتمالی با ثبات(RPC) به صورت الگوریتم 2 جدول بندی می شود. برای خوشه های کروی، پیچیدگی آن به صورت عملیات $O(NCp)$ به ازای هر تکرار باقی می ماند در حال یکه ثابت ها بزرک تر از ثابت های موجود در الگوریتم RKM می باشند. مشابه با RKM، تکرار های RPC تحت شرایط متعادل همگرا می شوند. همگرایی تکرار های RPC در فرض زیر اثبات می شوند

الگوریتم 2: خوشه بندی احتمال گرایانه قوی

مطلوب است: ماتریس داده های ورودی $\mathbf{X}$، تعداد خوشه های C و پارامتر $\lambda > 0$

1- به طور تصادفی $\mathbf{M}^{(0)}, \boldsymbol{\pi}^{(0)} \subset \mathcal{P}$ را مقدار دهی کرده و $\boldsymbol{\Sigma}^{(0)} = \delta \boldsymbol{I}_p$ $(\sigma^{(0)} = \sqrt{\delta})$ به ازای $\delta > 0$ را در نظر گرفته و $\mathbf{O}^{(0)}$ را برابر با صفر قرار دهید.

2- به ازای $t = 1, 2, \ldots$ موارد زیر را انجام دهید

3- آپدیت $\gamma_{nc}^{(t)}$ از طریق(18) به ازای همه n.c

4- آپدیت $\boldsymbol{\pi}^{(t)}$ از طریق(20)

5- آپدیت $\mathbf{M}^{(t)}$ از طریق(21)

6- آپدیت $\mathbf{O}^{(t)}$ با حل 22 و 23

7- آپدیت $\boldsymbol{\Sigma}^{(t)}$ $(\sigma^{(t)})$ از طریق 24 و 25

8- پایان

فرض 3: تکرار های RPC به کمینه مختصات محور از احتمال لگاریتمی منفی در (8) همگرا می شوند

اثبات: به پیوست C مراجعه کنید.

فرض 3 تضمین کننده این است که تکرار های RPC همگرا هستند. با این حال چون هر جمله مشتق ناپذیر $\|\mathbf{o}_n\|_{\Sigma^{-1}}$ دارای دو متغیر بهینه سازی متفاوت $\boldsymbol{\Sigma}$ و $\mathbf{o}_n$ می باشد، تکرار BCD را می توان در کمینه محلی مختصات محور قرار دارد که لزوما یک کیمنه محلی از (8) نیست. وقتی که تکرار ها همگرا شدند، γ_{nc}

نهایی را می توان به صورت خوشه نرم تفسیر کرد که به موجب آن تخصیص سخت را می توان از طریق ماکزیمم قاعده تشخیص پسین یعنی $\mathbf{x}_n \in \mathcal{X}_c$ به ازای $c = \arg\max_{c'} \gamma_{nc'}$ بدست آورد.

تبصره 4(انتخاب λ): تعدیل λ در صورتی امکان پذیر است که اطلاعات اضافی برای مثال در مورد درصد داده های پرت قابل دسترس باشند. الگوریتم خوشه بندی با ثبات برای توالی کاهشی مقادیر λ ($\{\lambda_g\}$) با استفاده از شروع گرم(14) تا زمانی که تعداد مورد انتظار از داده های پرت شناسایی شوند اجرا می شود. هنگام حل λ_g، شروع گرم اشاره به متغیر های بهینه سازی ای دارد که به راه حل بدست امده به ازای λ_{g-1} مقدار دهی شده اند. از این روی اجرای الگوریتم در $\{\lambda_g\}$ می تواند به طور کارامد صورت گیرد زیرا تعداد کمی از تکرار های BCD به ازای λ_g برای همگرایی کافی است.

پ: الگوریتم های خوشه بندی با ثبات وزنی

همان طور که قبلا ذکر شد، روش های خوشه بندی با ثبات مطرح شده تا کنون، برای تقریب پنالتی ناپیوسته $\mathrm{I}(|\mathbf{o}_n\|_2 > 0)$ by $\|\mathbf{o}_n\|_2$، از پارادایم CS تبعیت می کنند که در آن $\mathrm{I}(|x| > 0)$ با تابع محدب $|x|$ جایگزین می شود. با این حال، استدلال بر این است که توابع غیر محدب نظیر $\log(|x| + \epsilon)$ به ازای $c > 0$ کوچک می تواند تقریب دقیق تری را از $\mathrm{I}(|x| > 0)$(34) ارایه کند. این منطق موجب می شود تا $\|\mathbf{o}_n\|_2$ را در (5)،(6) و (8) با $\log(\|\mathbf{o}_n\|_2 + c)$ برای بهبود پراکندگی بلوک در $\mathbf{o}_n$ و نیز بهبود انعطاف پذیری به داده های پرت جایگزین کنیم.

تغییر منظم سازی موجب تغییر الگوریتم های BCD تنها در زمان کمینه سازی با توجه به O می شود. این مرحله ویژه در $\mathbf{o}_n$ تجزیه می شود ولی به جای $\phi^{(t)}(\mathbf{o}_n)$ در (10)، می توان معادله زیر را کمینه سازی کرد

$$\phi_w^{(t)}(\mathbf{o}_n) := \sum_{c=1}^{C} (u_{nc}^{(t-1)})^q \times \left(\|\mathbf{x}_n - \mathbf{m}_c^{(t)} - \mathbf{o}_n\|_2^2 + \lambda \cdot \log\left(\|\mathbf{o}_n\|_2 + \epsilon\right) \right) \quad (26)$$

که دیگر محدبنیست. بهینه سازی در(26) با استفاده از یک تکرار از رویکرد کمینه سازی- بیشینه سازی(MM)(25) انجام می شود. هزینه $\phi_w^{(t)}(\mathbf{o}_n)$ با تابع $f^{(t)}(\mathbf{o}_n;\mathbf{o}_n^{(t-1)})$ بیشینه سازی می شود بدین معنی که $\phi_w^{(t)}(\mathbf{o}_n) \leq f^{(t)}(\mathbf{o}_n;\mathbf{o}_n^{(t-1)})$ به ازای هر $\mathbf{o}_n$ و وقتی که $\phi_w^{(t)}(\mathbf{o}_n) = f^{(t)}(\mathbf{o}_n;\mathbf{o}_n^{(t-1)})$ $\mathbf{o}_n = \mathbf{o}_n^{(t-1)}$ می باشد.سپس $f^{(t)}(\mathbf{o}_n;\mathbf{o}_n^{(t-1)})$ با توجه به $\mathbf{o}_n$ برای بدست اوردن $\mathbf{o}_n^{(t)}$ کمینه سازی می شود.

به منظور یافتن یک بیشینه ساز برای $\phi_w^{(t)}(\mathbf{o}_n)$، تحدب لگاریتم استفاده می شود یعنی این که $\log x \leq \log x_o + \frac{x}{x_o} - 1$ به ازای هر x و x_o مثبت می باشد. با استفاده از نامساوی اخیر برای پنالتی و نادیده گرفتن جملات ثابت، کمینه سازی زیر را خواهیم داشت

$$f^{(t)}(\mathbf{o}_n;\mathbf{o}_n^{(t-1)}) := \sum_{c=1}^{C}(u_{nc}^{(t-1)})^q\left(\|\mathbf{x}_n - \mathbf{m}_c^{(t)} - \mathbf{o}_n\|_2^2 + \lambda_n^{(t)}\|\mathbf{o}_n\|_2\right) \quad (27)$$

که $\lambda_n^{(t)} := \frac{\lambda}{(\|\mathbf{o}_n^{(t-1)}\|_2+\epsilon)}$ می باشد.

مقایسه 27-10 نشان می دهد که منظم سازی جدید منجر به یک نسخه وزنی از مقدار اصلی می شود. تنها تفاوت بین الگوریتم های با ثبات مطرح شده قبلی و همتاهای وزنی آن، تعریف λ_n می باشد. در تکرار t، مقادیر بزرگ تر برای $\|\mathbf{o}_n^{(t-1)}\|_2$ منجر به آستانه های کوچک تر در قواعد تعیین آستانه(11-23) می شود که به موجب آن، $\mathbf{o}_n$ به صورت غیر صفر انتخاب می شود. الگوریتم های خوشه بندی با ثبات وزنی، $\mathbf{o}_n^{(0)}$ را به مقدار $\mathbf{o}_n$ از الکوریتم غیر وزنی مقدار دهی می کند. از این روی، برای اجرای RKM وزنی برای یک مقدار خاص λ_n RKM باید ابتدا اجرا شود. سپس، RKM وزنی با همه متغیر های مقدار دهی شده برای مقادیر بدست آمده با RKM ولی با $\lambda_n^{(1)}$ که قبلا تعریف شد اجرا می شوند.

مرحله MM ترکیب شده با الگوریتم های BCD توسعه یافته در این جا تحت مفروضات حداقل همگرا است. برای همین منظور،توجه کنید که توالی های مقادیر هدف برای RKM و RPC هر دو به صورت غیر افزایشی هستند. چون توابع هزینه مربوطه در زیر نشان داده شده اند، توالی ها به صورت همگرا در می آیند. تعیین نقاط و سرعت همگرایی در حوزه و فضای این مقاله نمی گنجد.

4- خوشه بندی داده های تفکیک پذیر غیر خطی و با بعد بالا

الگوریتم های خوشه بندی با ثبات استفاده شده تا کنون به طور کلی شامل عملیات $\mathcal{O}(NCp)$ به ازای هر تکرار بوده اند. با این حال، چندین کاربرد مستلزم خوشه بندی داده های نسبا کم ولی با بعد بالا در حضور داده های پرت می باشند در کاربرد های مربوط به تصویر برداری، می توان تصاویر $N = 500$ با $p = 800 \times 600 = 480,000$ را خوشه بندی کرد در حالی که در تحلیل ریز آرایه DNA، برخی از ده ها نمونه DNA(که به طور بالقوه دارای خطا بوده یا به ندرت رخ می دهند) بر اساس سطوح بیان خود در هزاران ژن(18) خوشه بندی می شود. در این سناریو های خوشه بندی که در آن ها $p \gg N$ می باشد، یک روش کارامد باید از ذخیره و پردازش داده های بعدی p(12) اجتناب کند. برای این منظور، الگوریتم های بخش 3 در این جا(29) هسته سازی می شوند. نشان داده شده است که این الگوریتم های هسته ای به ازای هر تکرار و فضای $\mathcal{O}(N^2)$ نیازمند عملیات $\mathcal{O}(N^3C)$ می باشند و از این روی وقتی که $p > N^2$ است مطلوب نمی باشند. این هسته سازی نه تنها ذخیره پردازشی را در رژیم داده های با بعد باال(بخش 4 الف) در اختیار می گذارد بلکه امکان شناسایی خوشه های تفکیک پذیر غیر خطی را می دهد(بخش 4 ب).

