انجمن کامپیوتر ایران
Computer Society of Iran

مرکز توسعه فناوری نیرو (متن)

ترکیب بردارهای ویژه در حوزه KPCA به منظور خوشه بندی داده ها

سهیلا اشک زری طوسی
دانشگاه فردوسی مشهد- گروه کامپیوتر
Soheila.Ashkezari@stu-mail.um.ac.ir

هادی صدوقی یزدی
دانشگاه فردوسی مشهد- گروه کامپیوتر
h-sadoghi@um.ac.ir

چکیده: خوشه بندی به عنوان یکی از تکنیکهای مهم در شناسایی الگو، پردازش تصویر و داده کاوی شناخته می شود. در فضاهایی با ابعاد بالا به علت وجود وابستگیهای غیر خطی بین ویژگیها، الگوریتمهای خوشه بندی معمولا با شکست مواجه می شوند. برای مقابله با این مشکل عموما با انتقال فضا به حوزه ویژگی- با ابعاد بالا سعی در به دست آوردن ویژگیهایی مناسب تر برای توصیف داده می شود. در این مقاله سعی شده است تا با انتقال داده ها به فضای Kernel PCA به توصیف مناسب تری از داده ها دست یابیم و از خصوصیات این فضا برای خوشه بندی بهتر استفاده نماییم. برای این منظور، پس از استخراج ویژگیهای جدید در فضای متعامد KPCA، به بررسی آنها پرداخته و ویژگیهای مناسب برای خوشه بندی را با انتساب وزن مناسب به آنها، انتخاب و ترکیب می نماییم و در انتها از روشی مبتنی بر رای گیری وزندار برای خوشه بندی داده ها استفاده می نماییم. نتایج آزمایشها بهبود مناسبی را در مقایسه با الگوریتمهای خوشه بندی فازیC-میانگین[1] و K-میانگین[1] نشان میدهد.

واژه های کلیدی: فضای تعامد، تحلیل مولفه های اصلی کرنل، خوشه بندی

۱- معرفی

تحلیل مولفه های اصلی[1] روشی قدرتمند برای استخراج ساختار از مجموعه داده های با ابعاد بالا می باشد. در این روش با حل یک مسئله مقادیر ویژه و یا با استفاده از الگوریتمهای تکراری، مولفه های اصلی استخراج می شوند. در حقیقت، PCA تبدیلی متعامد از دستگاه مختصاتی است که داده ها را توصیف می کند. دستگاه مختصات جدید از تصویر داده ها بر روی محورهایی که اصطلاحا محورهای اصلی[2] داده ها نامیده می شوند، به دست می آید. اغلب تعداد کمی از این مولفه های اصلی برای توصیف ساختارهای موجود در داده ها کافی است. اما در مواقعی که بین متغیرها در فضای ورودی وابستگیهایی با مرتبه های بالا وجود داشته باشد، استفاده از روش PCA کارا نخواهد بود. در این موارد استفاده از PCA در فضای متغیرها و یا ویژگیها که به طور غیر خطی به فضای ورودی وابسته است و از آن به عنوان Kernel PCA یاد می شود، ترجیح داده می شود[۱].

از جمله مزیتهای این فضای جدید ، استخراج ویژگیهای بیشتر با کیفیتی مناسب تر نسبت به مولفه های اصلی می باشد به صورتیکه برای دستیابی به دقتی برابر با دقت روش PCA، در KPCA به تعداد مولفه های کمتری نیاز داریم[۲].

در اغلب کارهایی که با هدف خوشه بندی داده ها در فضای تعامد انجام شده [۷،۶،۴،۳]، از PCA و یا KPCA برای استخراج ویژگی استفاده شده است و بعد از تصویر داده ها در فضای ویژگی، با استفاده از الگوریتمهای خوشه بندی مانند K-میانگین داده هاخوشه بندی می شوند. در[۳] به بررسی کارایی روشهای مختلف استخراج ویژگی در خوشه بندی پرداخته شده است. در [۶]سعی شده است تا با راهنمایی PCA خوشه بندی انجام شود. در [۷] هدف رفع مشکل گرفتار شدن الگوریتم K-میانگین در بهینه های محلی با استفاده از KPCA می باشد.

در ادامه این مقاله، در بخش ۲ به معرفی Kernel PCA و بررسی مقالاتی که برای خوشه بندی از بردارهای ویژه استفاده نموده اند می پردازیم. روش پیشنهادی برای خوشه بندی بر اساس ترکیب بردارهای ویژه در بخش ۳ بررسی شده و در بخش ۴، با پیاده سازی این روش به روی چهار مجموعه داده سنتز شده، کارایی آن با روشهای دیگر خوشه بندی مانند فازی C-میانگین و K-میانگین مقایسه شده است و در بخش پایانی نتیجه گیری و زمینه های تحقیقاتی آینده بیان شده اند.

