چهارمین سمینار آمار فضایی و کاربردهای آن
4th Seminar on Spatial Statistics and Its Applications
۵–۶ آبان ماه ۱۴۰۰
27-28 Oct. 2021

خوشه‌بندی داده‌های فضایی با ماشین‌های بردارپشتیبان

سمیرا زحمتکش، محسن محمدزاده
گروه آمار، دانشگاه تربیت مدرس

چکیده: امروزه یادگیری ماشین ابزارهای مهمی و جالب توجهی برای تحلیل هوشمندانه داده‌های فضایی، پردازش و به تصویر کشیدن آن‌ها فراهم می‌سازند. این تکنیک‌ها مکملی برای روش‌هایی مانند زمین‌آمار محسوب می‌شوند. یکی از این روش‌ها ماشین بردارپشتیبانی است که در تحقیقات زیادی نشان داده شده است در در بسیاری از مسائل کاربردی کارآمد است. در این مقاله به کاربردی از ماشین بردارپشتیبانی در خوشه‌بندی داده‌های فضایی پرداخته می‌شود، که در آن با استفاده از داده‌های اواقعی مدلی برای خوشه‌بندی نواحی طراحی می‌شود.

واژه‌های کلیدی: یادگیری ماشین، بردارپشتیبانی، داده‌های فضایی، خوشه بندی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62M30، 68Q32.

۱ مقدمه

یادگیری ماشین، به کارگیری مدل‌های آماری برای تشخیص الگوهای موجود در داده‌ها یا پیشگویی بر اساس الگوهای کشف شده در داده‌ها است. روش‌های یادگیری ماشین شامل دو ردهٔ نظارتی و غیرنظارتی است. در روش‌های نظارتی با استفاده از مجموعه داده‌های آموزشی[1] مدل های پیشگو ساخته می‌شود. در مقابل روش‌های غیرنظارتی بیشتر روی توصیف داده‌ها متمرکز می‌شوند. هر دو دسته از روش‌ها در زمینه‌های مختلف مانند فیزیک، بیولوژی، اکولوژی، جغرافیا و ... استفاده شده است. در عصر مواجهه با داده‌های بزرگ، تکنیک‌های یادگیری ماشین می‌توانند کاربرد موثری در تحلیل داده‌ها داشته باشند، زیرا نسبت به روش‌های متداول آماری مفروضات کمتری در مورد متغیرهای ورودی در نظر گرفته می‌شود. این روش‌ها برای اموری همچون پیشگویی رفتار مشتریان در آینده، خدمات مشاوره (پیشنهاد موسیقی، فیلم، خرید)، تشخیص چهره، رانندگی

[1] Training data set

* نام وآدرس پست الکترونیکی ارائه‌دهنده مقاله: سمیرا زحمتکش، samirazahmatkesh@yahoo.com