الف: کامینز با ثبات برای داده های با بعد بالا

بدون از دست دادن عمومیت، تاکید اصلی هسته ای سازی الگوریتم کامینز قوی می باشد. ماتریس $N \times C$ ، $\mathbf{U}_q$ با درایه $[\mathrm{U}_q]_{nc} := u_{nc}^q$ و گرامیان $\mathbf{K} := \mathbf{X}'\mathbf{X}$ تشکیل شده با همه محصولات درونی زوجی بین بردار های ورودی را در نظر بگیرید. اگرچه محاسبه هزینه های $\mathcal{O}(N^2p)$، k این تنها یک بار انجام می شود. توجه داشته باشید که آپدیت های 9، 11 و 13 مستلزم محصولات درونی بین بردار های p بعدی $\{\mathbf{o}_n, \mathbf{r}_n\}_{n=1}^N$ و $\{\mathbf{m}_c\}_{c=1}^C$ می باشد. در صورتی که $\{\mathbf{v}_i \in \mathbb{R}^p\}_{i=1}^2$ زوجی از این بردار ها باشد، هزینه

برای محاسبه $\mathbf{v}_1^T\mathbf{v}_2$ مشخصا $O(p)$ خواهد بود. با این حال، در صورتی که همه این بردار ها در $\text{range}(\mathbf{X})$ قرار گیرند، یعنی در صورتی که $\{\mathbf{w}_i \in \mathbb{R}^N\}_{i=1}^2$ وجود داشته باشد به طوری که $\{\mathbf{v}_i = \mathbf{X}\mathbf{w}_i\}_{i=1}^2$، آنگاه $\mathbf{v}_1^T\mathbf{v}_2 = \mathbf{w}_1^T\mathbf{K}\mathbf{w}_2$ است و محصول درونی را می توان به طور متناوب در $O(N^2)$ محاسبه کرد. با استناد به این مشاهده، ابتدا نشان داده می شود که همه بردار های $p \times 1$ در $\text{range}(\mathbf{X})$ قرار دارد. اثبات به صورت استقرایی است: در صورتی که در $(t-1)$ امین تکرار هر $\mathbf{o}_n^{(t-1)} \in \text{range}(\mathbf{X})$ و $\mathbf{U}^{(t-1)} \in \mathcal{U}_2$، باشد، نشان داده می شود که $\mathbf{o}_n^{(t)}, \mathbf{m}_c^{(t)}, \mathbf{r}_n^{(t)}$ آپدیت شده با RKM در دامنه $\text{range}(\mathbf{X})$ نیز قرار می گیرد.

فرض کنید که در t امین تکرار، ماتریس $\mathbf{U}^{(t-1)}$ تعریف کننده $\mathbf{U}_n^{(t-1)}$ ، در $\mathcal{U}_2$ قرار می گیرد در حالی که ماتریس $\mathbf{A}^{(t-1)}$ وجود دارد به طوری که $\mathbf{O}^{(t-1)} = \mathbf{X}\mathbf{A}^{(t-1)}$ است. سپس، آپدیت مراکز در(9) را می توان به صورت زیر بیان کرد

$$\mathbf{M}^{(t)} = (\mathbf{X} - \mathbf{O}^{(t-1)})\mathbf{U}_q^{(t-1)}\text{diag}^{-1}((\mathbf{U}_q^{(t-1)})^T\mathbf{1}_N) \quad (28)$$
$$= \mathbf{X}\mathbf{B}^{(t)} \quad (29)$$

که

$$\mathbf{B}^{(t)} := (\mathbf{I}_N - \mathbf{A}^{(t-1)})\mathbf{U}_q^{(t-1)}\text{diag}^{-1}((\mathbf{U}_q^{(t-1)})^T\mathbf{1}_N). \quad (30)$$

قبل از آپدیت $\mathbf{O}^{(t)}$، بردار های باقی مانده $\mathbf{r}_n$ باید از طریق(12) به روز رسانی شوند. با الحاق باقیمانده ها در $\mathbf{R}^{(t)} := [\mathbf{r}_1^{(t)} \dots \mathbf{r}_N^{(t)}]$، آپدیت در (12) را می توان در شکل ماتریس به صورت زیر بازنویسی کرد

$$\mathbf{R}^{(t)} = \mathbf{X} - \mathbf{M}^{(t)}(\mathbf{U}_q^{(t-1)})^T\text{diag}^{-1}(\mathbf{U}_q^{(t-1)}\mathbf{1}_C) \quad (31)$$
$$= \mathbf{X}\mathbf{\Delta}^{(t)} \quad (32)$$

که

$$\mathbf{\Delta}^{(t)} := \mathbf{I}_N - \mathbf{B}^{(t)}(\mathbf{U}_q^{(t-1)})^T\text{diag}^{-1}(\mathbf{U}_q^{(t-1)}\mathbf{1}_C). \quad (33)$$

از معادله 11، هر $\mathbf{o}_n^{(t)}$ یک نسخه مقیاس بندی شده از $\mathbf{r}_n^{(t)}$ بوده و مقیاس بندی بستگی به $\|\mathbf{r}_n^{(t)}\|_2$ دارد. بر اساس(31)، مورد اخیر را می توان به صورت $\|\mathbf{r}_n^{(t)}\|_2 = \sqrt{(\boldsymbol{\delta}_n^{(t)})^T \mathbf{K} \boldsymbol{\delta}_n^{(t)}} = \|\boldsymbol{\delta}_n^{(t)}\|_{\mathbf{K}}$، محاسبه کرد که در آن $\boldsymbol{\delta}_n^{(t)}$ نشان دهنده n امین ستون از $\boldsymbol{\Delta}^{(t)}$ است. با استفاده از اپراتور تعیین آستانه، می توان به آپدیت زیر رسید:

$$\mathbf{O}^{(t)} = \mathbf{X}\mathbf{A}^{(t)} \qquad (34)$$

که در آن n امین ستون $\mathbf{A}^{(t)}$ با معادله زیر بدست می اید

$$\boldsymbol{\alpha}_n^{(t)} = \boldsymbol{\delta}_n^{(t)} \left[1 - \frac{\lambda}{2\,\|\boldsymbol{\delta}_n^{(t)}\|_{\mathbf{K}}} \right]_+, \quad \forall n. \qquad (35)$$

بعد از اثبات مرحله قیاسی با(34)، استدلال در صورتی کامل است که متغیر های داده های پرت $\mathbf{O}$ به صورت $\mathbf{O}^{(0)} = \mathbf{X}\mathbf{A}^{(0)}$ به ازای مقداری از $\mathbf{A}^{(0)}$ از جمله مقدار دهی عملی و معنی دار در صفر ، مقدار دهی شوند. نتیجه اثبات شده در زیر نشان داده شده است.

فرض 4: با انتخاب $\mathbf{O}^{(0)} - \mathbf{X}\mathbf{A}^{(0)}$ به ازای هر مقدار $\mathbf{A}^{(0)} \subset \mathbb{R}^{N \times N}$ و $\mathbf{U}^{(0)} \in \mathcal{U}_2$، ستون های متغیر های ماتریس $\mathbf{O}, \mathbf{M}, \mathbf{R}$ به روز رسانی شده با RKM در دامنه $\mathrm{range}(\mathbf{X})$ قرار می گیرد یعنی $\mathbf{O}^{(t)} - \mathbf{X}\mathbf{A}^{(t)}$ $\mathbf{M}^{(t)} - \mathbf{X}\mathbf{B}^{(t)}$ $\mathbf{R}^{(t)} - \mathbf{X}\boldsymbol{\Delta}^{(t)}$ معلوم وجود به طوری که $\mathbf{A}^{(t)}, \mathbf{B}^{(t)}, \boldsymbol{\Delta}^{(t)}$ به ازای همه مقادیر t است.

آن چه که بایستی هسته ای شود، آپدیت ها برای تخصیص های خوشه می باشد. برای مرحله آپدیت(13) یا(14)، باید $\|\mathbf{x}_n - \mathbf{m}_c^{(t)} - \mathbf{o}_n^{(t)}\|_2^2$ و $\|\mathbf{o}_n^{(t)}\|_2$ را محاسبه کرد. با توجه به این که $\mathbf{x}_n = \mathbf{X}\mathbf{e}_n$ است،که در آن $\mathbf{e}_n$ به معنی n امین ستون از $\mathbf{I}_N$ می باشد و بر اساس آپدیت های هسته ای(28) و (34)، می توان به آسانی اثبات کرد که:

$$\|\mathbf{x}_n - \mathbf{m}_c^{(t)} - \mathbf{o}_n^{(t)}\|_2^2 = |\mathbf{X}(\mathbf{e}_n - \boldsymbol{\beta}_c^{(t)} - \boldsymbol{\alpha}_n^{(t)})\|_2^2 \quad (36)$$
$$- |\mathbf{e}_n - \boldsymbol{\beta}_c^{(t)} - \boldsymbol{\alpha}_n^{(t)}\|_{\mathbf{K}}^2 \quad (37)$$

و این به ازای هر n و c است که در آن $\boldsymbol{\beta}_c^{(t)}$، c امین ستون از $\mathbf{B}^{(t}$ است. همانند(34)، می توان داشت:

$$\|\mathbf{o}_n^{(t)}|_2 = \|\mathbf{X}\boldsymbol{\alpha}_n^{(t)}\|_2 = \|\boldsymbol{\alpha}_n^{(t)}\|_{\mathbf{K}}. \quad (38)$$

الگوریتم کامینز با ثبات هسته ای(KRKM) به صورت الگوریتم 3 خلاصه سازی شده است. در خصوص RKM، الگوریتم KRKM به صورت $\frac{\mathbf{M}^{(t)}-\mathbf{M}^{(t-1)}\|_F}{\|\mathbf{M}^{(t)}\|_F} \leq \epsilon_s$ به پایان رسیده و یا به طور معادل ناشی از(28) می باشد

$$\frac{\left(\sum_{c=1}^{C} \|\boldsymbol{\beta}_c^{(t)} - \boldsymbol{\beta}_c^{(t-1)}|_{\mathbf{K}}^2\right)}{\left(\sum_{c=1}^{C} \|\boldsymbol{\beta}_c^{(t)}|_{\mathbf{K}}^2\right)} \leq \epsilon_s^2$$

این معادله به ازای $\epsilon_s > 0$می باشد. KRKM مستلزم عملیات $\mathcal{O}(N^3C)$ در هر تکرار است در حالی که متغیر های ذخیره شده $\mathbf{A}, \mathbf{B}, \boldsymbol{\Delta}, \mathbf{U}_q$ و $\mathbf{K}$ فضای $\mathcal{O}(N^2)$ را اشغال می کنند. توجه کنید که اگر مراکز M صریحا مورد نیاز باشد(برای مثال برای اهداف تفسیری)،آن ها را می توان از طریق(28) بعد از اتمام الگوریتم KRKM بدست آورد.