۲- بررسی مولفه های اصلی در حوزه کرنل و خوشه بندی در فضای ویژگی

۲-۱ بررسی Kernel PCA

ایده KPCA استفاده از PCA برای تصویر غیرخطی داده ها با استفاده از حقه کرنل می باشد. اگر مجموعه داده ای با m نمونه داشته باشیم به صورت $X = [x_1,...,x_m]$ ، ماتریس کواریانس داده ها را در فضای

---

[1] Principal Component Analysis(PCA)
[2] Principal axis

ورودی به صورت $C = \frac{1}{m} XX^T$ تعریف می نماییم. بردارهای ویژه نرمال C، زیرفضای مولفه ها را در PCA تشکیل می دهند که داده ها به طور خطی بروی این زیرفضا تصویر خواهند شد. برای توسعه این روش به حالت غیرخطی، فضای ویژگی H را که به صورت غیر خطی وابسته به فضای ورودی است در نظر می گیریم:

$$\varphi : \mathrm{X} \to \mathrm{H}, \quad \mathrm{x} \to \varphi(\mathrm{x}) \qquad (۱)$$

فضای H می تواند ابعاد بزرگ دلخواه و حتی نامحدود داشته باشد.

فرض می کنیم داده ها مرکزدار شده باشند، $\sum_{\mathrm{j}=1}^{m} \varphi(x_i) = 0$.

ماتریس کواریانس را در فضای H محاسبه می تماییم:

$$\mathrm{C} = 1/\mathrm{m} \sum_{\mathrm{j}=1}^{m} \varphi(x_i)\varphi(x_j)^T \qquad (۲)$$

سپس مقادیر ویژه $\lambda \geq 0$ و بردارهای ویژه غیر صفر $V \in H \setminus \{0\}$ که در معادله زیر صدق می کنند را می یابیم.

$$\lambda \mathrm{V} = \mathrm{CV} \qquad (۳)$$

همه جوابهای V با $\lambda \geq 0$ در فضای $\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_m)$ قرار می گیرد. بنابراین :

$$\lambda \langle \varphi(\mathrm{x_n}), V \rangle = \langle \varphi(\mathrm{x_n}), CV \rangle \quad \text{for all } \mathrm{n} = 1, \ldots, \mathrm{m} \qquad (۴)$$

علاوه بر این ضرایب $\alpha_i (i = 1, \ldots, m)$ نیز وجود دارند به صورتیکه

$$V = \sum_{\mathrm{i}=1}^{m} \alpha_i \varphi(x_i) \qquad (۵)$$

با مقایسه (۴) و (۵) خواهیم داشت:

$$\lambda \sum_{i=1}^{m} \alpha_i \langle \varphi(x_n), \varphi(x_j) \rangle = \frac{1}{m} \sum_{i=1}^{m} \alpha_i \left\langle \varphi(x_n), \sum_{j=1}^{m} \varphi(x_j) \langle \varphi(x_j), \varphi(x_i) \rangle \right\rangle \qquad (۶)$$

برای n=1,…,m.

با تعریف ماتریس کرنل $K_{ij} = \langle \varphi(x_j), \varphi(x_i) \rangle$ معادله (۶) به این صورت بیان می شود:

$$m\lambda K\alpha = K^2\alpha \qquad (۷)$$

و α بردارهای ستونی $\alpha_1, \ldots, \alpha_m$ را مشخص می نماید. برای حل مسئله (۷)، مسئله دوگان آن (۸) را برای مقادیر ویژه مخالف صفر حل می کنیم.

$$m\lambda\alpha = K\alpha \qquad (۸)$$

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ مقادیر ویژه K و $\alpha^1, \ldots, \alpha^m$ مجموعه کامل بردارهای ویژه مرتبط با آنها می باشند. فرض می کنیم λ_p آخرین مقدار ویژه مخالف صفر باشد. با در نظر گرفتن نرمال بودن بردارهای وابسته به λ_p در فضای H، $\alpha^1, \ldots, \alpha^p$ را نرمال می کنیم.

$$\langle V^n, V^n = 1 \rangle, n = 1, \ldots, p \qquad (۹)$$

بنابر رابطه (۵) و (۸)، رابطه فوق به صورت زیر تبدیل می شود:

$$1 = \sum_{i=1}^{m} \alpha_i^n \alpha_j^n \langle \varphi(x_i), \varphi(x_j) \rangle = \sum_{i,j=1}^{m} \alpha_i^n \alpha_j^n K_{ij} = \langle \alpha^n, K\alpha^n \rangle = \lambda \langle \alpha^n, \alpha^n \rangle \qquad (۱۰)$$

برای استخراج مولفه های اصلی، باید تصویر داده ها بروی بردارهای ویژه را در فضای H محاسبه نماییم. برای این منظور، اگر نقطه آزمون x را داشته باشیم تصویر آن در H برابر خواهد بود با $\varphi(x)$. بنابراین،

$$\langle V^n, \varphi(x) \rangle = \sum_{i=1}^{m} \alpha_i^n \langle \varphi(x_i), \varphi(x) \rangle \qquad (۱۱)$$

$$\langle V^n, \varphi(x) \rangle = \sum_{i=1}^{m} \alpha_i^n K(x_i, x) \qquad (۱۲)$$

n=1,…,p

تصویر نقطه آزمون روی بردار ویژه nام را $Prj(n)$ می نامیم.

