دسته‌بند تک‌کلاسه مبتنی بر بردارهای پشتیبان برای داده‌های نویزی با استفاده از الگوریتم گروه میگوی آشوبی و تراکم محلی

محمدهادی قومنجانی[1]، دانشجوی کارشناسی ارشد؛ جواد حمیدزاده[2]، استادیار

۱- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی سجاد - مشهد - ایران - mghomanjani22@sadjad.ac.ir
۲- دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی سجاد - مشهد - ایران - j_hamidzadeh@sadjad.ac.ir

چکیده: هدف دسته‌بندی تک‌کلاسه، تشخیص و جداسازی داده‌های اصلی از داده‌های پرت است. دسته‌بند توصیف داده‌ها مبتنی بر بردار پشتیبان، یکی از روش‌های دسته‌بندی تک‌کلاسه است. این روش با تعریف اَبَرکُره‌ای در فضای ویژگی‌ها، سعی بر پوشش داده‌های اصلی در فضای اَبَرکُره دارد. سطح اَبَرکُره، مرز جداساز داده‌های اصلی از داده‌های پرت است. تعیین شعاع و مرکز مناسب برای اَبَرکُره در فضای ویژگی‌ها، یک مسئله بهینه‌سازی است. وجود داده‌های نویزی در مجموعه داده‌های اصلی و عدم توجه به تراکم داده‌ها در انتخاب مرکز، از چالش‌های روش فوق است که باعث ایجاد خطا در تعیین مرز اَبَرکُره خواهند شد. یکی از هدف‌های دسته‌بند پیشنهادی (KH-SVDD)، جستجوی مرکز مناسب برای اَبَرکُره، با استفاده از الگوریتم بهینه‌سازی گروه میگوی آشوبی است. همچنین با استفاده از تراکم محلی نقاط داده‌ها، اهمیت و تأثیر نقاط بر مرز دسته‌بند، به‌صورت یک وزن محاسبه می‌شود. این وزن، پارامتری کمکی برای تشخیص داده‌های اصلی و نویزی است. برای ارزیابی روش پیشنهادی، آزمایش‌های متعددی بر روی مجموعه داده‌های واقعی انجام شده است. نتایج حاصل از آزمایش‌ها نشان‌دهنده عملکرد بهتر روش پیشنهادی از نظر تشخیص داده‌های نویزی در مقایسه با الگوریتم‌های مرز دانش است.

واژه‌های کلیدی: دسته‌بند تک‌کلاسه، تشخیص داده‌های پرت و نویزی، گروه میگوی آشوبی، وزن‌دهی تراکمی.

One-class Classifier Based on Support Vectors for Noisy Data by Using Chaotic Krill Herd Algorithm and Local Density

M. H. Ghomanjani[1], MSc Student; J. Hamidzadeh[2], Assistant Professor

1- Faculty of Computer Engineering and Information Technology, Sadjad University of Technology, Mashhad, Iran, Email: mghomanjani22@sadjad.ac.ir
2- Faculty of Computer Engineering and Information Technology, Sadjad University of Technology, Mashhad, Iran, Email: j_hamidzadeh@sadjad.ac.ir

Abstract: The purpose of one-class classification is to detect and separate target data from outlier. Support vector data description classifier is one of the one-class data classification methods. This method creates a hyper-sphere in feature space and tries to cover target data in the hyper-sphere. The hyper-sphere surface is the discernment boundary between target and outlier data. Determining appropriate radius and center for the sphere is an optimization problem. Existence of the noise in the data set and lack of attention to data density for choosing the center is the challenge of this method that triggered the mistake in determining of detection boundary. In the proposed classifier (KH-SVDD) we tried to search appropriate center of sphere with the use of chaotic krill herd optimization algorithm. Also, a weight is calculated for the effectiveness of the points on the classifier boundary with the use of local density of data points. This weight is an auxiliary parameter to detect target data from noise. The results of the experiments have been compared with state-of-the-art methods, which show superiority of the proposed method in noise detection.

Keywords: One-class classifier, outlier and noise detection, chaotic krill herd, weighted density.

تاریخ ارسال مقاله: ۱۳۹۵/۰۹/۱۷
تاریخ اصلاح مقاله: ۱۳۹۵/۱۱/۱۷ و ۱۳۹۶/۰۲/۱۰
تاریخ پذیرش مقاله: ۱۳۹۶/۰۵/۰۹
نام نویسنده مسئول: جواد حمیدزاده
نشانی نویسنده مسئول: ایران - مشهد - بلوار جلال آل‌احمد ۶۴ - دانشگاه صنعتی سجاد - دانشکده مهندسی کامپیوتر و فناوری اطلاعات.

۱- مقدمه

تشخیص داده‌های پرت[۱] از مهم‌ترین فعالیت‌های داده‌کاوی محسوب می‌شود و در حوزه‌هایی مانند تشخیص بیماری، تشخیص نفوذ و اختلال سیستم، کاربردهای فراوانی دارد. یافتن داده‌های پرت در عمل کاری پرهزینه و در برخی موارد به‌علت در دسترس‌نبودن داده‌های پرت در مرحله آموزش، غیرممکن است. روش‌های متعددی از جمله مدل‌های آمار و احتمالی، مدل‌های خطی، مدل‌های تقریبی، مدل مارکوف، مدل K -نزدیک‌ترین همسایه و دسته‌بندی تک‌کلاسه برای تشخیص داده‌های پرت ارائه شده است. دسته‌بندی تک‌کلاسه[۲]، یکی از پرکاربردترین روش‌ها برای تشخیص داده‌های پرت است [۴-۱].

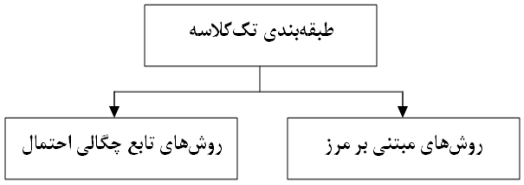
دسته‌بندی[۳]، انتساب یک نمونه داده جدید به یکی از کلاس‌های موجود است، این عملیات توسط روالی به نام دسته‌بند انجام می‌شود. به‌طورکلی در دسته‌بندی‌های تک‌کلاسه، مجموعه داده‌های در دسترس، داده‌های اصلی هستند و دانشی درباره داده‌های پرت وجود ندارد. در نتیجه داده‌های اصلی باید در یک کلاس (کلاس مثبت یا هدف) دسته‌بندی شوند. داده‌هایی که به کلاس تعیین‌شده، تعلق نداشته باشند به‌عنوان داده‌های پرت شناخته می‌شوند. از مهم‌ترین اهداف دسته‌بندی تک‌کلاسه، شناسایی و تشخیص مرز داده‌های اصلی از داده‌های پرت است [۵، ۶].

تفاوت دسته‌بندی تک‌کلاسه و دوکلاسه را می‌توان بدین‌صورت بیان کرد که در دسته‌بندی دوکلاسه (یا چندکلاسه)، ویژگی‌های دو کلاس اصلی (یا چند کلاس) در مرحله آموزش دسته‌بند، براساس مجموعه داده‌های برچسب‌دار تعیین می‌شود، سپس داده‌های دیگر با مقایسه با این دو کلاس (چند کلاس)، دسته‌بندی می‌شوند. در این نوع دسته‌بندی‌ها، همه داده‌ها باید در یکی از دو کلاس موجود، دسته‌بندی شوند، در نتیجه، در ماهیت دسته‌بندی دوکلاسه امکان تشخیص داده‌های پرت وجود ندارد [۷، ۸].

شکل ۱ تقسیم‌بندی کلی روش‌های دسته‌بندی تک‌کلاسه را نشان می‌دهد. اساس کار دسته‌بندها در روش‌های مبتنی بر مرز، به‌دست‌آوردن مرزی برای جداسازی داده‌های کلاس اصلی از داده‌های پرت است. گروه دوم، دسته‌بندهای مبتنی بر احتمال است. در این دسته‌بندها، با استفاده از تابع چگالی احتمال، یک مقدار عددی برای داده‌های موجود، محاسبه می‌شود. با مقایسه این مقدار عددی با یک مقدار آستانه، تعلق و عدم تعلق نمونه داده‌ها به کلاس اصلی مشخص می‌شود [۹].