الگوریتم 3: RKM هسته ای

مطلوب است: ماتریس گرامیان $\mathbf{K} \succ \mathbf{0}$، تعداد خوشه های C، $q \geq 1$ و $\lambda > 0$

1- مقدار دهی تصادفی $\mathbf{U}^{(0}$ در $\mathcal{U}_2$ و صفر قرار دادن $\mathbf{A}^{(0)}$

2- به ازای $t = 1, 2 \ldots$ موارد زیر را انجام دهید

3- آپدیت $\mathbf{B}^{(t)}$ از 30

4- آپدیت $\boldsymbol{\Delta}^{(t}$ از 33

5- آپدیت $\mathbf{A}^{(t)}$ از 35

6- آپدیت $\mathbf{U}^{(t)}$ و $\mathbf{U}_q^{(t)}$ از (13) یا (14)، (36) و (38)

7- پایان

ب:RKM هسته ای برای خوشه های تفکیک پذیر غیر خطی

به دلیل فاصله اقلیدسی مورد استفاده، کامینز استاندارد فرض می کند که خوشه ای اصلی دارای شکل کروی بوده و به طور خطی تفکیک پذیر هستند: خوشه بندی مبتنی بر GMM این محدودیت را نیز دارد. کامینز هسته ای این مانع را با نقشه یابی بردار های $\mathbf{x}_n$ با فضای بعدی بالاتر $\mathcal{H}$ برطرف می کند که موسوم به فضای ویژگی از طریق تابع هدف $\varphi : \mathbb{R}^p \rightarrow \mathcal{H}$ است(30). داده های نقشه یابی شده $\{\varphi(\mathbf{x}_n)\}_{n=1}^N$ دارای بعد $p' > p$ بوده و یا حتی نامتناهی است. کامینز در نسخه مرکزی خود به $\varphi(\mathbf{x}_n)$ اعمال می شود. از این روی، پارتیشن ها یا بخش های قابل تفکیک خطی در فضای ویژگی امکان تقسیم قابل تفکیک غیر خطی را در فضای داده های اصلی می دهد.

برای این که یک الگوریتم هسته ای پذیر باشد، محصولات درونی $\varphi^T(\mathbf{x}_n)\varphi(\mathbf{x}_m)$ باید به آسانی قابل محاسبه باشد. وقتی که نقشه خطی $\varphi(\mathbf{x}_n) = \mathbf{x}_n$ به طور ناچیز فرض شود، این محصولات درونی به سادگی، درایه های گرامیان $\mathbf{X}^T\mathbf{X}$ می باشند. وقتی که نقشه غیر خطی استفاده شود، ماتریس هسته K با درایه های $[\mathbf{K}]_{n,m} := \varphi^T(\mathbf{x}_n)\varphi(\mathbf{x}_m)$ جایگزین ماتریس گرامیان شده و باید معلوم باشد. بر اساس تعریف،K نیمه متناهی مثبت بوده و می تواند برای خوشه بندی استفاده شود و حتی زمانی که $\varphi(\mathbf{x}_n)$ دارای بعد بالا (بخش 4الف)، بعد نامتناهی یا حتی مجهول باشد(13). یک مورد جالب در شرایطی وجود دارد که در آن $\mathcal{H}$ فضای هیبلرت هسته می باشد. سپس، محصول درونی در $\mathcal{H}$ توسط تابع هسته معلوم $\kappa(\mathbf{x}_n, \mathbf{x}_m) := \varphi^T(\mathbf{x}_n)\varphi(\mathbf{x}_m)$ ارایه می شود(29، فصل 3). هسته های معمول برای داده بردار به صورت چند جمله ای و گاوسی است: هسته ها را می توان برای اشیای غیر برداری نیز تعریف کرد نظیر رشته ها یا گراف ها(29).

با استناد به KRKM توسعه یافته در بخش 4 الف، اکنون مدیریت هسته های دلخواه ساده نیست. با معلوم بودن X و هسته $\kappa(\mathbf{x}_n, \mathbf{x}_m)$، ماتریس K را می توان به آسانی محاسبه کرد. با استفاده از هسته مطابق با گرامیان،

الگوریتم 3 به آسانی به رژیم خوشه بندی غیر خطی اعمال می شود. با این حال به خاطر داشته باشید که بر عکس داده های با بعد بالای خوشه بندی، در خوشه بندی غیر خطی(با ثبات)، مراکز را نمی توان به طور کلی محاسبه کرد: حتی اگر بخواهیم مرکز فضای ویژگی را بازیابی کنیم، پیش تصویر فضای ورودی آن ممکن است وجود نداشته باشد(29، فصل 18).

ج: خوشه بندی احتمالی با ثبات هسته ای

هسته ای سازی RPC مانع از ایجاد اختلاف عمده در هسته ای سازی RKM می شود: آپدیت های GMM و RPCدر بخش 3 ب برای بردار های ویژگی زمانی معتبر هستند که تنها بعد P آن ها متناهی و معلوم باشند. اهمیت آن را می توان به شکل زیر توضیح داد. ابتدا، آپدیت واریانس در (25) مستلزم بعد p می باشد که وقتی در معرض هسته ای سازی قرار گرفت تبدیل به P می شود. دوما، ترکیبات(داده پرت-آگاه) از الگوریتم های گاوسی زمانی که در معرض مدل سازی بردار های تصادفی بعد نامتناهی قرار گیرد کاهش می یابد. به منظور غلبه بر این محدودیت، می توان از مفهوم نقشه هسته تجربی استفاده کرد(29، فصل 2.2.6) با توجه به بردار های ورودی در $\mathcal{X}$ و ماتریس هسته آن ها K، امکان جایگزینی φ بانقشه هسته تجربی $\hat{\varphi}$ وجود دارد: $\mathbb{R}^p \longrightarrow \mathbb{R}^N$ به صورت $\hat{\varphi}(\mathbf{x}) := (\mathbf{K}^{\frac{1}{2}})^{\dagger}[\kappa(\mathbf{x}_1, \mathbf{x}) \cdots \kappa(\mathbf{x}_N, \mathbf{x})]^T$ تعریف می شود که در آن $(\cdot)^{\dagger}$نشان دهنده شبه اینورس های مور-پنروس می باشند. فضای ویژگی $\hat{\mathcal{H}}$ ناشی از $\hat{\varphi}$ دارای بعدیت متناهی N می باشد در حالی که می توان تایید کرد $\hat{\varphi}^T(\mathbf{x}_n)\hat{\varphi}(\mathbf{x}_m) - \varphi^T(\mathbf{x}_n)\varphi(\mathbf{x}_m) - \kappa(\mathbf{x}_n, \mathbf{x}_m)$ به ازای همه مقادیر $\mathbf{x}_n, \mathbf{x}_m \in \mathcal{X}$ است.

در شرایط احتمالی هسته ای، $\hat{\varphi}(\mathbf{x}_n)$ از ترکیبی از توزیعات گاوسی چند متغیره C با $\boldsymbol{\Sigma} - \sigma^2 \boldsymbol{I}_N$ که در همه خوشه ها مشترک هستند گرفته می شود. آپدیت های مبتنی بر EM از RPC در بخش 3 ب بعد از جایگزینی بعد p در(25) با N و بردار های ورودی $\mathbf{x}_n$ با $\hat{\varphi}(\tilde{\mathbf{x}}_n)$که محصولات درونی آن درایه های K هستند به طور معتبر باقی می مانند. روش هسته ای سازی مشابه با روش در نظر گرفته شده برای RKM می باشد: ابتدا، ماتریس های کمکی $\mathbf{A}^{(t)}, \mathbf{B}^{(t)}$ و $\boldsymbol{\Delta}^{(t)}$ معرفی می شوند. با مقدار دهی تصادفی با $\sigma^{(0)}, \boldsymbol{\pi}^{(0)} \in \mathcal{P}, \mathbf{B}^{(0)} \in \mathbb{R}^{N \times C}$ و صفر قرار دادن $\mathbf{A}^{(0)}$، همان طور که در فرض 4 مطرح است، می

توان نشان داد که آپدیت های RPC هسته ای برای $\mathbf{R}^{(t)}$ و $\mathbf{O}^{(t)}, \mathbf{M}^{(t)}$ دارای ستون هایی در دامنه $\mathrm{range}(\mathbf{\Phi})$، می باشند که در آن $\mathbf{\Phi} := [\phi(\mathbf{x}_1) \cdots \phi(\mathbf{x}_N)]$ است.به جای ماتریس تخصیص U در KRKM، ماتریس $N \times C$ براورد های احتمال پسین $\mathbf{\Gamma}^{(t)}$ استفاده می شود که در آن $[\mathbf{\Gamma}^{(t)}]_{nc} := \gamma_{nc}^{(t)}$ دارای شرط $\mathbf{\Gamma}^{(t)} \mathbf{1}_C = \mathbf{1}_N \ \forall t$ است. الگوریتم RPC هسته ای به صورت الگوریتم 4 خلاصه سازی می شود. همانند KRKM، محاسبات آن $\mathcal{O}(N^3 C)$ به ازای هر تکرار است در حالی که متغیر های ذخیره شده $\mathbf{A}, \mathbf{B}, \mathbf{\Delta}, \mathbf{\Gamma}, \boldsymbol{\pi}, \mathbf{K},$ و σ فضای $\mathcal{O}(N^2)$ را اشغال می کند.