برای سادگی، فرض شده بود که همه داده ها مرکزدار شده اند. در فضای ورودی محاسبه مرکز دار نمودن آسان است اما در فضای ویژگی H مشکل است و به سادگی نمی توان مرکز مشاهدات نگاشت شده به فضای با ابعاد بالا را محاسبه نمود. اما با استفاده از فرمول اصلاح شده زیر برای $KPCA$ این امر محقق می گردد.

$$\tilde{K}_{ij} = (K - 1_m K - K 1_m + 1_m K 1_m)_{ij} \qquad (۱۳)$$

و $1_m = 1/m$ برای همه i و j ها[۱].

کرنلهای مختلفی مانند کرنل گوسی، چند جمله ای، کوادراتیک، سیگموئید و ... می توانند به کار گرفته شوند.

۲-۲ بررسی روشهای خوشه بندی با استفاده از بردارهای ویژه

ابتدا به بررسی کارهایی می پردازیم که با انتقال فضای ورودی به فضای ویژگی خطی و یا غیر خطی، سعی در استخراج ویژگیهایی مناسب تر برای خوشه بندی دارند. در [۳]، ابتدا با استفاده از روشهای PCA، KPCA، Sammon و CCA داده ها به فضای ویژگی انتقال یافته اند و سپس عملکرد این روشها در خوشه بندی بروی داده ها آزمایش شده است. برای مقایسه کارایی خوشه بندی این روشهای تصویر داده، از الگوریتم K-میانگین استفاده شده است، نتایج حاصل نشان دهنده برتری KPCA برای استخراج ویژگیهای مناسب نسبت به سه روش دیگر است. در [۴] الگوریتم خوشه بندی معمولی را در فضای کرنل از دیدگاه جدیدی انجام داده است، به جای اینکه الگوریتمها در فضای ویژگی غیرخطی کرنلیزه شوند، در فضای ویژگی خطی این کار انجام می شود، به بیان دیگر در این مقاله نشان داده شده است که کرنل-K-میانگین[۳] معادل با اعمال KPCA قبل از K-میانگین بروی داده ها می باشد. به این مفهوم که اگر نمونه ها را روی KPCA تصویر کنیم و بعد K-میانگین را روی آنها اعمال نماییم، فاصله بین نمونه ها معادل با فاصله آنها در فضای غیر خطی هیلبرت خواهد بود. در [۶] ارتباط بین PCA و روش خوشه بندی K-میانگین بررسی شده است و با قرار دادن حد آستانه بروی بردار ویژگی استخراج شده با تحلیل مولفه های اصلی، خوشه بندی اولیه را انجام می دهد و سپس سعی بر بهینه نمودن تابع هدف K-میانگین می نماید. به بیان دیگر، روشی برای خوشه بندی با الگوریتم K-میانگین و با راهنمایی تحلیل مولفه های اصلی ارائه شده است. در [۷] الگوریتم مکاشفه ای خوشه بندی K-میانگین مبتنی بر KPCA و برنامه نویسی پویا ارائه و سعی شده است تا با استفاده از KPCA یک ناحیه بهینه از نمونه های ورودی در جهت اصلی کرنل[۴] به عنوان ناحیه اولیه برای K-میانگین انتخاب می شود تا از گرفتار شدن الگوریتم K-میانگین در بهینه های محلی اجتناب شود. در دسته دوم مقالاتی قرار می گیرند که با الهام از KPCA به روشهای دیگری برای استخراج ویژگی و خوشه بندی می پردازند مانند [۵] که تحلیل مولفه های انتروپی[۵] در آن به عنوان روش جدیدی برای تبدیل داده ها و کاهش بعد معرفی شده است. بر خلاف KPCA این روش نیازی به مقادیر ویژه بزرگتر ندارد و ممکن است مجموعه داده تبدیل یافته کاملا متفاوتی نسبت به KPCA تولید کند. این روش می تواند جایگزین مفیدی برای حذف نویز الگوها باشد. بعد از تصویر داده ها به فضای ویژگی با استفاده از این روش، داده های تبدیل یافته با توجه به توزیع خوشه ها در جهتهای متفاوت، یک ساختار زاویه ای خواهند داشت که از همین خاصیت برای خوشه بندی استفاده شده است..