اتوماتیک، هوش مصنوعی و غیره کاربرد فراوانی دارند. زمین‌آمار یکی از راهکارهای متعارف برای تحلیل داده‌های وابسته فضایی و در صورت لزوم مدل‌بندی و پیش‌گویی آن‌ها است. به طور کلی زمین‌آمار رویکردی وابسته به مدل و بر اساس تحلیل اکتشافی و مدل‌بندی ساختارهای همبستگی فضایی است. در مقابل روش‌های یادگیری ماشین مبتنی بر داده هستند و تقریبا فارغ از مدل آماری هستند. در این روش‌ها با الگوریتمی مواجه هستیم که الگوها و روابط ناشناخته بین متغیرهای ورودی را کشف می‌کند. در این جا به معرفی روش ماشین بردار پشتیبانی[۲] (SVM) برای داده‌های فضایی پرداخته می‌شود. روش SVM یکی از روش‌های یادگیری نظارتی است که از آن برای خوشه‌بندی و رگرسیون استفاده می‌شود. این روش از جملهٔ روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای خوشه‌بندی داده‌ها نشان داده است. مبنای کار روش SVM، دسته‌بندی خطی داده‌ها است، که در آن سعی می‌شود خطی انتخاب شود که حاشیه اطمینان بیشتری داشته باشد. الگوریتم SVM جزو الگوریتم‌های تشخیص الگو دسته‌بندی می‌شود. از الگوریتم SVM در هر جایی که نیاز به تشخیص الگو یا دسته‌بندی اشیا در کلاس‌های خاص باشد می‌توان استفاده کرد. الگوریتم SVM اولین بار توسط ولادیمیر وپنیک در ۱۹۶۳ ابداع شد و در سال ۱۹۹۵ توسط وپنیک و کورتز برای حالت غیرخطی تعمیم داده شد (کورتز و وپنیک ، ۱۹۹۵؛ وپنیک ، ۱۹۹۵). روش SVM ابتدا با هدف خوشه‌بندی در دو کلاس مطرح شد و به تدریج به مسایل رگرسیونی (اسمولا و اسچوکوف ، ۱۹۹۸) و برآورد چگالی‌های احتمال (وستون و همکاران ، ۱۹۹۸) نیز تعمیم داده شد. الگوریتم SVM در مقایسه با سایر روش‌های یادگیری ماشین نسبتاً ساده است. بر خلاف شبکه‌های عصبی در ماکسیمم‌های موضعی گیر نمی‌افتد. برای داده‌های با ابعاد بالا تقریباً خوب جواب می‌دهد. مصالحه بین پیچیدگی دسته‌بندی‌کننده و میزان خطا به‌طور واضح کنترل می‌شود. کانوسکی و همکاران (۱۹۹۹) به ارایه کاربردی از الگوریتم SVM در داده‌های فضایی پرداختند و عملکرد آن در خوشه‌بندی را با روش‌های زمین‌آماری مانند کریکینگ مقایسه کردند. در این مقاله با هدف مطالعه کارایی الگوریتم SVM به نتایجی که کانوسکی و همکاران (۱۹۹۹) به دست آورده‌اند استناد خواهد شد. در ادامه روش ماشین‌های بردارپشتیبان خطی در بخش ۲ معرفی می‌شود. سپس روش ماشین‌های بردار پشتیبان غیر خطی در بخش ۳ ارائه می‌شود. آنگاه در بخش ۴ به ارائه کاربردی از SVM برای خوشه‌بندی داده‌های میزان رسوب آب در دریاچه جنوا که توسط کمیته بین‌المللی حفاظت آب جمع آوری شده است پرداخته می‌شود و عملکرد این الگوریتم با دو الگوی مختلف برای داده‌ها مورد بررسی و ارزیابی قرار می‌گیرد. در انتها به بحث و نتیجه‌گیری پرداخته خواهد شد.

۲ ماشین‌های بردارپشتیبان خطی

مجموعه‌ای از n زوج از داده های آزمایشی را به صورت

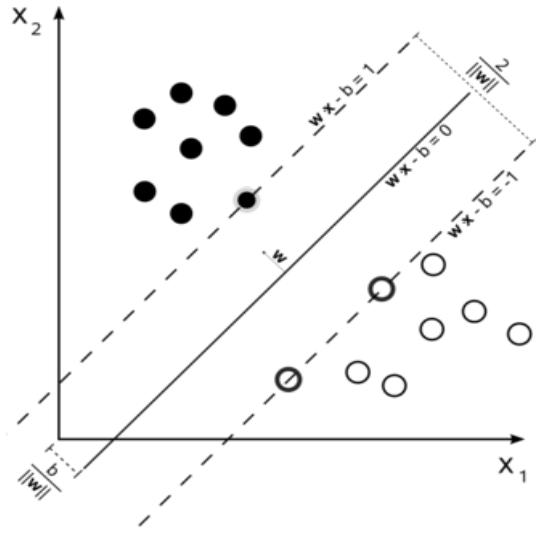
$$D = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in R^p, y_i \in \{-1, 1\}\}_{i=1}^{n}$$