الگوریتم 4: RPC هسته ای

مطلوب است ماتریس هسته یا گرامیان $\mathbf{K} \succ 0$ ، تعداد خوشه های C و $\lambda > 0$

1- مقدار دهی تصادفی $\sigma^{(0)}, \boldsymbol{\pi}^{(0)} \in \mathcal{P}$ و $\mathbf{B}^{(0)}$ ، و صفر قرار دادن $\mathbf{A}^{(0)}$

2- به ازای $t = 1, 2, \ldots$، موارد زیر را انجام دهید

3- آپدیت $\mathbf{\Gamma}^{(t)}$ از طریق 18 با استفاده از 36

4- آپدیت $\boldsymbol{\pi}^{(t)}$ به صورت $\boldsymbol{\pi}^{(t)} = (\mathbf{\Gamma}^{(t)})^T \frac{\mathbf{1}_N}{N}$

5- آپدیت $\mathbf{B}^{(t)}$ به صورت $\mathbf{B}^{(t)} = (\mathbf{I}_N - \mathbf{A}^{(t-1)}) \mathbf{\Gamma}^{(t)} \mathrm{diag}^{-1}(N \boldsymbol{\pi}^{(t)})$

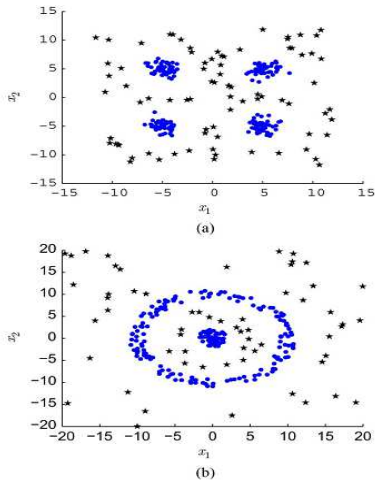
6- آپدیت $\mathbf{\Delta}^{(t)}$ به صورت $\mathbf{\Delta}^{(t)} - \mathbf{I}_N - \mathbf{B}^{(t)} (\mathbf{\Gamma}^{(t)})^T$

7- آپدیت ستون های $\mathbf{A}^{(t)}$ به صورت $\boldsymbol{\alpha}_n^{(t)} - \boldsymbol{\delta}_n^{(t)} \left[1 - \frac{\lambda \sigma^{(t-1)}}{\|\boldsymbol{\delta}_n^{(t)}\|_{\mathbf{K}}}\right]_+$ به ازای همه مقادیر n

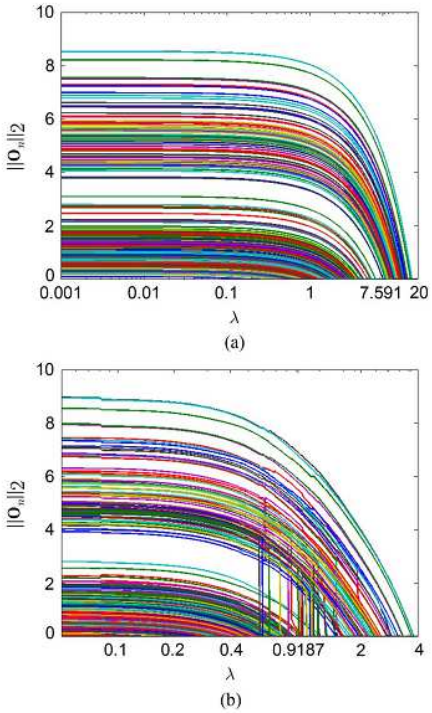
8- آپدیت $\sigma^{(t)}$ از طریق(25) که در آن p با N با استفاده از قواعد ℓ_2 محاسبه شده در مرحله 3 و با استفاده از $\|\mathbf{o}_n^{(t)}\|_2 - \|\boldsymbol{\alpha}_n^{(t)}\|_{\mathbf{K}}$ به ازای همه مقادیر n جایگزین می شود.

9- پایان

تبصره 5: (الگوریتم های هسته ای وزن دهی مجدد): همانند بخش 3 پ، نسخه های مجددا وزن دهی شده از KRKM و KRPC را می توان به سادگی با معرفی یک پارامتر وابسته به تکرار $\lambda_n^{(t)} = \frac{\lambda}{(\|\mathbf{o}_n^{t-1}\|_2 + c)}$ مشتق کرد.



شکل 1: مجموعه داده های مصنوعی: بردار های غیر داده پرت با دوایر توپر نشان داده شده اند(ستاره).الف: مجموعه داده با خوشه های کروی $C = 1$ و 80 داده پرت.ب: مجموعه داده با $C = 2$ حلقه متحد المرکز و 60 داده پرت



شکل 2:منحنی های $\|o_n\|_2$ به صورت تابعی از λ برای مجموعه داده ها در شکل 1 الف. الف: الگوریتم RKM به ازای $q = 1$ ب: الگوریتم RPC

5- تست های عددی

تست های عددی نشان دهنده عملکرد الگوریتم های جدید در هر دو مجموعه داده های مصنوعی و واقعی در این بخش ارایه می شوند. عملکرد از طریق توانایی آن ها برای شناسایی داده های پرت و کیفیت خود خوشه بندی ارزیابی می شود. کیفیت خوشه بندی به خودی خود از طریق شاخص رند تعدیل یافته(ARI) بین خوشه بندی حاصله و برچسب های واقعی داده ها(21) اندازه گیری می شود. برای روش هایی که قادر به شناسایی داده های پرت نمی باشند،ARI به طور اجتناب ناپذیری در همه داده ها محاسبه می شود. برای روش های با قابلیت های تشخیص داده های پرت،ARI بعد از اجرای داده های پرت محاسبه می شود. در هر آزمایش، λ با استفاده از جست و جوی شبکه مطرح شده در تبصره 4 تعدیل می شود. با استفاده از روش شروع گرم، مسیر راه حل برای همه نقاط شبکه در مقدار زمان مشابه با زمان مورد استفاده برای حل برای یک مقدار خاص λ محاسبه می شود.

الف: مجموعه داده های مصنوعی

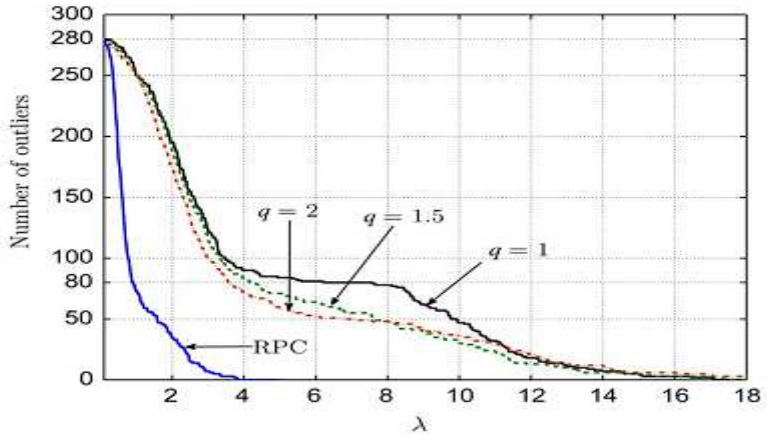
دو مجموعه داده مصنوعی استفاده شد. مجموعه داده نخست که در شکل 1 الف نشان داده شده است، متشکل از انتخاب تصادفی 200 بردار از توزیعات گاوسی دو متغیره $C = 4$ (50 بردار به ازای هر توزیع) و 80 بردار داده پرت($N = 280$) می باشد.توزیعات گاوسی دارای میانگین های مختلف و یک ماتریس کواریانس مشترک $0.8 \cdot \mathbf{I}_2$ می باشد. دومین مجموعه داده متشکل ازنقاط متعلق به حلقه های هم مرکز $C = 2$ که در شکل 1 ب نشان داده شده است می باشد. حلقه درونی(بیرونی) دارای 50 (150) نقطه بود. این هم چنین شامل 60 نقطه بود که بین حلقه ها و خارج از حلقه های خارجی متناظر با داده های پرت قرار گرفته اند($N = 260$). خوشه بندی این دومین مجموعه داده ها چالش بر انگیز است حتی اگر داده های پرت به دلیل شکل یا ماهیت چند مقیاس خوشه ها موجود نباشند.

اثر λ بر روی تعداد داده های پرت شناسایی شده برای مجموعه داده های با خوشه های کروی شناسایی شد. در شکل 2، مقادیر $\{\|o_n\|_2\}_{n=1}^{N}$ به صورت تابعی از λ نشان داده شده است(تبصره 4). منحنی قاعده داده پرت

نشان داده شده در شکل 2(الف) با الگوریتم RKM با $q = 1$ با استفاده از یک مقدار دهی تصادفی متناظر است. برای $\lambda > 17$ ، همه $\mathbf{o}_n$ برابر با صفر قرار داده شدند. با نزدیک شدن λ به صفر، تعداد بیشتری از $\mathbf{o}_n$ ها مقادیر غیر صفر را اختیار کردند. انتخاب $\lambda \in [6.2, 7.6]$ منجر به 80 داده پرت شد. شکل 2(ب) $\{\|\mathbf{o}_n\|_2\}_{n=1}^{N}$ را با تغییر λ برای الگوریتم RPC با فرض $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$ نشان می دهد. توجه داشته باشید که مسیر های دنبال شده با $\|\mathbf{o}_n\|_2$ با کاهش λ یک تغییر سریع را از صفر نشان می دهد. با تمرکز بر این نقاط، به طور تجربی مشاهده شد که وقتی آن ها دارای $\mathbf{o}_n$ صفر بودند، γ_{nc} پیشین آن ها برای تخصیص عضویت مبهم بود. با کاهش λ به طوری که $\|\bar{\mathbf{o}}_n\|_2 > 0$ باشد، یکی از γ_{nc} آن ها سریعا به سمت یک میل می کند در حالی که سایرین سریعا به صفر کاهش می یابند از این روی موجب افزایش سریع $\|\bar{\mathbf{o}}_n\|_2$ تا یک مقدار متناهی می شود(23). لازم به ذکر است که این رفتار مستلزم ناپایداری یا اختلال در شناسایی داده های پرت نیست.

در شکل 3، تعداد نقاط شناسایی شده به صورت داده های پرت یعنی تعداد $\|\mathbf{o}_n\|_2$ غیر صفر به صورت تابعی از λترسیم می شود. این منحنی ها هنگامی که مقدار λ برای شناسایی یک تعداد خاصی از داده های پرت برابر با صفر قرار داده شود مفید خواهد بود. هدف اصلی شناسایی داده های پرت $s = 80$ بود. هر دو RKM و RPC با λ تعدیل شده برای شناسایی 80 داده پرت، قادر به خوشه بندی صحیح داده ها و شناسایی داده های پرت بودند. اگرچه بدست اوردن منحنی ها در شکل 3 نیازمند حل مسائل خوشه بندی با ثبات مختلف است، که برای هر مقدار λ یکی در نظر گرفته می شود، آن ها را می توان به طور موثربا استفاده از شروع گرم که در تبصره 4 توصیف شد محاسبه کرد.

برای این آزمایش، شکل 3 نیز براورد تعداد داده های پرت را در مجموعه داده با بررسی شیب های منحنی نشان داد. هنگام کاهش λ برای RKM سخت، برای $\lambda \in [6.2, 7.6]$ یک منحنی صاف ایجاد شد و بعد از آن یک منحنی با شیب زیاد دیده شد.