۳- روش پیشنهادی خوشه بندی بر اساس ترکیب بردارهای ویژگی

در الگوریتمهای خوشه بندی که در فضای ورودی اجرا می شوند از ویژگیهایی که داده ها را توصیف می کنند برای هدف خوشه بندی استفاده می شود. در این مقاله سعی شده است تا به رویکردی جدید برای خوشه بندی برسیم، و بنابر این هدف به بررسی خوشه بندی در فضای تعامد یا ویژگیهای استخراج شده از KPCA پرداخته ایم. برای خوشه بندی داده ها، از بردارهای ویژه که محاسبه آنها در بخش قبل توضیح داده شد به عنوان روشی برای انجام این کار استفاده می نماییم بدین صورت که بعد از به دست آوردن این بردارهای ویژه در فضای KPCA، به دنبال یافتن بهترین های آنها برای خوشه بندی می گردیم.

مراحل انجام کار به صورت زیر می باشد:

۱- اولویت انتخاب بردارهای ویژه را بر مبنای مقادیر ویژه قرار می دهیم، یعنی بردارها را بر اساس ترتیب نزولی مقادیر ویژه مرتب می کنیم.

۲- تعداد مناسبی از بردارهای ویژه را انتخاب می نماییم(k). در انتخاب این تعداد عواملی مانند شکل مجموعه داده تاثیرگذار است.

۳- تصویر مجموعه داده ها را مطابق رابطه (۱۲) بروی بردارهای ویژه انتخاب شده محاسبه می کنیم و آنها را بردارهای نگاشت می نامیم(Prj).

۴- با یافتن حد آستانه مناسب، به طور جداگانه بر روی هریک از این بردارهای نگاشت عملیات خوشه بندی را انجام می دهیم.

۵- با توجه به میزان توانایی و موفقیت این بردارها در خوشه بندی، به هرکدام از آنها وزن w(i) ، i=1,…,k را اختصاص می دهیم.

۶- برای محاسبه خوشه بندی نهایی، مطابق رابطه (۱۴) برای رای گیری، میانگین وزندار خوشه بندی بر اساس بردارهای نگاشت را محاسبه می نماییم.

$$FinalVote = \frac{\sum_{i=1}^{k} w(i) \times \Pr j(i)}{\sum_{i=1}^{k} w(i)} \qquad (۱۴)$$

هر Prj(i) در رابطه فوق برداری است که نشان دهنده رای بردار ویژه iام برای خوشه بندی هر یک از داده های مجموعه داده می باشد. با انتخاب حد آستانه مناسب و اعمال آن بروی بردار FinalVote خوشه بندی نهایی را اعمال می کنیم. به عنوان مثال برای خوشه بندی داده ها به دو خوشه، حدآستانه را برابر با ۱.۵ در نظر می گیریم.

---

۳ KKMeans

۴ kernel principal direction

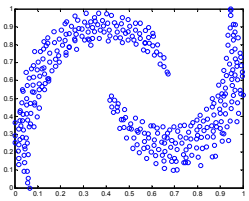
۵ Entropy Component Analysis

۴- نتایج آزمایشها

در این بخش به بررسی چند آزمایش می پردازیم. طبق الگوریتم ارائه شده در بخش ۳ بعد از مرتب نمودن نزولی بردارهای ویژه بر اساس مقادیر ویژه، تصویر مجموعه داده را بر روی بردارهای ویژه انتخاب شده محاسبه می نماییم و با قرار دادن حد آستانه مناسب روی بردارها، داده ها را خوشه می کنیم. سپس با انتساب دادن وزن مناسب به هر یک از این بردارها که نشان دهنده تعلق داده ها به خوشه ها می باشند، برای مشخص نمودن خوشه نهایی که داده به آن متعلق است از رای گیری استفاده می کنیم. برای انجام آزمایشها از کرنل گوسی استفاده نموده ایم که برای هرآزمایش مقدار σ مشخص شده است . برای ارزیابی روش پیشنهاد شده، میزان خطا (درصد نمونه هایی که نادرست خوشه بندی شده اند) در این روش را با الگوریتمهای خوشه بندی فازیC-میانگین و K-میانگین مقایسه نموده ایم و برای مقایسه بهتر، هر یک از الگوریتمهای فازی C-میانگین و K-میانگین را هم در فضای ورودی و هم در فضای حاصل از نگاشت داده ها بروی بردارهای ویژه KPCA اعمال می نماییم. در هر یک از آزمایشهای زیر، w_i ها نشان دهنده وزن منتسب به بردار ویژگی iام است. σ عرض کرنل گوسی به کار رفته را نشان می دهد و Thr حد آستانه مورد استفاده برای خوشه بندی روی بردارهای نگاشت می باشد.