در نظر بگیرید، که در آن هر $\mathbf{x}_i$ یک بردار حقیقی pـبعدی و y متغیری با مقادیر ۱ یا -1 است که قرار گیری نقاط داده در یکی از دو کلاس را نشان می‌دهد. در اینجا با فضای مختصات دوبعدی R^2 سر و کار خواهیم داشت. هدف پیدا کردن ابرصفحه [۳] جداکننده با بیشترین فاصله از نقاط حاشیه‌ای است که نقاط با $y_i = 1$ را از نقاط با $y_i = -1$ جدا کند. هر ابر صفحه می‌تواند به صورت مجموعه‌ای از نقاط $\mathbf{x}$ که برای آن شرط $\mathbf{w}.\mathbf{x} - b = 0$، برقرار باشد نوشته شود، که در آن $\cdot$ علامت ضرب داخلی و $\mathbf{w}$ یک بردار نرمال عمود بر ابرصفحه است. می‌خواهیم $\mathbf{w}$ و b را طوری انتخاب کنیم که بیشترین فاصله بین ابرصفحه‌های موازی که داده‌ها را از هم جدا می‌کنند ایجاد شود. این ابرصفحه‌ها با روابط

$$\mathbf{w}.\mathbf{x} - b = 1, \quad \mathbf{w}.\mathbf{x} - b = -1$$

[2] Support vector machines
[3] Hyperplanes

توصیف می‌شوند. اگر داده‌ها جدایی‌پذیر خطی باشند می‌توان دوابرصفحه در حاشیه نقاط به طوری که هیچ نقطه مشترکی نداشته باشند در نظر گرفت و سپس تلاش کرد تا فاصله آن‌ها حداکثر شود. به صورت هندسی ثابت می‌شود که فاصله بین این دو صفحه برابر $\frac{۲}{\|\mathbf{w}\|}$ است، که در آن $\|\mathbf{w}\|$ نرم اقلیدسی یا اندازه طول بردار $\mathbf{w}$ است. شکل ۱ اَبَرصفحه‌ای با حداکثر حاشیه برای یک ماشین بردارپشتیبانی برای داده‌هایی از دو دسته است. داده‌هایی که بر روی اَبَرصفحه حاشیه قرار دارند بردارهای پشتیبانی نام دارند. بنابراین هدف مینیمم کردن $\|\mathbf{w}\|$ است و برای جلوگیری از ورود نقاط به حاشیه (قرارگیری



شکل ۱ : ابرصفحه‌ای با حداکثر حاشیه برای یک ماشین بردارپشتیبانی که با داده‌هایی از دو دسته .

درست نقاط در هر کلاس) شرایط

$$\mathbf{w}.\mathbf{x_i} - b \geq ۱, \quad y_i = ۱$$
$$\mathbf{w}.\mathbf{x_i} - b \leq -۱, \quad y_i = -۱$$

به مساله اضافه می‌شود، که معادل

$$y_i(\mathbf{w}.\mathbf{x_i} - b) \geq ۱, \quad i = ۱, ..., n$$

است. به این ترتیب با کنار هم قرار دادن این دو، یک مساله بهینه‌سازی به صورت

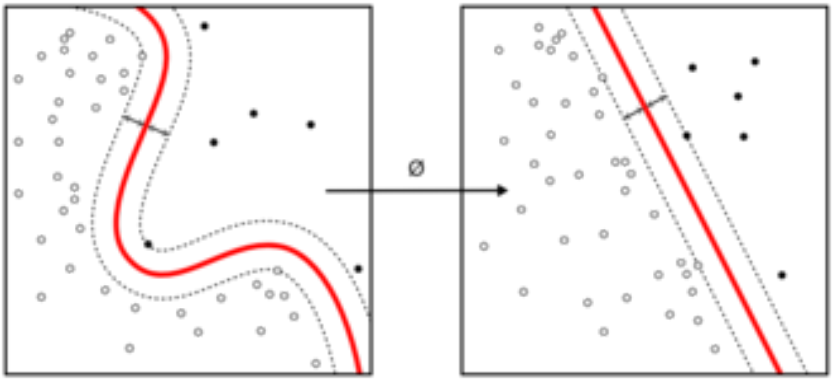
$$\min \|\mathbf{w}\|; \qquad \text{s.t.} \qquad y_i(\mathbf{w}.\mathbf{x_i} - b) \geq ۱, \qquad i = ۱, ..., n$$