دسته‌بند توصیف داده‌ها مبتنی بر بردارهای پشتیبان، یکی از دسته‌بندهای تک‌کلاسه است، که مهم‌ترین کاربرد آن تشخیص مرز بین داده‌های اصلی و داده‌های پرت است. در این مقاله یک دسته‌بند تک‌کلاسه (KH-SVDD[۴]) بر پایه دسته‌بند توصیف داده‌ها مبتنی بر بردارهای پشتیبان ارائه شده است. ایده اصلی این دسته‌بند، تشخیص درست داده‌های نویزی است.

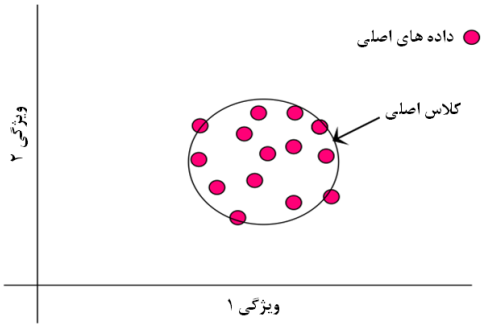
ساختار ادامه مقاله به‌شرح زیر است: در بخش دوم، مروری بر دسته‌بند توصیف داده‌ها مبتنی بر بردار پشتیبان، ارائه شده است. کارهای مرتبط پیشین در بخش سوم بررسی شده است. دسته‌بند پیشنهادی در بخش چهارم ارائه شده است. در بخش پنجم نتایج آزمایش‌ها نشان داده شده است. نتیجه‌گیری و کارهای آینده در بخش ششم مطرح گردیده است.



شکل ۱: تقسیم‌بندی روش‌های دسته‌بندی تک‌کلاسه

۲- دسته‌بند توصیف داده‌ها مبتنی بر بردار پشتیبان

دسته‌بند تک‌کلاسه توصیف داده‌ها مبتنی بر بردارهای پشتیبان (SVDD[۵])، از مهم‌ترین و پرکاربردترین روش‌ها جهت تشخیص داده‌های پرت است. این دسته‌بند جزء دسته دسته‌بندهای تک‌کلاسه مبتنی بر تشخیص مرز است. روش فوق به دنبال ساخت آبَرکُره‌ای برای پوشش داده‌های اصلی است. داده‌های درون آبَرکُره، به‌عنوان کلاس هدف و داده‌های خارج از آبَرکُره به‌عنوان داده‌های پرت، دسته‌بندی می‌شوند. مرز توصیف و تشخیص داده‌ها، همان رویه آبَرکُره است. این مرز با استفاده از نقاط داده‌ای که بر روی آن قرار دارند، توصیف می‌شود. به نقاط واقع بر روی مرز، بردارهای پشتیبان گفته می‌شود. شکل ۲ مثالی از آبَرکُره فوق، در فضای دوبعدی را نشان می‌دهد. داده‌های اصلی در شکل ۲ به‌صورت نقاط رنگی نشان داده شده‌اند. نقاط اصلی توسط یک دایره محاط شده‌اند. نقاطی که روی محیط دایره قرار دارند، بردارهای پشتیبان هستند و مرز دسته‌بند توسط این نقاط مشخص می‌شود [۱۰].

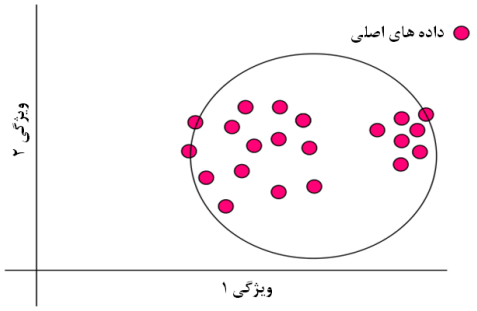


شکل ۲: آبَرکُره‌ای برای پوشش داده‌های اصلی

در این روش مرز آبَرکُره، با استفاده از لم هسته[۶] در فضای ویژگی‌های داده‌ها، قابل انعطاف می‌شود و امکان تفکیک‌پذیری خطی در فضاهای بالاتر (فضای ویژگی‌ها) افزایش می‌یابد. تفاوت مرز آبَرکُره،

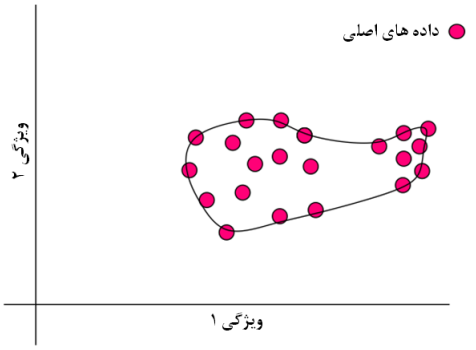
بدون و با استفاده از لـم هسـته در شـکل‌های ۳ و ۴ نشـان داده شـده است [۱۱].

شکل ۳ آبَرکُره‌ای محاط‌شده بر مجموعـه داده‌هـای اصـلی، بـدون استفاده لم هسته را نشان می‌دهد. در این حالت امکان تفکیک‌پذیری و انعطاف مرز آبَرکُره، در فضای ویژگی‌هـا وجـود نـدارد. شـکل ۴ همـان آبَرکُره، با استفاده لم هسته را نشان می‌دهـد. مـرز آبَرکُـره در شـکل ۴ منعطف شده و امکان تفکیک‌پذیری در فضـای ویژگی‌هـا را دارا اسـت [۱۲].



شکل ۳: مرز آبَرکُره بدون استفاده از لم هسته

روش SVDD به دنبال پوشش حداکثری داده‌های اصلی و کـاهش شانس پذیرش داده‌های پرت در کلاس هـدف اسـت. در SVDD، یـک مسئله بهینه‌سازی به‌صورت معادله (۱) تعریف می‌شود [۱۳].



شکل ۴: مرز آبَرکُره با استفاده از لم هسته

$$\min R_2 + C\sum_{i=1}^{N}\xi_i \qquad (۱)$$
$$s.t.\ \ \|\phi(x_i)-a\|^2 \le R^2+\xi_i \quad ,\xi_i \ge 0,\ i=1,2,...,N$$

هدف مسئله بهینه‌سازی فوق، تعریف آبَرکُره‌ای با حـداقل شـعاع R به مرکز a برای احاطه‌کردن داده‌های اصلی در فضای ویژگی‌هـا اسـت. داده‌هایی که خارج از آبَرکُره قرار گیرند، خطـایی بـه مقـدار ξ_i دارنـد. مقدار ξ_i در واقع فاصله نقاط بیرون از آبَرکُره تا مرز آبَرکُره است. نقـاط بیرون از آبَرکُره با ضریبی به مقـدار C، جریمـه می‌شـوند و در مسـئله بهینه‌سـازی اسـتفاده می‌شـوند تـا نقـاط مـرزی مناسـبی (بردارهـای پشتیبان) برای مرز آبَرکُره پیدا شود. تابع ϕ، یک تابع نگاشت است که یک نقطه را به ابعاد بالاتر در فضای ویژگی نگاشت می‌کند.

با توجه به معادله (۱)، هرگاه فاصله یک نقطه تا مرکز آبَرکُره، کمتر مساوی شعاع آبَرکُره باشد، آن نقطـه بـه کـلاس اصـلی تعلـق دارد. بـه عبارت دیگر، نقاطی که فاصله ویژگی‌های آنها تا مرکز آبَرکُره، بزرگ‌تـر از شعاع کره باشند، به‌عنوان نقاط پرت شناخته می‌شوند.

با حل مسئله بهینه‌سازی معادله (۱) با استفاده از ضرایب لاگرانـژ، معادله (۱) به‌فرم مشتق‌پذیر معادله (۲) تبدیل می‌شود [۱۳].

$$L = R^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[R^2+\xi_i-(\phi(x_i)-a)^2] - \sum_{i=1}^{N}\gamma_i\xi_i \quad ,\alpha_i,\gamma_i \ge 0 \qquad (۲)$$

در ایـن معادلـه α_i, γ_i ضـرایب لاگرانـژ هسـتند. هـدف معادلـه (۲)، کمینه‌سازی عبارت L بـا توجـه بـه کمینه‌سـازی متغیرهـای R,a,ξ و بیشینه‌سازی ضرایب لاگرانژ است. بـا مشـتق‌گیری جزئـی معادلـه (۲) نسبت به متغیرهای R,a,ξ، مقدار شعاع، موقعیت مرکز و مقدار خطای نقاط خارج از آبَرکُره حاصل می‌شود.

$$\frac{\partial L}{\partial R} = 0 \rightarrow \sum\alpha_i = 1 \qquad (۳)$$

معادله (۳) مشتق جزئی نسبت به متغیر شعاع است، در نتیجه مجموع ضرایب لاگرانژ در مسئله یک می‌شود.

$$\frac{\partial L}{\partial a} = 0 \rightarrow a = \frac{\sum\alpha_i\phi(x_i)}{\sum\alpha_i} = \sum\alpha_i\,\phi(x_i) \qquad (۴)$$

معادله (۴) مشتق جزئی نسبت به مرکز آبَرکُره است کـه بـا کمـک آن موقعیت مرکز آبَرکُره محاسبه می‌شود.

$$\frac{\partial L}{\partial \xi} = 0 \rightarrow \alpha_i + \gamma_i = C \rightarrow 0 \le \alpha_i \le C, \quad \gamma_i, C \ge 0 \qquad (۵)$$

با توجه به معادله (۵) محدودیتی برای ضریب لاگرانژ α تعیین می‌شود. بعـد از جـایگزینی معادلـه‌های فـوق در معادلـه (۲)، دوگـان مسـئله به‌صورت معادله (۶) ایجاد می‌شود.