۴-۱ آزمایش ۱

مجموعه داده دوبعدی نشان داده شده در شکل ۱ را در نظر بگیرید. پارامترهای در نظر گرفته شده برای این مجموعه داده در جدول ۱ نشان داده شده است.

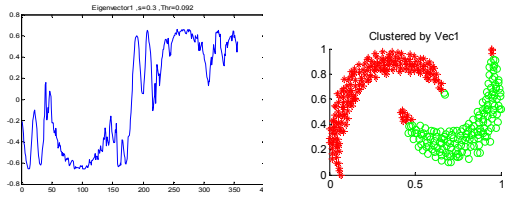


شکل۱- مجموعه داده آزمایش ۱

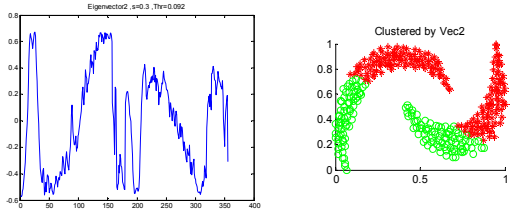
جدول۱ – پارامترهای مربوط به مجموعه داده شکل ۱

W_1	W_2	W_3	σ	Thr
۲	۱	۱	۰.۳	۰.۱

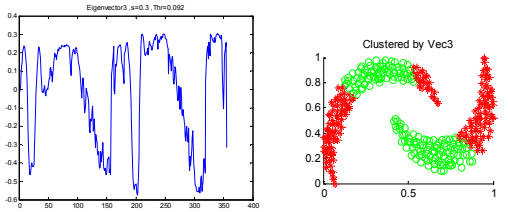
در شکل ۲ نگاشت مجموعه داده بر روی سه بردار ویژه اول و نتیجه حاصل از خوشه بندی داده ها بروی این بردارها نمایش داده شده است.



الف)تصویر داده ها بر روی بردار ویژه اول و نتیجه حاصل از خوشه بندی



ب)تصویر داده ها بر روی بردار ویژه دوم و نتیجه حاصل از خوشه بندی



پ)تصویر داده ها بر روی بردار ویژه سوم و نتیجه حاصل از خوشه بندی

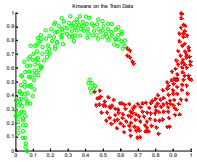
شکل۲- تصویر داده ها بروی بردارهای ویژه و نتیجه حاصل از خوشه بندی

همانگونه که در شکل۲ مشهود است توانایی بردار ویژه اول برای خوشه بندی داده ها نسبت به بردار دوم و سوم بیشتر می باشد، بنابراین با اختصاص دادن وزن ۲ به بردار اول و وزن۱ به دو بردار دوم و سوم(جدول۱)، طبق رابطه (۱۴) به محاسبه میانگین وزندار پرداخته و خوشه بندی نهایی را انجام می دهیم. مقایسه خطا در روش پیشنهادی و الگوریتمهای فازی C-میانگین و K-میانگین در جدول ۲ نشان داده شده است . نتایج نهایی خوشه بندی با این سه روش نیز در شکل ۳ مشاهده می شود.

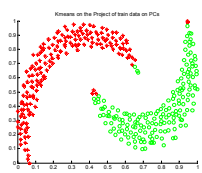
جدول ۲ - مقایسه میزان خطا

الگوریتم	درصد خطا در خوشه اول	درصد خطا در خوشه دوم	درصد خطای کلی
FCM (in input space)	۴.۲۷	۴.۵۴	۴.۴۰
FCM(in KPCA feature space)	۳.۰۴	۲.۲۷	۲.۶۵

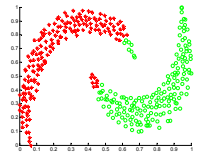
KMEANS(in input space)	۴.۸۷	۶.۰۶	۵.۴۶
KMEANS(in feature space)	۴.۸۷	۶.۰۶	۵.۴۶
KPCA-based Clustering	۱.۶۸	۱.۱۴	۱.۴۱



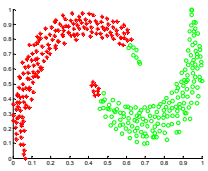
الف)خوشه بندی FCM در فضای ورودی



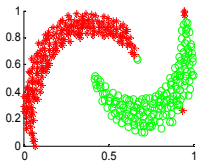
ب)خوشه بندی FCM در فضای ویژگی



پ)خوشه بندی KMEANS در فضای ورودی



ت)خوشه بندی KMEANS در فضای ویژگی



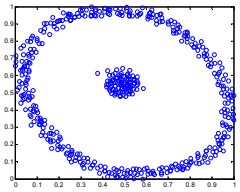
ث)خوشه بندی KPCA-based

شکل ۳ – مقایسه خوشه بندی با استفاده از روشهای FCM,KMEANS و روش پیشنهادی KPCA-based

همانگونه در جدول ۲ نشان داده شده است، خطا در روش پیشنهادی خوشه بندی مبتنی بر KPCA نسبت به روش فازی C-میانگین و K-میانگین کاهش یافته است.