به دست می‌آید. این یک مساله بهینه‌سازی دشوار است، زیرا به $\|\mathbf{w}\|$ وابسته است. حل این مساله با استفاده از روش‌های برنامه‌ریزی غیرخطی[۴] (برتسکاس و دیمیتری ، ۱۹۹۹) که روش‌های شناخته شده‌ای در حل مسایل محدودیت‌دار هستند صورت می‌گیرد. همچنین در ابعاد بالا به کمک قضیه دوگانی لاگرانژ (برگز و همکاران ، ۱۹۹۹) مساله مینیمم‌سازی به فضایی با ابعاد بالا تعمیم داده می‌شود.

[4]Quadratic programing

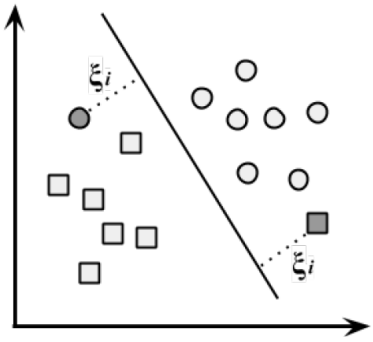
۳ ماشین‌های بردارپشتیبان غیر خطی

ابرصفحه جداکننده بهینه اولین بار توسط وپنیک (۱۹۶۳) ارائه شد که یک جداکننده خطی بود. بوزر و همکاران (۱۹۹۲) راهی را برای ایجاد خوشه‌بندی غیرخطی، با استفاده قرار دادن یک تابع هسته برای پیدا کردن ابرصفحه با بیشترین حاشیه، پیشنهاد دادند. الگوریتم پیشنهادی ظاهراً مشابه است، به جز آنکه تمام ضرب‌های نقطه‌ای با یک تابع هسته غیرخطی جایگزین شده‌اند. این خصوصیت اجازه می‌دهد، الگوریتم، برای ابرصفحه با بیشترین حاشیه در یک فضای ویژگی تغییرشکل داده، مناسب باشد. ممکن است، تغییرشکل غیرخطی باشد و فضای تغییر یافته، دارای ابعاد بالاتری باشد. به هر حال دسته‌بندی‌کننده، یک ابرصفحه در فضای ویژگی با ابعاد بالا است، که ممکن است در فضای ورودی نیز غیرخطی باشد (شکل ۲). در حالت ساده متغیر تضعیف شده ξ [5] که حاشیه‌ای بین دو دسته ایجاد می‌کند در نظر گرفته می‌شود. این متغیر



شکل ۲: استفاده از هسته برای خوشه‌بندی غیرخطی

اجازه می‌دهد که برخی نقاط در خوشه نادرست قرار گیرند. شکل ۳ این مساله را با جمله متناظر ξ_i نشان می‌دهد. مقدار



شکل ۳: نمایش متغیر تضعیف شده ξ در خوشه بندی غیرخطی SVM

ثابت C به عنوان هزینه برای تمام نقاطی که به اشتباه در خوشه‌ای قرار می‌گیرند در نظر گرفته می‌شود و به جای یافتن حداکثر حاشیه الگوریتم سعی می‌کند که هزینه کل را مینیمم کند. به این ترتیب مساله بهینه‌سازی به صورت

$$\min \|\mathbf{w}\| + C \sum \xi_i$$
$$y_i(\mathbf{w}.\mathbf{x}_i - b) \geq ۱ - \xi_i, ; \quad \forall \mathbf{x}_i, \quad \xi_i \geq ۰$$