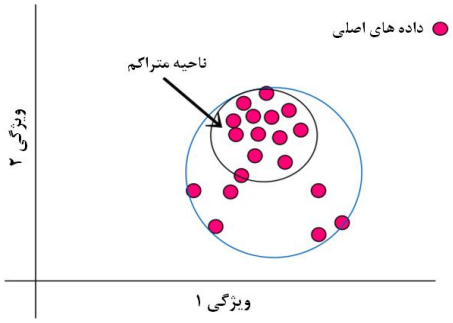
$$\max L = \sum_{i=1}^{N}\alpha_i(\phi(x_i).\phi(x_j)) - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha j(\phi(x_i).\phi(x_j)) \qquad (۶)$$
$$,\sum\alpha_i = 1, \quad 0 \le \alpha_i \le C$$

در معادله (۶)، $\phi(x_i).\phi(x_j)$ به‌صورت یک تابع هسته تعریف می‌شـود. معادله (۶) یک مسئله بهینه‌سازی محدب درجه دو است که یک جواب بهینه سراسری دارد. ارزیابی نقاط توسط این دسـته‌بند، بـا اسـتفاده از معادله (۷) صورت می‌گیرد. به‌ازای هر نقطه، اگر $f(x) \le 0$ شـود، آن نقطه به کلاس هدف تعلق دارد در غیر این صورت به‌عنوان نقطه پـرت معرفی می‌شود.

$$f(x) = \|\phi(x_i)-a\|^2 - R^2 \qquad (۷)$$

با توجه به معادله فوق نقاطی که ضریب لاگرانژ α_i آنها صفر باشد، به‌عنوان نقاط داخل آبَرکُره و نقاطی که ضریب لاگرانژ آنها بین صفر و C باشد، نقاط مرزی آبَرکُره هستند و به‌عنوان بردارهای پشتیبان شناخته می‌شوند. هم‌چنین اگر ضریب لاگرانژ برابر C باشد، نقاط خارج از آبَرکُره قرار دارند و به‌عنوان نقاط پرت شناخته می‌شوند.

در این مقاله یک دسته‌بند تک‌کلاسه برای تشخیص داده‌های نویزی (KH-SVDD) ارائه شده است. یکی از نقاط ضعف دسته‌بند SVDD عدم توجه به تراکم نقاط در مجموعه داده‌ها، برای انتخاب مرکز آبَرکُره است. یک ناحیه متراکم نزدیک به مرز آبَرکُره باعث خطا در انتخاب مرز، در نتیجه افزایش خطا در تشخیص داده‌های اصلی از پرت می‌شود [۲، ۳]. عدم توجه به تراکم در انتخاب مرکز آبَرکُره در شکل ۵ نشان داده شده است. در این شکل مرز آبَرکُره به ناحیه پرتراکم داده‌ها نزدیک است.



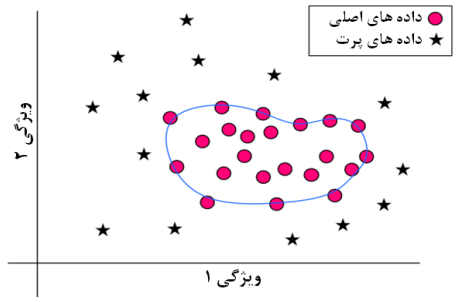
شکل ۵: ناحیه متراکم داده‌ها نزدیک به مرز آبَرکُره

در دسته‌بند پیشنهادی، از ایده الگوریتم غذایابی میگوی آشوبی برای یافتن مرکز بهینه آبَرکُره استفاده شده است. در طبیعت، میگوها در اطراف یک منبع غذایی یک ناحیه متراکم ایجاد می‌کنند. یک میگو با جستجو و انتخاب یک ناحیه متراکم میگوها، به سمت مرکز تراکم (محل غذا) حرکت می‌کند. از ایده فوق برای یافتن مرکز متراکم و مناسب برای آبَرکُره استفاده شده است.

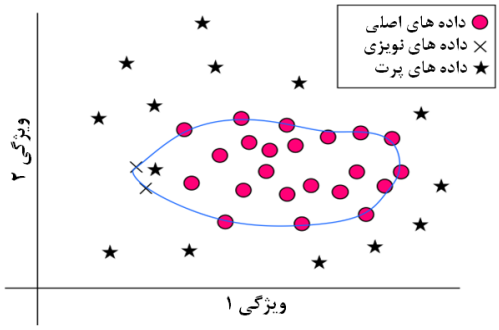
وجود نویز در مجموعه داده‌ها، یکی از چالش‌های مهم در دسته‌بندی است. عوامل مختلفی از جمله خطا در دستگاه‌های اندازه‌گیری ، نیروی انسانی و نویزهای محیطی مانند دما، فشار و گرما موجب ایجاد خطا و نویز در مجموعه داده‌های اصلی جمع‌آوری‌شده، می‌شوند [۱۴].

حضور داده‌های نویزی در مجموعه داده‌های اصلی، چالشی بسیار مهم برای دسته‌بندهای تک‌کلاسه است. رفتار دسته‌بند با داده‌های نویزی مانند داده‌های اصلی است. در نتیجه، با توجه به تأثیر داده‌های نویزی در مرحله آموزش دسته‌بند، مرز دسته‌بند دچار خطا می‌شود (شکل ۷). شکل ۶ مرز دسته‌بند بدون حضور داده‌های نویزی را نشان می‌دهد، اما در شکل ۷ داده‌های نویزی موجب خطا، در تشخیص مرز دسته‌بند شده‌اند. همان طور که در شکل ۷ مشخص است یک نقطه داده پرت داخل آبَرکُره کلاس اصلی است که نشان از خطای دسته‌بند است [۱۵، ۱۶].

در دسته‌بند پیشنهادی با آگاهی از این اصل که داده‌های نویزی از تراکم کمتری نسبت به داده‌های اصلی در مجموعه داده‌ها برخوردار هستند [۱۷]، از معیارهای تراکم محلی نقاط و فاصله تا مرکز آبَرکُره استفاده شده، تا علاوه بر تشخیص داده‌های پرت توسط دسته‌بند، نقاط نویزی در مجموعه داده‌های اصلی شناسایی شوند و از تأثیر آنها بر مرز دسته‌بند، جلوگیری شود.



شکل ۶: مرز آبَرکُره بدون وجود نویز در مجموعه داده‌های اصلی



شکل ۷: تأثیر نویز موجود در مجموعه داده‌ها بر مرز آبَرکُره

۳- مرور و بررسی کارهای مرتبط پیشین

تلاش‌های زیادی جهت بهبود روش توصیف داده‌ها به کمک بردارهای پشتیبان صورت گرفته است. در سال ۲۰۰۴ تَکس و دوین در روشی به نام NSVDD، با فرض این‌که داده‌های پرت نیز در دسترس می‌باشند، از آنها در فرآیند یادگیری استفاده کردند. هدف این روش ساخت آبَرکُره‌ای برای پوشش داده‌های اصلی و عدم پوشش داده‌های پرت است [۱۳].

در سال ۲۰۰۷ لی و همکاران [۱۸] یک روش جدید بر پایه SVDD معرفی کردند، در این الگوریتم رویکرد جدیدی برای معیار فاصله براساس نزدیک‌ترین همسایه و پنجره پارزن[۷] و درجه تراکم تمام نقاط داده‌ها، ارائه شد.