۲-۴ آزمایش ۲

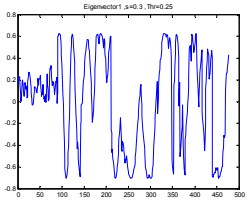
مجموعه داده شکل ۴ را در نظر بگیرید.

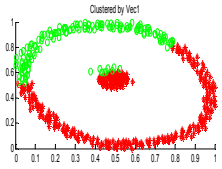


شکل۴- مجموعه داده آزمایش۲

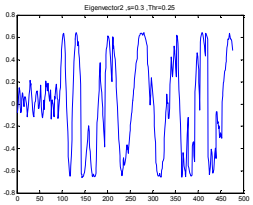
جدول۳ – پارامترهای مربوط به مجموعه داده شکل ۴

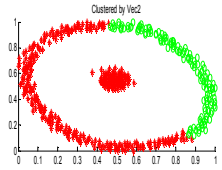
W_1	W_2	W_3	σ	Thr
۰	۰	۱	۰.۳	۰.۱



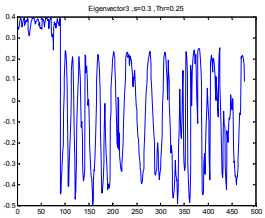


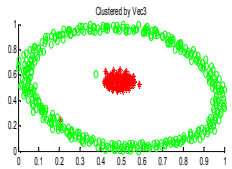
الف)تصویر داده ها بر روی بردار ویژه اول و نتیجه حاصل از خوشه بندی





ب)تصویر داده ها بر روی بردار ویژه دوم و نتیجه حاصل از خوشه بندی



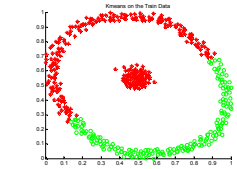


پ)تصویر داده ها بر روی بردار ویژه سوم و نتیجه حاصل از خوشه بندی

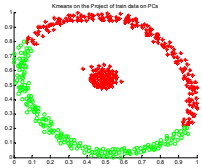
شکل۵- تصویر داده ها بروی بردارهای ویژه و نتیجه حاصل از خوشه بندی

همانگونه که در شکل۵-پ مشاهده می شود، بردار ویژگی سوم به خوبی می تواند داده ها را به دو خوشه کند، بنابراین مطابق جدول ۳ به بردار ویژگی سوم وزن ۱ و به دو بردار دیگر وزن ۰ را اختصاص می دهیم.

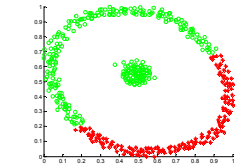
جدول ۴ - مقایسه میزان خطا			
الگوریتم	درصد خطا در خوشه اول	درصد خطا در خوشه دوم	درصد خطای کلی
FCM (in input space)	۰	۵۲.۵۹	۲۶.۲۹
FCM(in KPCA feature space)	۰	۴۷.۶۶	۲۳.۸۴
KMEANS(in input space)	۰	۵۵.۶۹	۲۷.۸۴
KMEANS(in feature space)	۰	۵۵.۶۹	۲۷.۸۴
KPCA-based Clustering	۱.۱۱	۰.۲۶	۰.۶۹



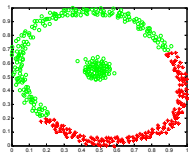
الف)خوشه بندی FCM در فضای ورودی



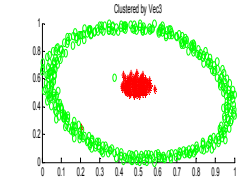
ب)خوشه بندی FCM در فضای ویژگی



پ)خوشه بندی KMEANS در فضای ورودی



ت)خوشه بندی KMEANS در فضای ویژگی



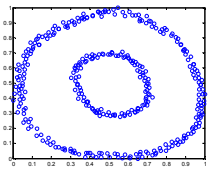
ث)خوشه بندی KPCA-based

شکل ۶ – مقایسه خوشه بندی با استفاده از روشهای FCM,KMEANS و روش پیشنهادی KPCA-based

با توجه به شکل۶ و جدول ۴ موفقیت آمیز بودن روش ارائه شده کاملا مشهود می باشد.

۳-۴ آزمایش ۳

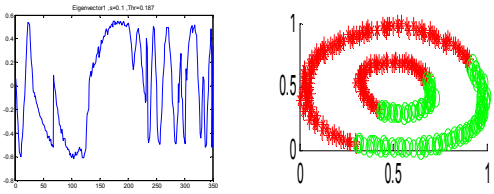
مجموعه داده شکل ۷ را در نظر بگیرید.