[5] Slack variable

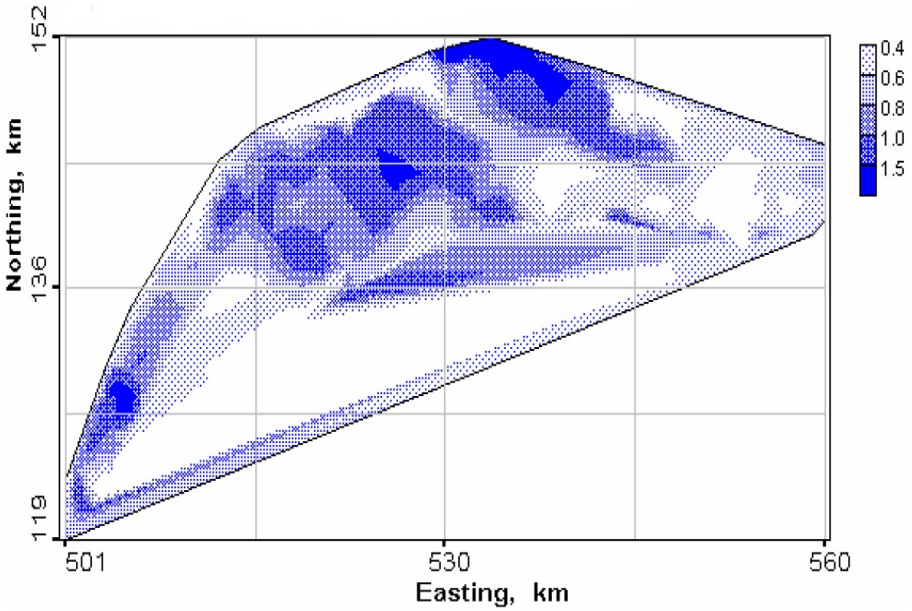
اصلاح می‌شود. عموماً توابع هسته $K : R^n \to R^n$ به صورت $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i).\phi(\mathbf{x}_j)$ در نظر گرفته می‌شوند، که در آن تابع ϕ داده‌ها را در فضای دیگری تصویر می‌کند. با پیروی از ایده بوزر و همکاران (۱۹۹۲) مساله بهینه‌سازی مجددا قابل بازنویسی به صورت

$$\min\ K(w,w) + C\sum \xi_i$$
$$y_i(K(w,x_i) - b) \geq ۱ - \xi_i, \quad \forall x_i, \quad \xi_i \geq ۰$$

است. این مساله نیز یک مساله غیرخطی دشوار است که با روش‌هایی که در بخش قبل ذکر شد حل خواهد شد.