در سال ۲۰۰۹ وانگ و همکاران با تولید داده‌های پرت بیرون از مرز آبَرکُره، به‌صورت مصنوعی و دخالت آنها در مرحله آموزش دسته‌بند توصیف بردارهای پشتیبان، عملکرد آن را بهبود دادند [۱۹].

هم‌چنین در سال ۲۰۰۹ در [۲۰]، با ترکیب فیلترینگ متعامد و روش SVDD، سعی در بهبود عملکرد دسته‌بند در برخورد با داده‌های

نویزی شد. فیلترینگ متعامد به‌عنوان یک پیش‌پردازش در این روش است. این فیلتر انحراف معیار ناخواسته در داده‌ها را حذف می‌کند.

در روشی دیگر [۲۱] با یک روش پساپردازشی[۸]، سعی در بهبود عملکرد SVDD شد. در این روش سعی شده است بار محاسباتی لم هسته در منعطف‌سازی مرز اَبَرکُره حذف شود. در این الگوریتم برای اصلاح مرزهای اَبَرکُره از الگوریتم نزدیکترین همسایه استفاده شده است.

در تحقیق [۲۲]، روشی براساس SVDD، برای توصیف داده‌های دو کلاس معرفی شد. در این روش برای هر کلاس به‌طور جداگانه، اَبَرکُره‌ای در نظر گرفته می‌شود و داده‌های خارج از این دو اَبَرکُره به‌عنوان داده‌های پرت تلقی می‌شوند.

در سال ۲۰۱۳ لیو و همکاران یک روش جدید بر پایه SVDD معرفی کردند. در این روش برای نقاط داده‌های اصلی یک امتیاز اعتبار، تعریف می‌شود. این امتیاز، احتمال درست‌نمایی[۹] نقاط برای عضویت در کلاس هدف را مشخص می‌کند [۲۳].

چا و همکاران در سال ۲۰۱۴ روشی به نام DW-SVDD را براساس الگوریتم توصیف داده‌ها مبتنی بر بردارهای پشتیبان، ارائه کردند [۲۴]. در این روش نگاهی به توصیف نقاط داده‌ها براساس چگالی آنها شده است. رویکرد استفاده‌شده در این روش برای توزیع تراکم داده‌ها، استفاده از روش KNN است. با توجه به تراکم نقاط به آنها وزنی مشخص داده می‌شود تا اولویت تأثیر آنها بر مرز دسته‌بند مشخص شود. در این روش برای انتخاب مرکز اَبَرکُره از تراکم داده‌ها استفاده نمی‌شود.

در سال ۲۰۱۵ چن و همکاران روشی به نام R-SVDD را بر پایه الگوریتم توصیف داده‌ها مبتنی بر بردار پشتیبان، برای دسته‌بندی در مجموعه داده‌های نویزی (هنگامی که فقط مجموعه داده‌های اصلی در دسترس است) معرفی کردند [۲۵]. در این روش از معیار فاصله‌ای به نام Cut-off استفاده شده است و با کمک همین معیار فاصله، تراکم محلی نقاط محاسبه شده و سپس از آن در تابعی برای جریمه نقاط داده‌ها استفاده می‌شود. از اشکالات این روش عدم توجه به تراکم داده‌ها در انتخاب مرکز اَبَرکُره است. هم‌چنین برای تشخیص داده‌های نویزی از داده‌های اصلی از تابع معادله (۸) برای جریمه نقاط استفاده می‌شود. در این تابع توجهی به فاصله نقاط تا مرکز اَبَرکُره نشده است و فقط از تراکم محلی نقاط ($\rho(x_i)$) استفاده شده است.

$$w(x_i) = \rho(x_i) / \max(\rho(x_i)) \qquad (۸)$$

کیم و همکاران در سال ۲۰۱۵ روشی جدیدی با استفاده از یادگیری عمیق برای توصیف داده‌ها معرفی کردند [۲۶]. در این روش برای غلبه بر مشکل بیش‌برازش[۱۰] از یک مدل چندلایه شبکه عصبی با یادگیری عمیق استفاده شده است. در این مدل لایه‌های پنهان وظیفه استخراج ویژگی‌های لازم برای SVDD را بر عهده دارند. هم‌چنین هر لایه به‌عنوان دسته‌بند یک کلاس، برای مجموعه داده‌ها است.

۴- روش پیشنهادی

دسته‌بند پیشنهادی (KH-SVDD) از دو گام اصلی تشکیل شده است. در گام اول، هدف پیداکردن یک مرکز بهینه برای اَبَرکُره با توجه به تراکم نقاط است و هدف گام دوم، ساخت مرز اَبَرکُره با کمترین شعاع ممکن و بیشترین دقت است. هم‌چنین مرحله آموزش دسته‌بند پیشنهادی در این دو گام، با فرض در دسترس‌بودن داده‌های اصلی انجام می‌شود. روش پیشنهادی در دو بخش، براساس گام‌های اصلی آن بررسی می‌شود. بخش ۴-۱ به بررسی، چگونگی جستجوی مرکز برای ساخت اَبَرکُره می‌پردازد. در بخش ۴-۲ نحوه تعیین مرز اَبَرکُره مشخص می‌شود.

۴-۱- جستجوی مرکز اَبَرکُره

هدف از گام اول، جستجوی مرکزی مناسب برای اَبَرکُره است. مرکز یک ناحیه متراکم، می‌تواند به‌عنوان یک جواب بهینه برای مرکز اَبَرکُره باشد. جستجوی نقطه‌ای متراکم، در مسئله با ابعاد بالا (تعداد ویژگی‌های زیاد) و تعداد نقاط زیاد، یک کار زمان‌بر است. در نتیجه استفاده از روش‌هایی که کل مجموعه داده‌ها را برای شمارش تراکم نقاط، پیمایش کنند یک کار زمان‌بر است. در روش پیشنهادی از الگوریتم گروه میگوی آشوبی برای یافتن نقاط متراکم محلی استفاده شده است.

در این مقاله از الگوی رفتار میگوها (در جستجوی غذا و پیوستن به یک دسته میگو) در جهت پیداکردن یک مرکز اَبَرکُره بهینه، برای توصیف داده‌ها الهام گرفته شده است [۲۷]. مسئله دسته‌بندی تک‌کلاسه مانند یک مسئله غذایابی گروه میگوها در نظر گرفته شده است.

یکی از مشکلات الگوریتم SVDD، عدم توجه به تراکم نقاط در انتخاب مرکز اَبَرکُره است. در دسته‌بند پیشنهادی (KH-SVDD) تلاش شده است تا مرکز اَبَرکُره به سمت ناحیه پرتراکم حرکت کند. این ایده موجب می‌شود مرز اَبَرکُره از نواحی متراکم دور شود که موجب کاهش خطا در تشخیص نقاط اصلی از نقاط پرت و نویزی می‌شود. در دسته‌بندی‌های تک‌کلاسه که فقط داده‌های یک کلاس موجود هستند، تراکم نقاط داده‌های اصلی، بسیار بیشتر از تراکم نقاط نویزی است.

هر نقطه داده (مانند یک میگو که به یک دسته میگو برای غذایابی ملحق می‌شود) در جستجوی یک مرکز اَبَرکُره، برای الحاق به آن است. رفتار غذایابی یک میگو (که در جستجوی غذا در ناحیه پرتراکم است و به سمت مرکز ناحیه متراکم حرکت می‌کند) مانند رفتار مرکز اَبَرکُره (مرکز اولیه با استفاده از الگوریتم SVDD مقداردهی می‌شود) است که تمایل دارد نقش خود را با یک نقطه متراکم‌تر عوض کند.

در الگوریتم گروه میگوی آشوبی، تابع ارزیابی تغییرات حرکت میگو، براساس فاصله میگو تا غذا (F)، تراکم میگوها (D) و حرکت ایجادشده توسط سایر میگوها (N) به‌صورت معادله (۹) در فضای n-بعدی تعریف می‌شود.

$$\frac{dx_i}{dt} = N_i + F_i + D_i \qquad (۹)$$

در دسته‌بند پیشنهادی، تغییرات زمان (dt) وجود ندارد و مانند سایر الگوریتم‌های تکاملی، جایگزین تغییرات زمان، تغییر در تکرار اجراهای الگوریتم است. در نتیجه تغییرات زمان (اختلاف تکرارهای الگوریتم) یک واحد است ($dt = 1$).