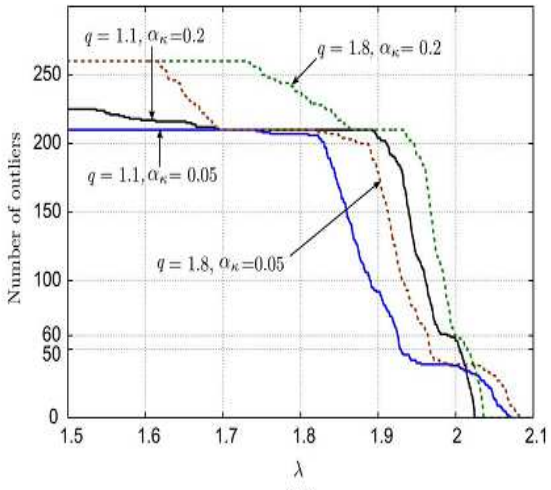
شکل 3: تعداد داده های پرت شناسایی شده به صورت تابعی از λ برای مجموعه داده در شکل 1(الف)، این حالت مسطح و بدون شیب،معرف یک منطقه گذار بین شناسایی صحیح داده های پرت و تعیین اشتباه داده های غیر پرت به صورت داده های پرت است. تغییرات شیب منحنی مشابه برای RKM و RPC نرم اطراف مقادیر λ که منجر به تعداد صحیحی از 80 نقطه پرت شد مشاهده شده است

تعداد داده های پرت

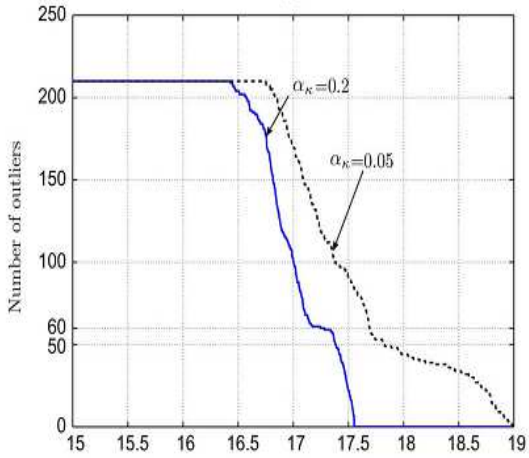
خطای جذر میانگین ریشه(RMSE) بین مراکز خوشه براورد شده برای روش های خوشه بندی و میانگین نمونه برای هر خوشه به عنوان یک معیار شایستگی استفاده شد. جدول 1 مینیمم RMSE بدست آمده را در 100 مقدار تصادفی برای چندین مقادیر از آلودگی داده پرت نشان می دهد. همه الگوریتم های تست شده، مقادیر مشترکی را داشتند. الگوریتم های تست شده RKM(WRKM) و RPC(WRPC) وزنی، کامینز سخت، کامینز نرم با $q = 1.5$، EM، خوشه بندی نویز(NC)(10)، آلفا کات(36) و خوشه بندی احتمالی(24) بودند. فاصله نویز در NC انتخاب شد به طوری که خوشه نویز و تعداد داده های پرت و پارامتر های تعدیل کننده برای خوشه بندی احتمالی برای همه خوشه ها برابر با 2 قرار داده شد. RKMو RPC نسبت به همتا های غیر با ثبات خود و آلفا کات به RMSE کم تری دست یافتند و قادر به شناسایی صحیح داده های پرت در همه موارد بودند. بهبود قابل توجهی با WRKM و WRPC مشاهده شد که بهترین عملکرد را در میان همه الگوریتم های تست شده نشان داد. توجه داشته باشید که ARI الگوریتم های خوشه بندی با ثبات جدید متناظر با مقادیر RMSE در جدول 1 برابر با یک بود. نکته جالب این است که NC اکتشافی، عملکرد خوشه بندی رقابتی بعد از تعدیل دقیق

پارامتر آن را ارایه کرد. اگرچه خوشه بندی احتمالی نیز عملکرد رقابتی را ارایه می کند، به طور تجربی مشاهده شد که به مقدار دهی و تعدیل پارامتر حساس است.

سپس، مجموعه داده های با دوایر هم مرکز نشان داده شده در شکل 1(ب) با استفاده از هسته گاوسی $\kappa(\mathbf{x}_n, \mathbf{x}_m) = \exp(-\alpha_\kappa \|\mathbf{x}_n - \mathbf{x}_m\|_2^2)$ خوشه بندی شد که در آن $\alpha_\kappa > 0$ یک پارامتر مقیاس بندی است. پارامتر α_κ^{-1} به عنوان براورد واریانس با ثبات از مجموعه داده های کل توصیف شده در (8) انتخاب شد. هر KRKM و KRPC قادر به شناسایی 60 نقطه پرت بود. در شکل 4 نمودار تعداد داده های پرت شناسایی شده با KRKM و KRPC به صورت تابعی از λ برای مقادیر مختلف α_κ ترسیم شده است. شکل 5 مقادیر $\|\mathbf{o}_n\|_2$ را برای WKRKM و WKRPC هنگام جست و جوی 60 داده پرت نشان می دهد. نقاط احاطه شده توسط یک دایره متناظر با بردار های شناسایی شده به صورت داده های پرت وهر شعاع دایره با مقدار $\|\mathbf{o}_n\|_2$ متناظر آن متناسب است.



(a)



(b)

تعداد داده های پرت

شکل 4: تعداد داده های پرت شناسایی شده به صورت تابعی از λ برای مجموعه داده در شکل 1 ب، الف: الگوریتم KRKM، ب: الگوریتم KRPC

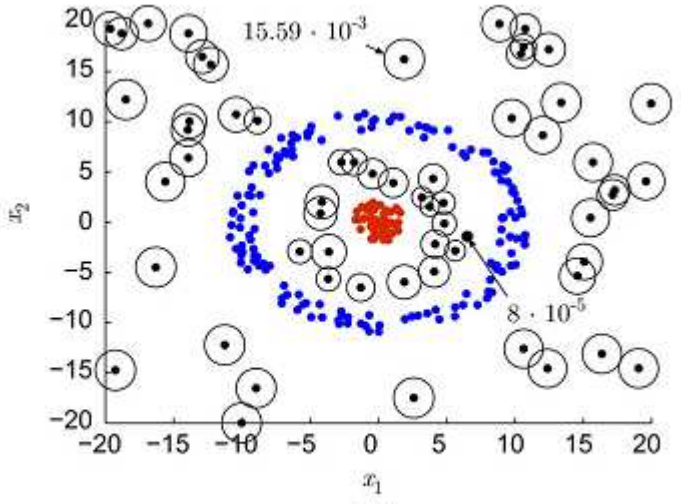
جدول 1: عملکرد RMSE الگوریتم های خوشه بندی برای مجموعه داده با خوشه های کروی C=4

	RMSE				
Outliers/N	**10**/210	**20**/220	**40**/240	**60**/260	**80**/280
hard K-means	0.6002	0.7713	1.2009	1.3927	1.5856
soft K-means ($q = 1.5$)	0.5156	0.6546	1.2006	1.5014	1.4558
EM	0.6003	0.7607	1.1267	1.2961	1.5271
hard RKM	0.2505	0.3660	0.6242	0.800	1.0126
hard WRKM	**0.0710**	**0.0627**	**0.0739**	**0.0461**	**0.0723**
soft RKM ($q = 1.5$)	0.2162	0.2129	0.3170	0.3706	0.4981
soft WRKM ($q = 1.5$)	**0.0521**	**0.0389**	**0.0304**	**0.0359**	**0.0407**
RPC	0.2984	0.3393	0.4483	0.5597	0.6652
WRPC	**0.0366**	**0.0572**	**0.0019**	**0.0029**	**0.0615**
α-cut ($\alpha = 0.5$)	0.6002	0.7713	1.2070	1.4264	1.6646
α-cut ($\alpha = 0.7$)	0.6078	0.7754	1.1671	1.3603	1.6911
α-cut ($\alpha = 0.9$)	0.5375	0.7006	1.1196	1.3088	1.6190
soft NC ($q = 1.5$)	0.0479	0.0598	0.0526	0.0493	0.0696
possibilistic ($q = 1.5$)	0.3207	0.3208	0.3207	0.3202	0.3201

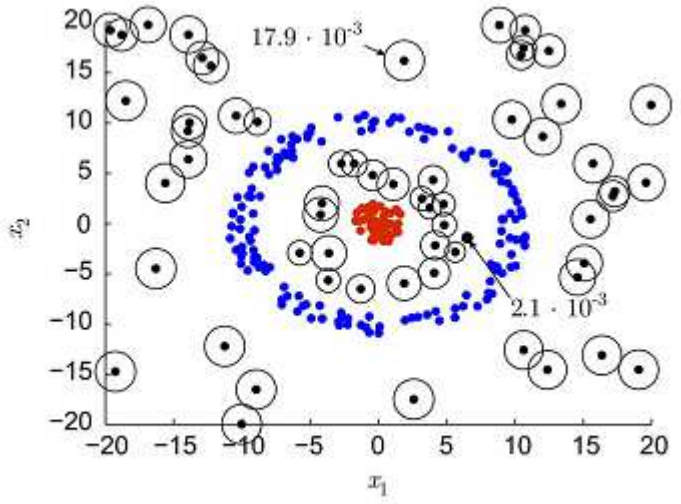
داده های پرت، کامینز سخت، کامینز نرم،RKM سخت، WRKM سخت، آلفا کات، احتمالی

ب: مجموعه داده های USPS

در این زیر بخش، الگوریتم های خوشه بندی با ثبات بر روی مجموعه تشخیص اعداد دست نوشته(USPS) سرویس پستی ایالات متحده تست شدند. این مجموعه حاوی تصاویر رقومی مقیاس خاکستری با 16×16 پیکسل با شدت نرمال شده با $[-1, 1]$ بودند.