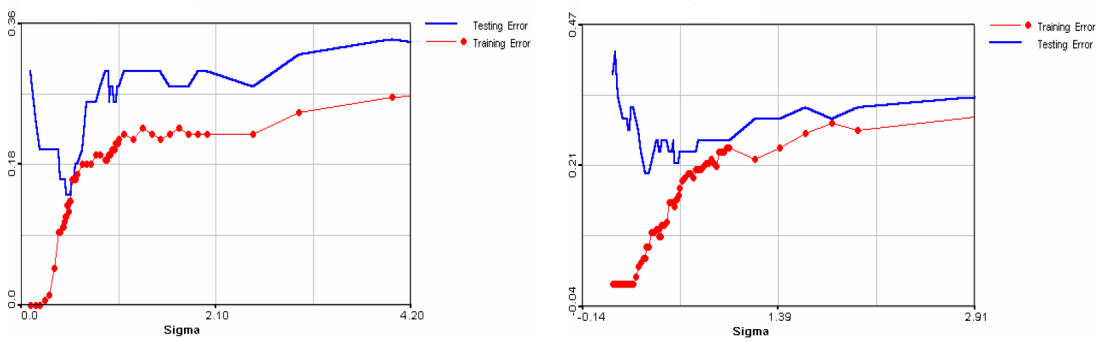
۴ تحلیل داده‌های رسوب آب دریاچه

داده‌های رسوب آب دریاچه جنوا توسط کمیته بین‌المللی حفاظت از آب این دریاچه طی سال‌های ۱۹۷۸، ۱۹۸۳ و ۱۹۸۸ اندازه‌گیری شده و از طریق سایت https://waterdata.usgs.gov/nwis/annual قابل دسترس است. در این‌جا داده‌های سال ۱۹۸۸، شامل اطلاعاتی درباره انواع مختلف رسوبات و اندازه‌گیری میزان برخی فلزات سنگین و مولکول‌های ارگانیک در آب دریاچه، که در ۲۰۰ موقعیت فضایی جمع‌آوری شده است، مورد تحلیل آماری قرار گرفته است (کانوسکی و همکاران ، ۱۹۹۹). تحلیل توصیف شبکه‌های نظارتی و خوشه‌بندی آن‌ها گام مهمی از تحلیل داده‌های فضایی است. غلظت کادمیم بر حسب $\frac{\mu g}{g}$ در تمام موقعیت‌های مکانی در شکل ۴ نشان داده شده است. مشاهده می‌شود که در ساحل



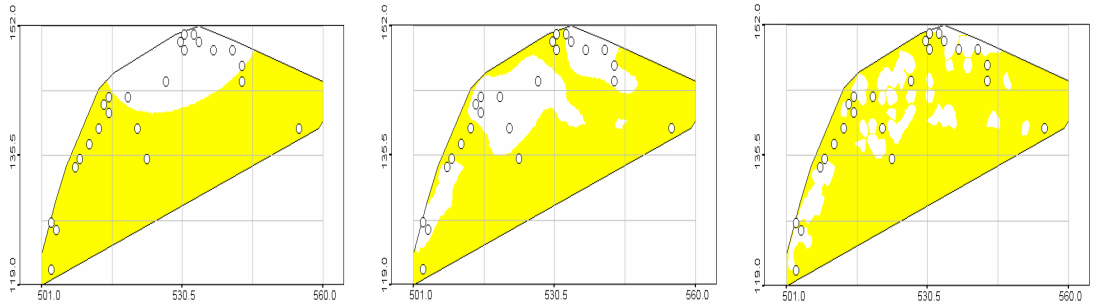
شکل ۴: غلظت کادمیم در موقعیت‌های مختلف مکانی

شمالی، قسمت میانی و در قسمت جنوب غربی دریاچه غلظت کادمیم بیشتر از سایر قسمت‌های دریاچه است. اکنون لازم است داده‌ها به دو بخش آموزشی و آزمایشی تقسیم‌بندی شود. انتخاب ۱۵۰ نقطه داده آموزشی و ۵۰ نقطه داده آزمایشی مورد هدف است. می‌توان این نقاط را به تصادف انتخاب کرد. اما به دلیل وابستگی فضایی داده‌ها بهتر است ابتدا ناحیه با مشبکه‌ای منظم پوشش داده شود و سپس از هر سلول نمونه‌ای برای داده‌های آموزشی و آزمایشی انتخاب شود. به این ترتیب مجموعه داده انتخاب شده همگن تر است. برای مشخص کردن دسته ها در داده‌ها علاقه‌مند هستیم که داده‌ها را به دو بخش بالا و پایین از یک حد آستانه تقسیم‌بندی نماییم. بنابراین تابع نشانگری تعریف می‌شود که مقدار آن صفر است اگر



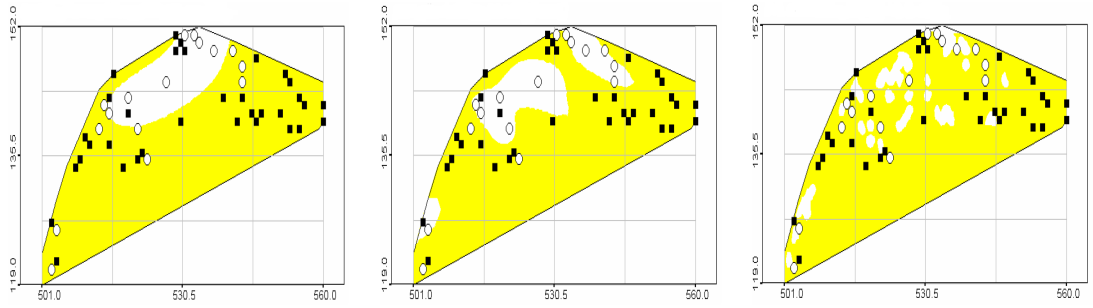
(ب) (الف)