ارزیابی تغییرات حرکت نقاط مرکز، در روش پیشنهادی توسط معادله (۱۰) بر پایه معادله (۹) محاسبه می‌شود، که در آن فاصله یک نقطه داده از مرکز قبلی (F)، تراکم محلی یک نقطه داده (D) و حرکت ایجادشده سایر نقاط (N) است. حرکت سایر نقاط در واقع جابه‌جایی نقش مرکز آبَرکُره، توسط نقاط است، که به سمت مرکز ناحیه متراکم جابه‌جا می‌شوند.

$$dx_i = N_i + F_i + D_i \qquad (۱۰)$$

با توجه به این‌که در روش پیشنهادی نقاط با تراکم بالا برای مسئله دارای اهمیت می‌باشند، تراکم محلی نقاط داده، با استفاده از نامساوی (۱۱) محاسبه می‌شود. هر دو نقطه از داده‌ها که در نامساوی (۱۱) صدق کنند به‌عنوان همسایه یکدیگر شناخته می‌شوند. به تعداد نقاطی که در همسایگی یک نقطه باشند، تراکم آن نقطه گفته می‌شود.

$$\frac{1}{d}\sum_{f=1}^{d}\left\|x_{if} - x_{jf}\right\| \le \eta \qquad (۱۱)$$

در نامساوی (۱۱)، d تعداد ویژگی‌های داده‌ها و x_{if} ویژگی f–ام از نقطه i–ام است. در معادله فوق ابتدا میانگین فاصله بین تمام ویژگی‌های دو نقطه از هم محاسبه می‌شود، اگر مقدار محاسبه‌شده از یک مقدار آستانه (η) کمتر یا مساوی باشد، دو نقطه در همسایگی یکدیگر قرار دارند. هم‌چنین از معادله سمت چپ نامساوی (۱۱) برای محاسبه فاصله یک نقطه تا مرکز آبَرکُره نیز استفاده می‌شود.

در الگوریتم گروه میگو، حرکت میگوها به سمت مرکز تراکم (محل غذا) است. معادله این حرکت با استفاده از تراکم محلی نقطه فعلی و تراکم محلی نقطه هدف، با کمک معادله (۱۲) تقریب زده می‌شود.

$$N_i^{new} = \alpha_i . N^{max} . \omega . N^{old} \qquad (۱۲)$$

در معادله (۱۲)، ω یک وزن تصادفی در بازه صفر و یک است. N^{max} بیشترین سرعت و N^{old} آخرین حرکت انجام شده است. جهت حرکت (α_i)، به کمک معادله (۱۳) محاسبه می‌شود.

$$\alpha_i = \alpha_i^{local} + \alpha_i^{target} \qquad (۱۳)$$

در روش پیشنهادی صورت معادله‌های (۱۲) و (۱۳) تغییر نمی‌کند و پارامترهای آن به‌این‌صورت می‌باشند: N^{max} بیشترین فاصله بین دو مرکز آبَرکُره قبلی و جدید، در تمام تکرارهای الگوریتم است. N^{old} فاصله بین مرکز قبلی و جدید، در تکرار اخیر الگوریتم است. هم‌چنین پارامتر α_i معادله (۱۳)، برای روش پیشنهادی، جهت حرکت نقاط مرکز (جهت تغییر نقش نقاط برای مرکزشدن در فضای ویژگی‌های داده‌ها) را نشان می‌دهد.

$$x_i(t + \Delta t) = x_i(t) + \Delta t . \frac{dx_i}{dt} \qquad (۱۴)$$

معادله حرکت یک میگو در طول زمان با کمک معادله (۱۴) بیان می‌شود، در روش پیشنهادی معادله جابه‌جایی مرکز آبَرکُره در تکرارهای مختلف (براساس معادله (۱۴))، با استفاده از معادله (۱۵) نشان داده می‌شود، که در آن t تعداد دفعات اجرای الگوریتم، را نشان می‌دهد.

$$x_i(t+1) = x_i(t) + dx_i \qquad (۱۵)$$

به‌طورکلی در گام اول روش پیشنهادی، یک بردار ویژگی‌ها، برای مرکز اولیه آبَرکُره (مانند یک کروموزوم) مقداردهی می‌شود. این بردار توسط تابع برازش (۱۰) مورد ارزیابی قرار می‌گیرد. در یک گام اجرای الگوریتم تراکم یک نقطه (مرکز فعلی) با نقاط همسایه‌اش مقایسه می‌شود و نقطه‌ای که بیشترین مقدار تابع برازش را به‌دست آورد به‌عنوان نسل برگزیده، مرکز آبَرکُره را تغییر می‌دهد. هم‌چنین در روش پیشنهادی از عملگرهای تولیدمثل، بازترکیب و جهش معرفی‌شده در الگوریتم گروه میگوی آشوبی [۲۲]، استفاده می‌شود.

یکی از مشکلات روش‌های جستجو، گرفتارشدن در نقاط متراکم محلی است. در روش پیشنهادی برای جلوگیری انتخاب نقطه متراکم محلی به‌جای سراسری، از توابع نگاشت آشوب برای تولید تصادفی گام جستجو، استفاده می‌شود. توابع آشوب این امکان را فراهم می‌سازند تا با در اختیارداشتن تابع نگاشت آشوب و مقدار اولیه تابع، اعداد تصادفی دوباره تولید شوند. این قابلیت برای جلوگیری از تولید نقاط تصادفی تکراری در الگوریتم که باعث انتخاب تکراری یک نقطه متراکم شود، بسیار مفید است.

از تابع آشوب معادله (۱۶) (Logistic map[20]) برای تغییر موقعیت تصادفی مرکز آبَرکُره، استفاده می‌شود. (r =۳/۹۹۹۹)

$$X_{n+1} = r . X_n . (1 - X_n) \qquad (۱۶)$$

۴-۲- تعیین مرز آبَرکُره (تعیین شعاع)

هدف این گام، یافتن یک شعاع کمینه برای تشکیل مرز آبَرکُره است. در ابتدا یک مقدار جریمه برای نقاط، علاوه بر میزان خطای ξ_i، تعریف می‌شود. این مقدار براساس تراکم محلی نقاط و فاصله تا مرکز آبَرکُره با توجه به معادله (۱۷) محاسبه می‌شود. هر چقدر میزان این جریمه کمتر باشد، نقطه مورد نظر تاثیری کمتری بر مرز دسته‌بند در مرحله آموزش دارد.

$$\omega(xi) = \frac{D_i}{\max(D_i)} . \frac{1}{dist(x_i)} \qquad (۱۷)$$

در معادله (۱۷)، D_i تراکم محلی یک نقطه و $dist(x_i)$ فاصله نقطه مورد نظر تا مرکز آبَرکُره است. مقدار جریمه محاسبه‌شده، عددی بین صفر و یک است. با استفاده از جریمه محاسبه‌شده معادله (۱) را به‌صورت معادله (۱۸) تغییر می‌دهیم.

$$\min R_2 + C\sum_{i=1}^{N}\omega(x_i)\xi_i$$
$$s.t.\ \ \|\phi(x_i)-a\|^2 \le R^2+\xi_i \quad ,\xi_i \ge 0,\ i=1,2,...,N \tag{۱۸}$$

با اضافه کردن ضرایب لاگرانژ به معادله (۱۸)، معادله بهینه‌سازی (۱۹) حاصل می‌شود.

$$L = R^2 + C\sum_{i=1}^{N}\omega(x_i)\xi_i - \sum_{i=1}^{N}\alpha_i[R^2+\xi_i-(\phi(x_i)-a)^2] - \sum_{i=1}^{N}\gamma_i\xi_i \quad ,\alpha_i,\gamma_i \ge 0 \tag{۱۹}$$

با مشتق‌گیری جزئی از متغیرهای شعاع، مرکز و خطا معادله (۱۹)، دوگان مسئله به‌صورت معادله (۲۰) تبدیل می‌شود.

$$\max\ L = \sum_{i=1}^{N}\alpha_i(\phi(x_i).\phi(x_j)) - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha\, j(\phi(x_i).\phi(x_j)) \quad ,\sum\alpha_i = 1,\quad 0\le\alpha_i\le\omega(x_i)C \tag{۲۰}$$

شعاع اَبَرکُره با توجه به بردارهای پشتیبان (SVs) با کمک معادله (۲۱) محاسبه می‌شود.

$$R = \sqrt{\frac{1}{|SVs|}\sum\|\phi(x_i)-a\|^2} \tag{۲۱}$$

گام‌های دسته‌بند KH-SVDD در الگوریتم ۱ نشان داده شده است.