(a)



(b)

شکل 5: نتایج خوشه بندی مجموعه داده در شکل 1(ب) با استفاده از یک هسته گوسی با $\sigma_k = 0.2$. نقاط احاطه شده با یک دایره به صورت داده های پرت در نظر گرفته شدند، شعاع دایره با مقدار $\|\mathbf{o}_k\|_2$ متناسب است. کوچک ترین و بزرگ ترین مقادیر $\|\mathbf{o}_k\|_2$ نشان داده شده است. الف: الگوریتم KRKM $(q = 1.1)$ ب: الگورتیم KRPC

جدول 2: ARI برای مجموعه داده‌USPS(C=6)

هسته	کامینز	کامدین	RKM	RPC

خطی	0.6469	0.5382	0.6573	0.6508
چند جمله ای	0.5571		0.6978	0.6965

این به 7201 نمونه آموزشی و 2007 نمونه آزمایشی از رقم های بین 0-9 تقسیم شد. اگرچه مجموعه دارای برچسب های کلاسی بود، آن ها ظاهرا متناقض بودند: برخی ارقام به اشتباه برچسب گذاری شده بودند در حالی که برخی از تصاویر حتی توسط انسان به سختی قابل طبقه بندی بودند(29، برنامه A). در این آزمایش، زیر مجموعه ای از ارقام 0-5 استفاده شدند. برای هر رقم، هر دو مجموعه داده های آموزشی و آزمایشی به یک مجموعه ترکیب شده و سپس 300 تصویر به طور یکنواخت و به شکل تصادفی نمونه برداری شده و منجر به یک مجموعه داده با 1800 تصویر شد. هر تصویر با یک بردار 256 بعدی نرمال شده دارای یک قاعده ℓ_2 واحد شد.

$\mathrm{RKM}\ (q = 1)$ و RPC برای پارتیشن بندی و تقسیم مجموعه داده ها به $C = 6$ خوشه و شناسایی $s = 100$ داده پرت استفاده شد. مجموع 20 مونته کارلو با مقدار دهی تصادفی که در همه الگوریتم ها مشترک است اجرا می شود. خوشه بندی نهایی به صورت یک خوشه بندی با کم ترین هزینه در (5) انتخاب شد. ARI برای کامینز، کامدین و طرح های پیشنهادی در جدول 2 نشان داده شده است. هر دوی RKM و RPC بهبود عملکرد خوشه بندی را نسبت به الگوریتم های غیر با ثبات نشان داد. هم چنین، ARI بدست آمده توسط WRKM(WRPC) برابر با ARI بدست آمده با RKM(RPC) می باشد. توجه داشته باشید که الگوریتم کامدین قادر به یافتن یک پارتیشن بندی برای داده های نشان دهنده 6 رقم موجود حتی بعد از 100 تکرار مونته کارلو بود.

مجموعه داده های USPS با استفاده از RKM و WRKM تعدیل شده برای شناسایی 100 داده پرت خوشه بندی شد.WRKM با نتایج بدست آمده با RKM مقدار دهی شد.اگرچه RKM و WRKM منجر به تصاویرپرت یکسانی می شود، اندازه w_n متفاوت بود و به این ترتیب برای WRKM به صورت یکنواخت در نظر گرفته شد. مجموعه داده های USPS نیز با استفاده از الگوریتم های RPC و WRPC خوشه بندی شد. شکل 6 (الف) مراکز خوشه ای بدست آمده با RPC و WRPC را نشان می دهد. شکل 6(ب) 100 داده پرت شناسایی شده را نشان می دهد. داده های پرت شناسایی شده توسط الگوریتم های RPC و WRPC نیز منطبق با هم می باشند.

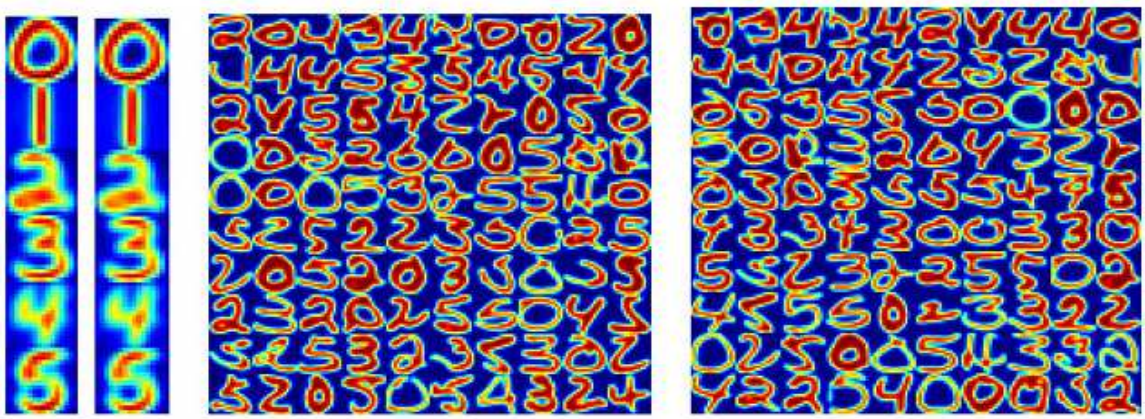
موقعیت تصاویر پرت در موزاییک متناظر با رتبه آن ها بر اساس اندازه $\mathbf{0}_{\mathbf{D}}$ متناظر آن ها می باشد(بزرگ ترین تا کوچک ترین از چپ به راست، بالا به پایین). توجه داشته باشید که همه داده های پرت شناسایی شده دارای یک صفتی می باشند که آن ها را از تصویر میانگین در هر خوشه متمایز می کند. در میان 100 داده پرت شناسایی شده توسط RKM و RPC،97 مورد در هر دو رویکرد مشترک بودند.

نسخه های هسته ای الگوریتم ها نیز در مجموعه داده های USPS استفاده شدند. مشابه با(29)، هسته چند جمله ای همگن مرتبه سوم، یعنی $\kappa(\bar{\mathbf{x}}_m, \bar{\mathbf{x}}_m) - (\mathbf{x}_n^T \bar{\mathbf{x}}_m)^3$ استفاده شد. امتیازات ARI بدست آمده با الگوریتم های خوشه بندی با ثبات هسته ای در جدول 2 نشان داده شده اند. بر اساس این امتیازات، دو مشاهده مهم وجود دارد: 1- کامینز هسته ای به داده های پرت حساس تر از کامینز است ولی 2- KRKM برای هسته خاص به یک عملکرد خوشه بندی بهتر نسبت به RKM منتهی می شود. در نهایت، 100 داده پرت شناسایی شده توسطKRKM در شکل 6 پ نشان داده شده است.

پ: خوشه بندی سند

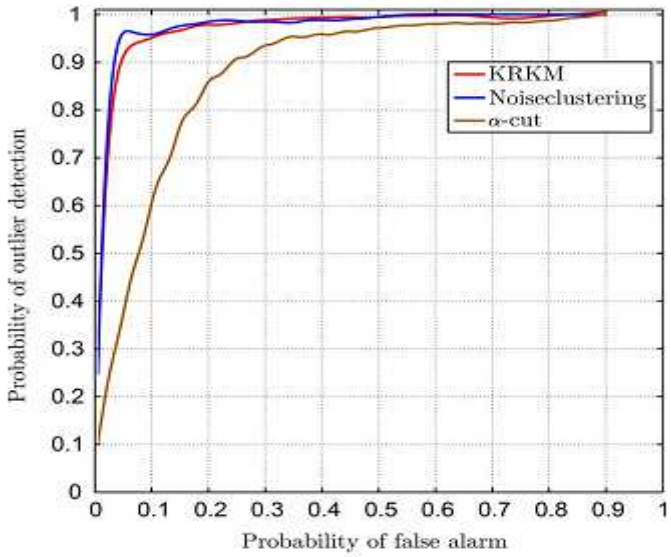
الگوریتم KRKMتوسعه یافته در بخش 4 الف در ادامه ارزیابی شده است. زمینه یک خوشه بندی سند در مجموعه داده TDT2 استاندارد می باشد. مجموعه TDT2 متشکل از داده های جمع اوری شده در طی نیمه اول 1998 از شش منبع خبری(6)می باشد. این شامل اسناد طبقه بندی شده به 96 مقوله معنایی است. هر سند با یک بردار حاوی تعداد دفعاتی است که هر یک از $p = 36\,771$ در سند اتفاق می افتند. بعد از کنار گذاشتن اسناد تخصیص داده شده به بیش از یک مقوله، 10212 بردار سند حاصل می شود که متعاقبا طوری نرمال سازی می شود که دارای قاعده ℓ_2 واحد باشد. مقوله ها متغیر از 1 تا 1844 سند می باشد. برای هر آزمایش خوشه بندی، 4 مورد از 19 بزرگ ترین مقوله به طور تصادفی انتخاب می شود و 100 سند از هر یک از این 4 مقوله به طور یکنواخت نمونه برداری شدند. یک نمونه تصادفی افزایشی شامل 20 سند از 30 مقوله موچک شامل داده های پرت افزایشی بود که منجر به مجموعه $N = 420$ شد. با پنهان کردن برچسب های مقوله از الگوریتم ها، وظیفه اصلی خوشه بندی مشترک 400 سند به 4 مقوله بزرگ و شناسایی 20 سند

استخراج شده از مقوله های کوچک تر به صورت داده های پرت است. هسته گاوسی با پهنای باند واحد استفاده می شود یعنی $[\mathbf{K}]_{n,m} = \exp(-||\mathbf{x}_n - \mathbf{x}_m|\frac{2}{2})$ و سند ها به $C = 4$ خوشه تقسیم می شوند(6). KRKM سخت(الگوریتم 3) با NC و آلفا کات(10، 36) مقایسه می شود. الگوریتم های تست شده برای پارتیشن بندی به شکل زیر مقدار دهی می شود: ماتریس $N \times C$ حاوی بردار های ویژه متناظر با بزرگ ترین مقادیر C از K می باشد که ورودی الگوریتم کامینز استاندارد است که ردیف N آن را به کلاس های C تخصیص می دهد(12). معیار شایستگی در این جا، قابلیت تشخیص اسناد پرت می باشد. برای هر مجموعه داده نمونه، همه سه الگوریتم برای یک شبکه از مقادیر برای پارامتر های تعدیل کننده آن ها اجرا می شود(λ برای KRKM، فاصله تا خوشه نویز در NC و درصد θ برای روش آلفا کات). از این روی، برای هر مجموعه داده، جفت احتمال تشخیص صحیح و خطای هشدار بدست می آید که متناظر با نقاط منحنی مشخصه عملکرد سیستم(یا منحنی عملیاتی دریافت کنندهROC) می باشد. منحنی های ROC از 50 مجموعه داده نمونه از طریق اسپلاین های هموار با پارامتر 0.9999 میانگین گیری شده و نمودار نتایج در شکل 7 نشان داده شده است.KRKM عملکرد تشخیص مشابه را با NC نشان می دهد اگرچه هر دوی آن ها عملکرد بهتری از روش آلفا کات داشتند.