شکل ۵: منحنی خطای SVM برای داده‌های آموزشی و آزمایشی و تابع نشانگر با آستانه الف- ۰/۸، ب- ۱.



(ج) (ب) (الف)

شکل ۶: خوشه‌بندی SVM با هسته RBF و حد آستانه ۰/۸، الف- بیش برازش، $\sigma = ۰/۰۳$، ب- بهینه، $\sigma = ۰/۳۵$ و ج- بیش همواری، $\sigma = ۳$.



(ج) (ب) (الف)

شکل ۷: خوشه‌بندی SVM با هسته RBF و حد آستانه ۱، الف- بیش برازش، $\sigma = ۰/۱$، ب- بهینه، $\sigma = ۰/۵$ و ج- بیش همواری، $\sigma = ۱۰$.

داده‌ها بالای یک حد آستانه باشد و در غیر اینصورت برابر ۱ است. این کار برای دو دسته داده‌های آموزشی و آزمایشی به طور جداگانه اجرا می‌شود. در اینجا دو مقدار $C_۱ = ۰/۸\frac{\mu g}{g}$ و $C_۲ = ۱\frac{\mu g}{g}$ انتخاب شده است. انتخاب مقدار ۰/۸ (که بسیار نزدیک به مقدار میانگین داده‌ها است) ناحیه‌ای همگن‌تر از مقادیر را نتیجه می‌دهد اما انتخاب مقدار ۱ باعث می‌شود

که اطلاعات بیشتر در موقعیت‌هایی قرار بگیرد که غلظت کادمیم بالاست. روش SVM برای دو الگوی مختلف ایجاد شده در داده‌ها انجام می‌شود تا کارایی الگوریتم ارزیابی شود. لازم به ذکر است در هر دو مورد با مساله خوشه‌بندی غیرخطی مواجه هستیم.

برای خوشه‌بندی لازم است ابتدا یک هسته مناسب انتخاب و پارامترهای آن مشخص شود. سپس ضرایب بردارهای پشتیبان بر اساس داده‌های آموزشی توسط بهینه‌گر محاسبه می‌شود. سرانجام کارایی ضرایب و پارامترهای هسته با استفاده از داده‌های آزمایشی برآورد می‌شود. اگر N تعداد کل داده‌ها باشد و m تعداد داده‌هایی که در دسته اشتباه قرار گرفته‌اند، نسبت $\frac{m}{N}$ به عنوان معیاری برای سنجش کیفیت نتیجه برای داده‌های آموزشی و آزمایشی استفاده می‌شود.

توابع هسته مختلفی مانند هسته چندجمله‌ای همگن و ناهمگن، هسته گاوسی،هسته تانژانت هذلولی، هسته تابع پایه شعاعی[۶] (RBF) و ... وجود دارند. در اینجا پس از بررسی این توابع هسته، در این مساله با هسته RBF در نظر گرفته شده است که نتایج بهتری نسبت به سایر هسته‌ها بدست می‌دهد. این هسته به صورت