الگوریتم ۱: الگوریتم دسته‌بند KH-SVDD

1 **At the first epoch:** calculate center of the hyper-sphere (with SVDD Eq. (4))
First Step:
2 **while** (not Stop-Condition)
3 **for** each data point is neighbor with center of the hyper-sphere
4 Calculate density of data points by Eq. (11)
5 Select maximum local density as new center
6 Evaluate fitness function for center by Eq. (10)
7 **end for**
8 use Eq. (16) for generate random center of hyper-sphere for find dense point
9 **end while**
Second Step:
10 **for** each data point calculate Eq. (17)
11 Determine SVs by Eq. (7)
12 Calculate R by Eq. (21)
13 **end for**

۵- ارزیابی و مقایسه روش پیشنهادی

در این بخش نتایج حاصل از آزمایش‌ها بر روی دسته‌بند KH-SVDD با روش‌های SVDD[13], DW-SVDD[24], R-SVDD[25] آورده شده است. مجموعه داده‌های جدول ۱ برگرفته از UCI به‌عنوان مجموعه داده‌های واقعی، برای آزمایش‌ها استفاده شده است [۲۸]. مجموعه داده‌ها در دو کلاس، شماره ۱ و ۲ دسته‌بندی شده‌اند که در زمان آموزش فقط از یک از این دو کلاس برای آموزش دسته‌بند استفاده شده است. در این حالت یک کلاس از مجموعه داده‌ها (کلاس ۱ یا کلاس ۲ مشخص‌شده در جدول ۱) به‌عنوان کلاس هدف در مرحله آموزش دسته‌بند استفاده می‌شود و کلاس دیگر به‌عنوان کلاس پرت در نظر گرفته می‌شود و در مرحله آزمایش دسته‌بند استفاده می‌شود. برای افراز مجموعه داده‌ها به مجموعه آموزشی و آزمایشی از روش اعتبارسنجی متقاطع ۱۰-تایی[11] استفاده شده است.

برای مقایسه عملکرد دسته‌بندها از معیار صحت دسته‌بند (معادله (۲۲)) و معیار نرخ مثبت صحیح ((معادله (۲۳)) استفاده شده است [۲۹].

جدول ۱: مجموعه داده‌های استفاده‌شده در آزمایش‌ها برگرفته از UCI

شماره	مجموعه داده	تعداد ویژگی	کلاس۱		کلاس۲	
			نام دسته	تعداد نمونه	نام دسته	تعداد نمونه
۱	Breast Cancer	۳۲	Benign	۳۵۷	Malignant	۲۱۲
۲	Ionosphere	۳۴	Good	۲۲۵	Bad	۱۲۶
۳	Pima Indians	۸	Yes	۵۰۰	No	۲۶۸
۴	Spambase	۵۷	Non-spam	۲۷۸۸	Spam	۱۸۱۳
۵	Sonar	۶۰	Rock	۹۷	Metal	۱۱۱

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{۲۲}$$

$$Recall = TPR = \frac{TP}{TP+FN} \tag{۲۳}$$

تعریف اصطلاحات مثبت صحیح (TP)، مثبت کاذب (FP)، منفی صحیح (TN) و منفی کاذب (FN) براساس جدول ۲ است.

جدول ۲: ماتریس درهم‌ریختگی[12] نقاط پیش‌بینی‌شده

	داده‌های پرت پیش‌بینی‌شده	داده‌های اصلی پیش‌بینی‌شده
داده‌های پرت واقعی	مثبت صحیح	منفی کاذب
داده‌های اصلی واقعی	مثبت کاذب	منفی صحیح

آزمایش‌ها بر روی دسته‌بندها در دو فرضیه صورت می‌پذیرد. در فرضیه اول مجموعه داده‌های اصلی فاقد نویز می‌باشند و در فرضیه دوم در مجموعه داده‌های اصلی، ۱۰ درصد از نقاط آن را به‌صورت تصادفی تغییر می‌دهیم تا مجموعه داده‌ها شامل نویز شود (با استفاده از نویز توزیع گوسی بر روی ویژگی‌ها [۳۰]). جدول ۳ و ۴ نتایج آزمایش‌ها بر روی مجموعه داده‌ها برای فرضیه اول (داده‌های بدون نویز) را نشان می‌دهد (براساس آموزش کلاس ۱ به‌عنوان کلاس اصلی). در جدول ۳ میزان صحت دسته‌بندها در مجموعه داده‌ها ارائه شده است.

جدول ۳: میانگین صحت دسته‌بندها در فرضیه داده‌های بدون نویز (کلاس۱- کلاس اصلی)

مجموعه داده‌ها	SVDD[13]	DW-SVDD[24]	R-SVDD[25]	KH-SVDD
Breast Cancer	۰/۹۲۶ (۴)	۰/۹۶۹ (۲)	۰/۹۶۷ (۳)	۰/۹۷۲ (۱)
Ionosphere	۰/۹۱۶ (۴)	۰/۹۸۲ (۱)	۰/۹۷۹ (۳)	۰/۹۸۱ (۲)
Pima Indians Diabetes	۰/۸۹۵ (۴)	۰/۹۵۱ (۱)	۰/۹۴۹ (۳)	۰/۹۵۰ (۲)
Spambase	۰/۸۷۹ (۴)	۰/۹۶۲ (۲)	۰/۹۶۲ (۱)	۰/۹۶۱ (۳)
Sonar	۰/۸۵۱ (۴)	۰/۹۴۲ (۲)	۰/۹۴۰ (۳)	۰/۹۴۲ (۱)

در جدول ۳ رتبه هر دسته‌بند داخل پرانتز نشان داده شده است. دسته‌بند پیشنهادی KH-SVDD، DW-SVDD و R-SVDD به‌علت استفاده از پارامتر تراکم در دسته‌بندی، نتایج بهتری را نسبت به دسته‌بند SVDD به‌دست آورده‌اند. هم‌چنین صحت دسته‌بندهای فوق نزدیک هم می‌باشند. جدول ۴ میانگین نرخ مثبت صحیح دسته‌بندها را در فرضیه اول نشان می‌دهد.

جدول ۴: میانگین نرخ مثبت صحیح دسته‌بندها در فرضیه داده‌های بدون نویز (کلاس۱- کلاس اصلی)

مجموعه داده‌ها	SVDD[13]	DW-SVDD[24]	R-SVDD[25]	KH-SVDD
Breast Cancer	۰/۹۴۱ (۴)	۰/۹۶۹ (۲)	۰/۹۶۷ (۳)	۰/۹۷۲ (۱)
Ionosphere	۰/۹۲۹ (۴)	۰/۸۸۵ (۲)	۰/۹۸۲ (۳)	۰/۹۸۸ (۱)
Pima Indians Diabetes	۰/۹۰۷ (۴)	۰/۹۶۳ (۱)	۰/۹۵۴ (۳)	۰/۹۵۸ (۲)
Spambase	۰/۸۹۱ (۴)	۰/۹۶۹ (۱)	۰/۹۶۷ (۲)	۰/۹۶۳ (۳)
Sonar	۰/۸۶۲ (۴)	۰/۹۵۱ (۳)	۰/۹۵۹ (۱)	۰/۹۵۷ (۲)

جدول ۵ میانگین صحت دسته‌بندها و جدول ۶ میانگین نرخ مثبت صحیح در فرضیه اول (داده‌های بدون نویز) براساس آموزش دسته‌بندها با استفاده از کلاس ۲ به‌عنوان کلاس اصلی را نشان می‌دهند.

دسته‌بند پیشنهادی KH-SVDD در بیشتر آزمایش‌های داده‌های بدون نویز، از نظر صحت و نرخ مثبت صحیح رتبه اول و دوم را دارا است.

جدول ۷ نتایج صحت دسته‌بندها و جدول ۸ نتایج نرخ مثبت صحیح، در فرضیه دوم (داده‌های دارای ۱۰درصد نویز) را نشان می‌دهند (کلاس ۱ به‌عنوان کلاس اصلی فرض شده است).

جدول ۹ نتایج صحت دسته‌بندها و جدول ۱۰ نتایج نرخ مثبت صحیح، در فرضیه دوم (داده‌های دارای ۱۰درصد نویز) را نشان می‌دهند. در این آزمایش کلاس ۲ به‌عنوان کلاس اصلی فرض شده است.