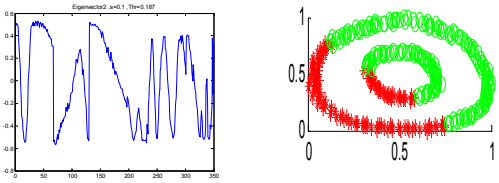
شکل۷- مجموعه داده آزمایش ۳

جدول۵ – پارامترهای مربوط به مجموعه داده شکل ۷

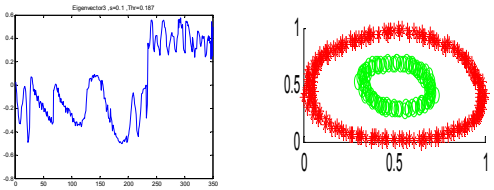
W_1	W_2	W_3	σ	Thr
۰	۰	۱	۰.۱	۰.۱۸۷



الف)تصویر داده ها بر روی بردار ویژه اول و نتیجه حاصل از خوشه بندی



ب)تصویر داده ها بر روی بردار ویژه دوم و نتیجه حاصل از خوشه بندی

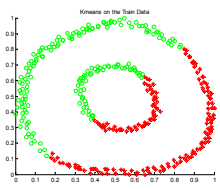


پ)تصویر داده ها بر روی بردار ویژه سوم و نتیجه حاصل از خوشه بندی

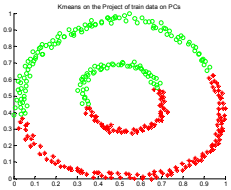
شکل ۸-تصویر داده ها بروی بردارهای ویژه و نتیجه حاصل از خوشه بندی

در این مثال نیز بردار ویژگی سوم به خوبی داده ها را خوشه بندی می کند.

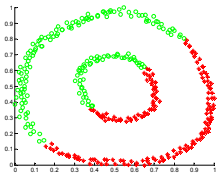
جدول ۶ – مقایسه میزان خطا			
الگوریتم	درصد خطا در خوشه اول	درصد خطا در خوشه دوم	درصد خطای کلی
FCM (in input space)	۴۸.۷۲	۵۰	۴۹.۳۶
FCM (in feature space)	۵۰	۷۲.۴۱	۶۱.۲۰
KMEANS(in input space)	۴۸.۲۹	۵۰	۴۹.۱۴۵
KMEANS(in feature space)	۴۸.۲۹	۵۰	۴۹.۱۴۵
KPCA-based Clustering	۰	۰	۰



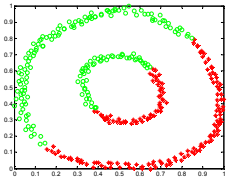
الف)خوشه بندی FCM در فضای ورودی



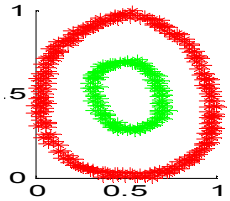
ب)خوشه بندی FCM در فضای ویژگی(با ۳ ویژگی اول)



پ)خوشه بندی KMEANS در فضای ورودی



ت)خوشه بندی KMEANS در فضای ویژگی



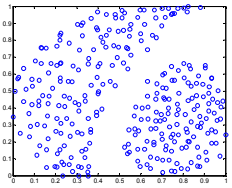
ث)خوشه بندی KPCA-based

شکل ۹ – مقایسه خوشه بندی با استفاده از روشهای FCM,KMEANS و روش پیشنهادی KPCA-based

در شکل ۹ مشاهده می شود که خوشه بندی مبتنی بر KPCA داده ها را به درستی خوشه نموده است.

۴-۴ آزمایش ۴

مجموعه داده شکل ۱۰ را درنظر بگیرید.

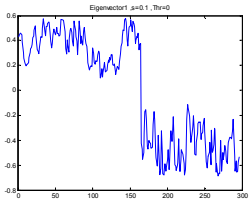


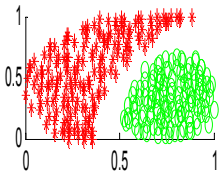
شکل۱۰- مجموعه داده آزمایش ۴

جدول۷ – پارامترهای مربوط به مجموعه داده شکل ۴

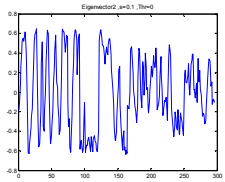
W_1	W_2	W_3	σ	Thr
۱	۰	۰	۰.۱	۰

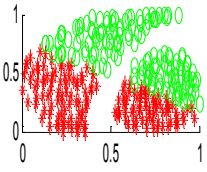
همانطور که در شکل ۱۱ مشاهده می شود، بردار ویژگی اول به خوبی می تواند داده ها را خوشه بندی نماید بنابراین وزن ۱ را به آن اختصاص می دهیم و دو بردار دیگر وزن ۰ می گیرند و تاثیری در خوشه بندی ندارند. در شکل ۱۲ نتایج نهایی حاصل از خوشه بندی نشان داده شده است.