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_i\|^۲/۲\sigma^۲\right)$$

است، که در آن σ پارامتر واریانس (پهنای باند) تابع پایه شعاعی است. برای خوشه‌بندی بهینه و مشاهده تغییرپذیری خطاهای آموزشی و آزمایشی در مقابل پارامتر واریانس هسته، منحنی‌های خطا محاسبه وا در شکل ۵ رسم شده‌اند. منحنی‌های خطا برای دو آستانه ۰/۸ و ۱ تقریبا مشابه است. و قابل تقسیم به سه بخش است. ابتدا در مقادیر پایین پارامتر هسته (واریانس) خطای آزمایشی در سطح بالایی قرار دارد در حالی‌که خطای آموزشی تقریبا صفر است در این قسمت بیش‌برازش رخ داده است. سپس خطای آزمایشی به همان سرعتی که خطای آموزشی در حال افزایش است، کاهش می‌یابد و بعد از این کاهش هر دو خطای آموزشی و آزمایشی شروع به افزایش می‌کنند. مقادیر بهینه پارامتر وقتی حاصل می‌شوند که خطای آزمایشی به کمترین مقدار خود برسد. در نهایت منحنی‌های خطا با حالت مسطح اما در سطح بالایی از خطا می‌رسند. در این قسمت هیچ تصمیمی نمی‌توان گرفت زیرا خوشه‌بندی به درستی انجام نخواهد شد. به این ناحیه بیش از حد هموار گفته می‌شود. نتایج خوشه بندی با روش SVM برای مقادیر مختلف پارمتر هسته که منجر به خوشه‌بندی بیش‌برازش، بهینه و بیش از حد هموار می‌شود برای مقدار آستانه ۰/۸ در شکل ۶ و برای آستانه ۱ در شکل ۷ نمایش داده شده است. در این شکل‌ها نقاط سفید مقادیر بالای حد آستانه و نقاط سیاه مقادیر پایین حد آستانه از داده‌های آزمایشی است. مناطق سفید خوشه‌بندی نیز متناظر با مقادیر بالای حد آستانه است.

بحث و نتیجه‌گیری

در این مقاله نتایج حاصل از خوشه‌بندی داده‌های فضایی دو کلاسه با استفاده از SVM قابل قبول بود و کیفیت اطلاعات مستخرج از داده‌ها را می‌توان با تغییر پارامترهای هسته و با استفاده از مجموعه داده‌های آزمایش کنترل کرد. مساله مهمی که برای تحقیقات آینده وجوددارد این است که پارامتر واریانس هسته مناسب به طور خودکار از داده‌ها قابل دستیابی باشد. در بحث خوشه‌بندی داده‌های فضایی با چندکلاس در داده‌ها مواجه خواهیم شد که نیاز به تعیین چندین حد آستانه خواهد بود. همچنین برای ارزیابی کیفیت عملکرد الگوریتم نحوه انتخاب داده‌های آموزشی و آزمایشی در داده‌های فضایی و اعتبارسنجی متقابل فضایی حایز اهمیت است که در مطالعات تکمیلی قرار داده شده است.

[6] Radial Basis Function

مراجع

Bertsekas, Dimitri P. (1999), Nonlinear Programming (Second eddition), Cambridge, MA. Athena Scientific.

Boser, B., Guyon, I., and V. N. Vapnik. (1992), A Training Algorithm for Optimal Margin Classifiers, 5th Annual ACM Workshop on COLT, 144–152.

Burges, C. J., Scholkopf, B., and Smola, A. J. (1999), Advances in Kernel Methods: Support Vector Learning, 89-116.

Cortes C., and Vapnik, V. (1995), Support Vector Networks, *Machine Learning*, **20**, 273–297.

Kanevski, M., Gilardi, N., Mayoraz, E., and Maignan, M. (1999), Environmental Spatial Data Classification with Support Vector Machines (No. REPWORK). IDIAP.

Lantz, B. (2019). Machine Learning with R: Expert Techniques for Predictive Modeling. Packt publishing ltd.

Smola A.J., and Schölkopf., B. (1998), A Tutorial on Support Vector Regression, *NeuroCOLT2 Technical Report Series*, NC2-TR-1998-030.

Vapnik, V. (1963) Pattern Recognition Using Generalized Portrait Method, *Automation and Remote Control*, 774-780.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag , New York.

Weston J., Gammerman, A., Stitson, M., Vapnik, V., Vovk, V., Watkins, C. (1998) Density Estimation using Support Vector Machines. Technical Report, Csd-TR-97-23. February.