جدول ۵: میانگین صحت دسته‌بندها در فرضیه داده‌های بدون نویز (کلاس۲- کلاس اصلی)

مجموعه داده‌ها	SVDD[13]	DW-SVDD[24]	R-SVDD[25]	KH-SVDD
Breast Cancer	۰/۹۴۹ (۴)	۰/۹۷۴ (۲)	۰/۹۷۱ (۳)	۰/۹۷۹ (۱)
Ionosphere	۰/۹۴۲ (۴)	۰/۹۸۵ (۲)	۰/۹۷۹ (۳)	۰/۹۸۶ (۱)
Pima Indians Diabetes	۰/۹۳۱ (۴)	۰/۹۵۳ (۳)	۰/۹۵۹ (۱)	۰/۹۵۹ (۲)
Spambase	۰/۹۴۳ (۴)	۰/۹۶۶ (۱)	۰/۹۶۳ (۳)	۰/۹۶۵ (۲)
Sonar	۰/۹۵۱ (۴)	۰/۹۵۷ (۲)	۰/۹۵۴ (۳)	۰/۹۶۰ (۱)

همان طور که در جدول‌های ۱۰-۷ مشخص است صحت و نرخ مثبت صحیح دسته‌بندها در دسته‌بندی داده‌های دارای نویز به‌میزان قابل توجهی کاهش یافته است، دسته‌بند R-SVDD و دسته‌بند پیشنهادی KH-SVDD به‌دلیل داشتن رویکردی برای تشخیص داده‌های نویزی، نتایج بهتری را در مواجه با مجموعه داده‌های دارای نویز، از خود نشان داده‌اند. هم‌چنین نتایج دسته‌بند پیشنهادی به‌میزان قابل قبولی از دیگر دسته‌بندها بهتر است. در آزمایش دیگر، میزان نویز در مجموعه داده‌ها به‌صورت پله‌ای افزایش یافته است، تا تأثیر افزایش نویز بر رفتار دسته‌بندها بررسی شود.

جدول ۶: میانگین نرخ مثبت صحیح دسته‌بندها در فرضیه داده‌های بدون نویز (کلاس۲- کلاس اصلی)

مجموعه داده‌ها	SVDD[13]	DW-SVDD[24]	R-SVDD[25]	KH-SVDD
Breast Cancer	۰/۹۶۲ (۴)	۰/۹۸۱ (۲)	۰/۹۷۹ (۳)	۰/۹۸۳ (۱)
Ionosphere	۰/۹۵۹ (۴)	۰/۹۸۹ (۱)	۰/۹۸۲ (۳)	۰/۹۸۵ (۲)
Pima Indians Diabetes	۰/۹۴۹ (۴)	۰/۹۶۹ (۲)	۰/۹۶۷ (۳)	۰/۹۷۱ (۱)
Spambase	۰/۹۴۹ (۴)	۰/۹۶۹ (۱)	۰/۹۶۷ (۲)	۰/۹۶۵ (۳)
Sonar	۰/۹۵۴ (۴)	۰/۹۶۲ (۱)	۰/۹۵۶ (۳)	۰/۹۵۹ (۲)

در این آزمایش، روند تغییرات صحت دسته‌بندها در چهار مرحله بررسی شده است. مرحله اول مجموعه داده‌ها بدون نویز است، در مراحل دیگر با افزایش نویز، به ترتیب ۱۰،۲۰ و ۳۰ درصدی به مجموعه داده‌ها عملکرد دسته‌بندها مقایسه شده است. همان‌طور که در شکل ۸ مشخص است، صحت تمام دسته‌بندها، با افزایش نویز به‌میزان قابل توجهی کاهش یافته است.

جدول ۷: میانگین صحت دسته‌بندها در فرضیه داده‌های دارای نویز (کلاس۱- کلاس اصلی)

مجموعه داده‌ها	SVDD[13]	DW-SVDD[24]	R-SVDD[25]	KH-SVDD
Breast Cancer	۰/۷۹ (۴)	۰/۸۱ (۳)	۰/۸۵ (۱)	۰/۸۴ (۲)
Ionosphere	۰/۷۸ (۴)	۰/۸۱ (۳)	۰/۸۶ (۲)	۰/۸۷ (۱)
Pima Indians Diabetes	۰/۷۵ (۴)	۰/۸۲ (۳)	۰/۸۴ (۲)	۰/۸۶ (۱)
Spambase	۰/۷۲ (۴)	۰/۸۳ (۳)	۰/۸۵ (۲)	۰/۸۷ (۱)
Sonar	۰/۷۳ (۴)	۰/۸۱ (۳)	۰/۸۳ (۲)	۰/۸۵ (۱)

در شکل ۸ تغییرات صحت دسته‌بندها با افزایش نویز بر روی مجموعه داده‌های Ionosphere نشان داده شده است. شیب خط عملکرد دسته‌بند پیشنهادی (KH-SVDD) و R-SVDD ملایم‌تر از دو روش دیگر است. برای مجموعه داده‌های Ionosphere، اختلاف درصد صحت تشخیص نمونه‌های اصلی و پرت در دسته‌بندی پیشنهادی بین دو بازه نویز صفر و ۳۰ درصد، تقریباً ۳۲ درصد است. این مقدار برای دسته‌بند R-SVDD، تقریباً ۳۷ درصد است.

جدول ۸: میانگین نرخ مثبت صحیح دسته‌بندها در فرضیه داده‌های دارای نویز (کلاس۱- کلاس اصلی)

مجموعه داده‌ها	SVDD[13]	DW-SVDD[24]	R-SVDD[25]	KH-SVDD
Breast Cancer	۰/۸۲ (۴)	۰/۸۴ (۳)	۰/۸۹ (۱)	۰/۸۷ (۲)
Ionosphere	۰/۸۰ (۴)	۰/۸۴ (۳)	۰/۹۰ (۱)	۰/۸۹ (۲)
Pima Indians Diabetes	۰/۷۸ (۴)	۰/۸۲ (۳)	۰/۸۵ (۲)	۰/۸۸ (۱)
Spambase	۰/۷۷ (۴)	۰/۸۵ (۳)	۰/۸۷ (۲)	۰/۹۰ (۱)
Sonar	۰/۷۶ (۴)	۰/۸۴ (۳)	۰/۸۶ (۲)	۰/۸۸ (۱)

در شکل ۹، تغییرات صحت دسته‌بندها با افزایش نویز بر روی مجموعه داده‌های Breast cancer نشان داده شده است. افزایش ۳۰ درصدی نویز در مجموعه داده‌های فوق، باعث کاهش تقریبی عملکرد دسته‌بندها به‌میزان ۳۰ درصد در KH-SVDD، ۳۴ درصد در R-SVDD، ۵۲ درصد در DW-SVDD و کاهش ۵۵ درصدی در دسته‌بند SVDD شده است.

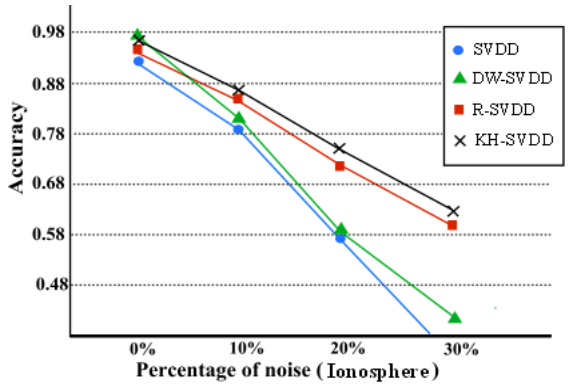
جدول ۹: میانگین صحت دسته‌بندها در فرضیه داده‌های دارای نویز (کلاس۲- کلاس اصلی)

مجموعه داده‌ها	SVDD[13]	DW-SVDD[24]	R-SVDD[25]	KH-SVDD
Breast Cancer	۰/۸۰ (۴)	۰/۸۴ (۳)	۰/۸۶ (۲)	۰/۸۷ (۱)
Ionosphere	۰/۷۹ (۴)	۰/۸۵ (۳)	۰/۸۷ (۱)	۰/۸۶ (۲)
Pima Indians Diabetes	۰/۷۸ (۴)	۰/۸۶ (۳)	۰/۸۷ (۲)	۰/۸۸ (۱)
Spambase	۰/۷۹ (۴)	۰/۸۵ (۳)	۰/۸۷ (۲)	۰/۸۸ (۱)
Sonar	۰/۷۹ (۴)	۰/۸۵ (۳)	۰/۸۶ (۲)	۰/۸۹ (۱)

جدول ۱۰: میانگین نرخ مثبت صحیح دسته‌بندها در فرضیه داده‌های دارای نویز (کلاس۲- کلاس اصلی)