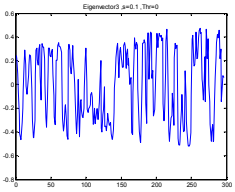


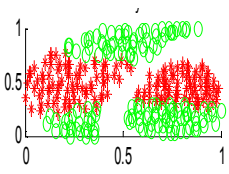
الف)تصویر داده ها بر روی بردار ویژه اول و نتیجه حاصل از خوشه بندی





ب)تصویر داده ها بر روی بردار ویژه دوم و نتیجه حاصل از خوشه بندی

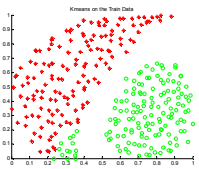




پ)تصویر داده ها بر روی بردار ویژه سوم و نتیجه حاصل از خوشه بندی

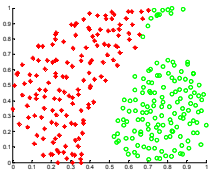
شکل ۱۱-تصویر داده ها بروی بردارهای ویژه و نتیجه حاصل از خوشه بندی

جدول ۸ - مقایسه میزان خطا			
الگوریتم	درصد خطا در خوشه اول	درصد خطا در خوشه دوم	درصد خطای کلی
FCM (in input space) S=0. 1	۷.۹۳	۰	۳.۹۶
FCM (in feature space)	۰	۰	۰
KMEANS(in input space)	۰	۷.۳۱	۳.۶۵
KMEANS(in feature space)	۰	۷.۳۱	۳.۶۵
KPCA-based Clustering	۰	۰	۰

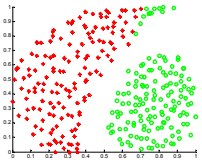


ب)خوشه بندی FCM در فضای ویژگی (با ۳ ویژگی اول)

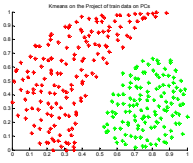
الف)خوشه بندی FCM در فضای ورودی



ت)خوشه بندی KMEANS در فضای ویژگی



پ)خوشه بندی KMEANS در فضای ورودی



ث)خوشه بندی KPCA-based

شکل ۱۲ – مقایسه خوشه بندی با استفاده از روشهای FCM,KMEANS و روش پیشنهادی KPCA-based

۵- نتیجه گیری

در این مقاله به بررسی روشی جدید برای خوشه بندی داده ها بر اساس ترکیب بردارهای ویژه حاصل از انتقال داده ها به فضای با ابعاد بالا توسط KPCA پرداخته شد. با توجه به مجموعه داده مورد بررسی، تعدادی مناسب از بردارهای ویژه را انتخاب نموده و با توجه به میزان توانایی هر بردار در خوشه بندی، وزنهای متناسب را به آنها نسبت می دهیم و آنها را باهم ترکیب می نماییم و عملیات خوشه بندی مبتنی بر رای گیری وزن دار انجام می دهیم.

در این مقاله وزنهایی که به هریک از بردارهای ویژگی اختصاص یافته است بر اساس آزمون و خطا و به صورت تجربی به دست آمده است، اما در آینده هدف تعیین این وزنها به صورت خودکار می باشد. همچنین تاثیر چینش متفاوت داده ها نیز نیازمند بررسی می باشد.

مراجع

[1] Bernhard Sch¨olkopf, Alexander J. Smola, *Learning with Kernels ,Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2002.

[2] Bernhard Sch¨olkopf , Alexander Smola, and Klaus Robert Muller, *Nonlinear Component Analysis as a Kernel Eigenvalue Problem*, Technical Report, Max-Planck-Institut –für biologische Kybernetik,1996.

[3] Alissar Nasser, Denis Hamad, *K-means Clustering Algorithm in Projected Spaces*,IEEE conference, 2006.

[4] Jing Li, Xuelong Li, Dacheng Tao, *KPCA for semantic object extraction in images*, Pattern Recognition, Vol. 41,pp. 3244 - 3250, 2008.

[5] Robert Jenssen, "Kernel Entropy Component Analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 Apr. 2009

[6] Chris Ding, Xiaofeng He, *K-means Clustering via Principal Component Analysis*, Proceedings of the 21 st International Conference on Machine Learning, Banff, Canada, 2004.

[7] Mantao Xu and Pasi Franti, *A Heuristic K-MEANS Clustering algorithm by Kernel PCA*, IEEE International Conference on Image Processing (ICIP),2004.