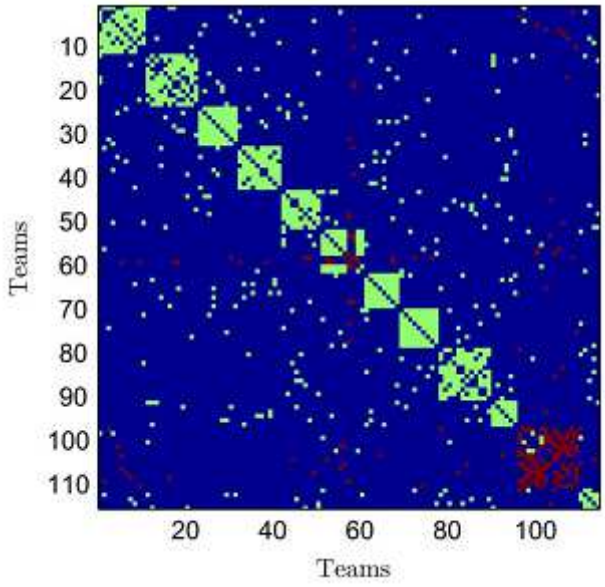
(a) (b) (c)

شکل6: خوشه بندی و داده های پرت برای مجموعه داده USPS با $C = 6$تعدیل شده برای شناسایی$s = 100$ داده پرت الف: مراکز RPC و WRPC، ب: داده های پرت شناسایی شده با RPC و WRPC پ: داده های پرت شناسایی شده با KRKM با استفاده از هسته چند جمله ای مرتبه سوم



شکل 7: شناسایی اسناد داده پرت در مجموعه TDT2

احتمال تشخیص داده پرت(محور افقی)، احتمال هشدار کاذب(محور عمودی)



شکل 8: ماتریس هسته برای شبکه فوتبال کالج جایگشتی با استفاده از خوشه بندی KRKM. درایه های صفر دارای رنگ آبی و داده های پرت با رنگ قرمز نشان داده شده اند.

لغات داخل شکل: تیم ها

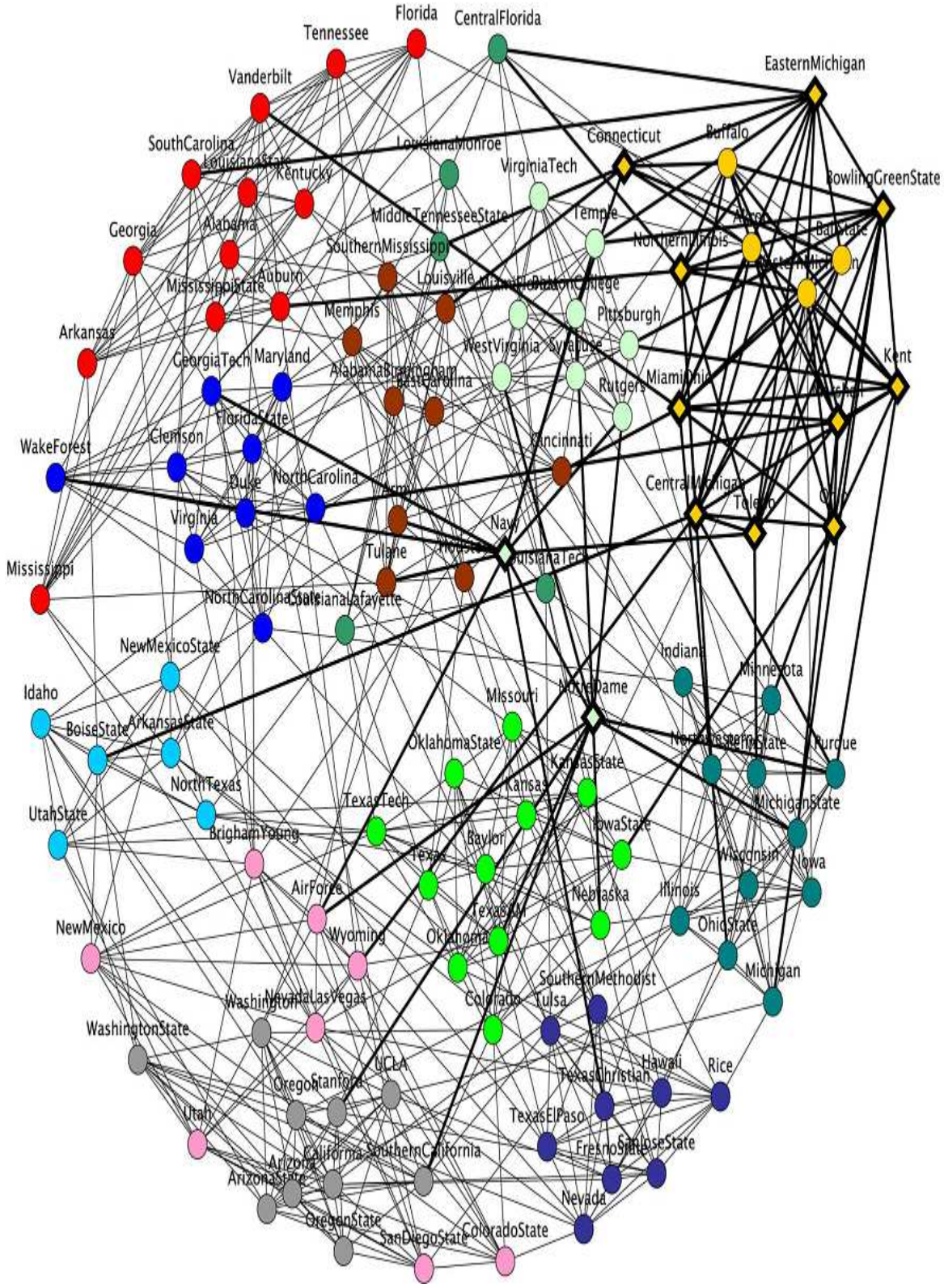
ت: شبکه فوتبال کالج

KRKM برای پارتیشن بندی و شناسایی داده های پرت در یک شبکه $N = 115$ تیم فوتبال کالج که در 12 کنفرانس مختلف برای بازی های بخش 1 در طی سال 2000 بازی می کردند استفاده شد(17). در این جدول زمانی، تیم ها اغلب در برابر تیم های همان کنفرانس بازی می کردند. هر گره در شبکه متناظر با یک تیم می باشد و یک لینک بین دوتیم در صورتی وجود دارد که آن ها در برابر هر تیم دیگر در طی فصل بازی می کردند. ساختار شبکه با ماتریس مجاورت E $N \times N$ خلاصه شده است.

به منظورشناسایی گروه ها و داده های پرت، ارتباط بین کامینز هسته و خوشه بندی طیفی برای پارتیشن بندی گراف استفاده شد(12). برطبق این ارتباط، الگوریتم خوشه بندی طیفی مرسوم با کامینز هسته ای با استفاده از ماتریس هسته خاص جایگزین می شود. ماتریس هسته مورد استفاده $\mathbf{K} = \nu \bar{I_N} + \bar{\mathrm{D}}^{\frac{-1}{2}}\mathbf{E}\bar{\mathrm{D}}^{\frac{-1}{2}}$ بود که در آن $\mathbf{D} := \mathrm{diag}(\mathbf{E}1_N)$ بوده و ν طوری انتخاب شد که $\mathbf{K} \succ \mathbf{0}$ باشد. تیم ها به $C = 12$ گروه تقسیم شدند. KRKM از طریق خوشه بندی طیفی مقدار دهی شده و λ برای شناسایی $s = 12$ داده پرت تعدیل شد. شکل 8 درایه های ماتریس هسته $\mathbf{K}$ را بعد از جایگشت ردیف و ستون نشان می دهد به طوری که تیم ها دریک خوشه ظاهرا گروه بندی می شوند.ARI بدست آمده با KRKM برابر با 0.9218 بود.

تیم های شناسایی شده به صورت داده های پرت مرتب شده به صورت نزولی بر اساس مقادیر $\|\mathbf{o}_{v_i}\|_2$ به شکل زیر می باشد: کانکتیکات، نیروی دریایی، نوتردام، ایلینوی شمالی، تولدو، میامی (اوهایو)، بولینگ گرین استیت، میشیگان مرکزی، میشیگان شرقی، کنت، اوهایو، و مارشال. سه مورد از آن ها یعنی کانکتیکات، نیروی دریایی، نوتردام به صورت تیم های مستقل بودند. کانکتیکات به کنفرانس آمریکای میانه تخصیص داده شد ولی بازی های زیادی با تیم های مربوط به این کنفرانس(4 بازی) همانند سایر تیم های مربوط به همین کنفرانس(حدود 8 بازی) انجام دادند. نوتردام و نیرویدریایی تعداد بازی برابری را با تیم های مربوط به دو کنفرانس متفاوت بازی کردند به طوری که آن ها به یک گروه متفاوت تخصیص داده شدند. چندین تیم مربوط به کنفرانس آمریکای میانه به صورت داده های پرت طبقه بندی شدند. به طور کلی این موضوع را می توان با تقسیم بندی کنفرانس به کنفرانس های آمریکای شرقی و غرب میانه توجیه کرد. تیم ها در هر یک از زیر کنفرانس های آمریکای میانه ، تعداد بازی یکسانی با تیم های مربوط به زیر کنفرانس خود و بقیه تیم ها داشتند. نکته جالب این که زیر بخش

کنفرانس آمریکای میانه با استفاده از KRKM با $C - 13$ ضمن جست و جوی 12 داده پرت شناسایی شد. در این مورد، ARI برای بخش مربوطه برابر با 0.9110 بود. سه تیم مستقل کانکتیکات، نوتردام و نیروی دریایی مجددا در میان 12 داده پرت شناسایی شده بودند.



شکل 9: خوشه بندی شبکه فوتبال کالج بدست آمده با KRKM برای C=12. داده های پرت با گره های الماسی شکل نشان داده شده اند.

6-نتیجه گیری

الگوریتم ها برای خوشه بندی با ثبات بر ساسا مدل داده های اصلی که داده های پرت را پوشش می دهد ایجاد شدند. هر دو شرایط خوشه بندی تقسیمی احتمالی و قطعی بر اساس الگوریتم های کامینز و مبتنی بر GMM در نظر گرفته شدند. با استناد به این حقیقت که داده های پرت به ندرت در داده ها یافت می شوند، یک ارتباط با الگوریتم های سیگنال پراکندگی-آگاه ایجاد شد. این موضوع منجر به توسعه الگوریتم های کارامد از نظر محاسباتی و الگوریتم های خوشه بندی با ثبات و همگرا شد. نسخه های هسته ای الگوریتم ها که برای داده های با بعدیت بالا مناسب هستند و یا زمانی که اطلاعات تشابه تنها در میان اشیا قابل دسترس باشد، نیز ایجاد شدند. عملکرد الگوریتم های خوشه بندی قوی از طریق ازمایشات عددی بر روی هر دو مجموعه داده های مصنوعی و واقعی اعتبار سنجی شدند.

پیوست A: اثبات قضیه

اثبات فرض1

چون $\sum_{c-1}^{C}(u_{nc}^{(t)})^q > 0$ به ازای همه مقادیر n و t به دلیل (c3) می باشد، اولین جمع وند از $\bar{\phi}^{(t)}(\mathbf{o}_n)$ در(10) یک تابع محدب $\mathbf{o}_n$ می باشد.از این روی، $\phi^{(t)}(\mathbf{o}_n)$ یک تابع شدید محدب بوده و کمینه ساز آن منحصر به فرد است. سپس، به خاطر داشته باشید که یک بردار $\mathbf{o}_n^{(t)}$ ، کمینه ساز (10)است اگر و تنها اگر $\mathbf{0} \in \partial\phi^{(t)}(\mathbf{o}_n^{(t)})$ می باشد که در آن $\partial\phi^{(t)}(\mathbf{o}_n)$ زیر دیفرانسیل $\phi^{(t)}(\mathbf{o}_n)$ می باشد. برای $\mathbf{o}_n \neq \mathbf{0}$، که هزینه در(10) مشتق پذیر است $\partial\phi^{(t)}(\mathbf{o}_n)$ به طور ساده گرادیان است. در $-2\sum_{c=1}^{C}(u_{nc}^{(t-1)})^q\left(\mathbf{x}_n - \mathbf{m}_c - \left(1+\frac{\lambda}{2\|\mathbf{o}_n\|_2}\right)\mathbf{o}_n\right)$ $\mathbf{o}_n = \mathbf{0}$، زیر دیفرانسیل قاعده ℓ_2 $\|\mathbf{o}_n\|_2$ مجموعه ای از بردار های $\{\mathbf{v}_n : \|\mathbf{v}_n\|_2 \leq 1\}$ می باشد و سپس زیر دیفرانسیل $\phi^{(t)}(\mathbf{o}_n)$ is $\partial\phi^{(t)}(\mathbf{o}_n) = \left\{-2\sum_{c\ 1}^{C}(u_{nc}^{(t-1)})^q\left(\mathbf{x}_n - \mathbf{m}_c - \frac{\lambda}{2}\mathbf{v}_n\right) : \|\mathbf{v}_n\|_2 \leq 1\right\}$ می باشد.

وقتی که کمینه ساز $\mathbf{o}_n^{(t)}$ غیر صفر باشد. شرط $\mathbf{0} \in \partial\phi^{(t)}(\mathbf{o}_n^{(t)})$ به طور ضمنی نشان می دهد که

$$\left(1 + \frac{\lambda}{2\|\mathbf{o}_n^{(t)}\|_2}\right)\mathbf{o}_n^{(t)} = \mathbf{r}_n^{(t)} \qquad \text{(A.39)}$$

است به طوری که $\mathbf{r}_n^{(t)}$ در (12) تعریف شده است. معادله(A39) نشان می دهد که $\mathbf{o}_n^{(t)}$ یک نسخه مقیاس بندی شده مثبت از $\mathbf{r}_n^{(t)}$ می باشد. مقیاس بندی را می توان به آسانی با قرار دادن قاعده ℓ_2 در دو طرف A.39 مشاهده کرد یعنی $\|\mathbf{o}_n^{(t)}\|_2 = \|\mathbf{r}_n^{(t)}\|_2 - \frac{\lambda}{2}$ که برای $\|\mathbf{r}_n^{(t)}\|_2 > \frac{\lambda}{2}$ معتبر است. با جایگزینی این معادله در A.39، $\mathbf{o}_n^{(t)} = \mathbf{r}_n^{(t)}\left(1 - \frac{\lambda}{2\|\mathbf{r}_n^{(t)}\|_2}\right)$ حاصل می شود.

به ازای $\mathbf{o}_n^{(t)} = \mathbf{0}$ ، یک $\mathbf{v}_n^{(t)}$ وجود دارد که برای آن $\|\mathbf{v}_n^{(t)}\|_2 \leq 1$ و $\mathbf{r}_n^{(t)} = \left(\frac{2}{\lambda}\right)\mathbf{v}_n^{(t)}$ صادق است. این در صورتی محتمل است که $\|\mathbf{r}_n^{(t)}\|_2 \leq \frac{\lambda}{2}$ باشد. این دو مورد برای $\mathbf{o}_n^{(t)}$ از طریق(11) بیان می شود

پیوست ب

اثبات فرض 2

با تعریف $f_s(c)$ به صورت صفر در زمانی که استدلال بولین c به صورت حقیقی باشد و در غیر این صورت با ∞ قرار دادن آن، مسئله در(6) را می توان در یک شکل غیر مقید یا نامحدود نوشت.

$$\min_{\mathbf{M},\mathbf{O},\mathbf{U}} \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}^{q}\left(\|\mathbf{x}_n - \mathbf{m}_c - \mathbf{o}_n\|_2^2 + \lambda\|\mathbf{o}_n\|_2\right) + f_s(\mathbf{U} \in \mathcal{U}_2). \qquad \text{(B.40)}$$

هزینه در (B.40) که آن را به صورت $f(\mathbf{M}, \mathbf{O}, \mathbf{U})$ تعریف می کنیم، یک تابع مناسب و نیمه پیوسته می باشد که به طور ضمنی نشان می دهد که مجمو سطح نا تهی بسته می شود. هم چنین چون f به صورت اجباری می شود، مجموع سطح آن به صورت محدود یا کران در تعیین می شود. از این روی، مجموعه سطوح ناتهی از f فشرده می باشند. به ازای $q > 1$، تابع $f(\mathbf{M}, \mathbf{O}, \mathbf{U})$ دارای یک کمینه ساز منحصر به فرد به ازای هر متغیر

بلوک بهینه سازی $\mathbf{M}, \mathbf{O},$ و $\mathbf{U}$ است. سپس، همگرایی الگوریتم RKM به نقطه کمینه مختصات محور از(6) از(33،4.1) تبعیت می کند.

وقتی که $q - 1$ است، اولین جمع وند را در (B.40) به صورت $f_0(\mathbf{M}, \mathbf{O}, \mathbf{U}) := \sum_{n-1}^{N} \sum_{c-1}^{C} u_{nc} \|\mathbf{x}_n - \mathbf{m}_c - \mathbf{o}_n\|_2^2$ تعریف کنید که یک بخش مشتق پذیر از F است. تابع f_0 دارای یک دامنه باز است و بخش غیر مشتق پذیر باقی مانده f با توجه به بلوک های بهینه سازی قابل تفکیک است. از این روی مجددا بر اساس 33، می توان گفت که الگوریتم RKM با $q = 1$ به یک کمینه محلی $(\mathbf{M}^*, \mathbf{O}^*, \mathbf{U}^*)$ از(6) همگرا می شود.

تا کنون نشان داده شده است که به ازای $q = 1$، یک تکرار BCD به یک کمینه محلی از(6) همگرا می شود. مرحله BCD برای آپدیت $\mathbf{U}$ یک قاعده سخت در(14) است. از این روی، این الگوریتم BCD 1- یک $\mathbf{U}^*$ با درایه های دو دویی را می دهد و 2- ضرورتا، آپدیت های BCD را برای حل(5) اجرا می کند. چون یک کمینه محلی از(6) با تخصیص های دو دویی، نیز یک کمینه محلی از(5) می باشد،ادعای مطرح شده در فرضیه صادق است.

اثبات فرض 3

ترکیب دو مرحله الگوریتم EM، یعنی(18-19)، می توان به راحتی تایید کرد که الگوریتم با یک توالی از تکرار های BCD برای بهینه سازی زیر معادل است

$$\min_{\mathbf{\Gamma}, \boldsymbol{\Theta}'} - \sum_{n\ 1}^{N} \sum_{c\ 1}^{C} \gamma_{nc} \log \left(\frac{\pi_c \mathcal{N}\left(\mathbf{x}_n; \mathbf{m}_c + \mathbf{o}_n, \mathbf{\Sigma}\right)}{\gamma_{nc}} \right) \qquad \text{(C.41)}$$

$$+ \lambda \sum_{n-1}^{N} \|\mathbf{o}_n\|_{\mathbf{\Sigma}^{-1}} + f_s(\mathbf{\Gamma} \subset \mathcal{U}_2) + f_s(\boldsymbol{\pi} \subset \mathcal{P}) + f_s(\mathbf{\Sigma} \succ 0)$$

که در آن $\boldsymbol{\Theta}' := \{\boldsymbol{\pi}, \mathbf{M}, \mathbf{O}, \mathbf{\Sigma}\}$ ماتریس $N \times C$ از $\mathbf{\Gamma}$ دارای درایه های $[\mathbf{\Gamma}]_{nc} := \gamma_{nc} > 0,$ می باشد و همانند (B.40) که $f_s(c)$ وقتی که شرط C صادق باشد صفر است، و در غیر این صورت ∞ می باشد. این که $\{\gamma_{nc}\}$ مثبت است، بعد از استفاده از قاعده بیزی برای

استدلال این که $\tilde{\gamma}_{nc} \propto \pi_c \mathcal{N}(\mathbf{x}_n; \mathbf{m}_c + \mathbf{o}_n, \tilde{\mathbf{\Sigma}})$ می باشد و در نظر گرفتن این که 1- $\mathcal{N}(\mathbf{x}_n; \mathbf{m}_c + \mathbf{o}_n, \mathbf{\Sigma})$ به ازای همه مقادیر $\mathbf{x}_n$ مثبت است و 2- همه مقادیر π_c باید مثبت باشد به طوری که هزینه در (C.14) به صورت متناهی باشد، اثبات خواهد شد.

تابع هدف این مسئله کمینه سازی، یک تابع نیمه پیوسته پایین، کردار دار پایین زیر و مناسب است، یعنی مجموع سطح ناتهی آن بسته است. چون این تابع نیز اجباری است، مجموع سطح آن به صورت کران دار است. از این روی مجموع سطح ناتهی فشرده است. به علاوه، تابع هدف دارای کمینه سازی منحصر به فرد برای بلوک بهینه سازی $\boldsymbol{\pi}, \mathbf{M}, \mathbf{O}$ و $\sum_{n=1}^{N} \tilde{\gamma}_{nc} > 0$, است. به ویژه، کمینه ساز بلوک $\mathbf{M}$ منحصر به فرد است زیرا $c \in \mathbb{N}_C$ به ازای همه مقادیر است. سپس بر اساس(33، قضیه 4.1 c)، الگوریتم RPC به نقطه کمینه مختصات محور (7) همگرا می شود.