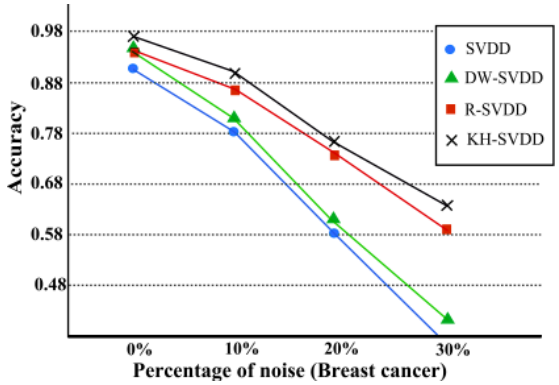
مجموعه داده‌ها	SVDD[13]	DW-SVDD[24]	R-SVDD[25]	KH-SVDD
Breast Cancer	۰/۸۶ (۴)	۰/۸۸ (۳)	۰/۹۰ (۲)	۰/۹۱ (۱)
Ionosphere	۰/۸۱ (۴)	۰/۸۵ (۳)	۰/۹۰ (۱)	۰/۸۷ (۲)
Pima Indians Diabetes	۰/۸۱ (۴)	۰/۸۸ (۳)	۰/۸۹ (۲)	۰/۹۱ (۱)
Spambase	۰/۷۷ (۴)	۰/۸۴ (۳)	۰/۸۸ (۲)	۰/۸۹ (۱)
Sonar	۰/۸۳ (۴)	۰/۸۷ (۳)	۰/۹۰ (۲)	۰/۹۲ (۱)

در شکل ۱۰، تغییرات صحت دسته‌بندها با افزایش نویز بر روی مجموعه داده‌های Pima indians diabetes نشان داده شده است. با افزایش ۳۰ درصدی نویز در مجموعه داده‌های فوق، عملکرد دسته‌بندهای KH-SVDD ۳۷ درصد، R-SVDD ۴۱ درصد، DW-SVDD ۵۹ درصد و SVDD ۶۱ درصد تقریباً کاهش یافته است.



شکل ۸: تغییرات صحت دسته‌بندها با افزایش نویز(Ionosphere)

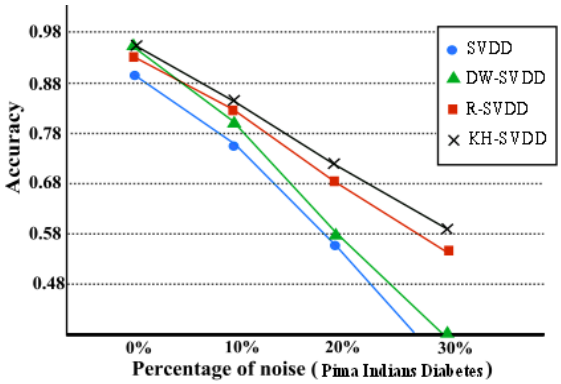
در شکل ۱۱ برای مجموعه داده‌های Spambase، کاهش تقریبی عملکرد KH-SVDD ۳۸ درصد، R-SVDD ۴۰ درصد، DW-SVDD ۵۶ درصد و SVDD ۵۸ درصد با افزایش ۳۰ درصدی نویز، به‌دست آمده است.



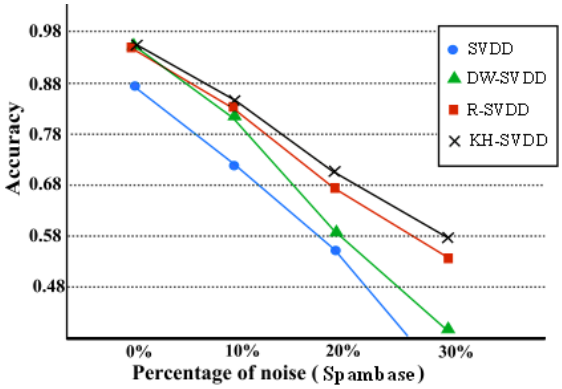
شکل ۹: تغییرات صحت دسته‌بندها با افزایش نویز(Breast)

در شکل ۱۲، تغییرات صحت دسته‌بندها با افزایش نویز بر روی مجموعه داده‌های Sonar نشان داده شده است. با افزایش ۳۰ درصدی نویز، صحت دسته‌بندهای KH-SVDD ۳۴ درصد، R-SVDD ۳۵

درصد، DW-SVDD ۵۲ درصد و SVDD ۵۲ درصد به‌طور تقریبی کاهش یافته است.



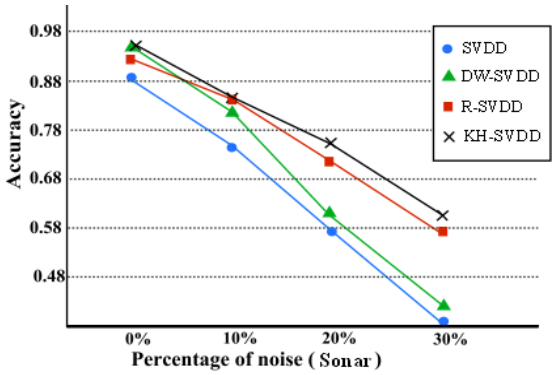
شکل ۱۰: تغییرات صحت دسته‌بندها با افزایش نویز(Diabetes)



شکل ۱۱: تغییرات صحت دسته‌بندها با افزایش نویز(Spambase)

نتایج آزمایش‌ها نشان می‌دهد که دسته‌بند پیشنهادی نسبت به دسته‌بندهای بررسی‌شده دیگر، در هنگام افزایش نویز نتایج بهتری را به‌دست آورده است. کارآیی دو دسته‌بند SVDD و DW-SVDD به‌علت نداشتن رویکردی برای مقابله با داده‌های نویزی، در زمان حضور نویز دچار افت شدیدی شده‌اند. روش R-SVDD و KH-SVDD با توجه به داشتن رویکردی برای مقابله با نویز، نسبت به دو روش دیگر کارآیی بهتری را در آزمایش‌ها کسب کردند.

یکی از مشکلات روش پیشنهادی افزایش زمان اجرای الگوریتم در مرحله آموزش دسته‌بند است. یافتن مرکز مناسب برای کلاس اصلی داده‌ها توسط الگوریتم میگو انجام می‌شود که یک سربار اضافی بر روی دسته‌بند ایجاد می‌کند.



شکل ۱۲: تغییرات صحت دسته‌بندها با افزایش نویز(Sonar)

۶- نتیجه‌گیری و کارهای آینده

در این مقاله، یک دسته‌بند تک‌کلاسه برای داده‌های نویزی ارائه شد. موقعیت مرکز اَبَرکُره در فضای ویژگی‌های داده‌ها با استفاده از الگوریتم SVDD و الگوریتم گروه میگوی آشوبی محاسبه شد. هم‌چنین برای کاهش تأثیر داده‌های نویزی بر مرز اَبَرکُره، از تراکم محلی نقاط و فاصله آنها تا مرکز اَبَرکُره به‌عنوان یک تابع جریمه استفاده شده است. در دسته‌بند KH-SVDD، سعی بر تشخیص داده‌های نویزی در مجموعه داده‌های اصلی شده است. فرض کلی که در این روش وجود دارد این است که تنها مجموعه داده‌های اصلی موجود است و دانشی درباره داده‌های پرت وجود ندارد. نتایج آزمایش‌ها نشان می‌دهد روش پیشنهادی نسبت به روش‌های مشابه دیگر، از صحت بالاتری در تشخیص داده‌های پرت در مجموعه داده‌های نویزی، برخوردار است. در کارهای آینده برای تشخیص داده‌های نویزی، می‌توان علاوه بر تراکم از پارامترهای آماری مانند میانگین، مد و انحراف معیار نیز استفاده کرد.

مراجع

[۱] محمدعلی زارع چاهوکی و سیدحمیدرضا محمدی، «بهینه‌سازی هسته‌های چندگانه در ماشین بردار پشتیبان جفتی برای کاهش شکاف معنایی تشخیص صفحات فریب‌آمیز»، مجله مهندسی برق دانشگاه تبریز، شماره ۴ جلد ۴۶، ۱۳۵-۱۴۵، ۱۳۹۵.

[۲] محمدامیر عباسیان و حسین نظام‌آبادی‌پور، «الگوریتم جستجوی گرانشی چندهدفه مبتنی بر مرتب‌سازی جبهه‌های مغلوب‌نشده»، مجله مهندسی برق دانشگاه تبریز، شماره ۱ جلد ۴۱، ۶۸-۸۰، ۱۳۹۰.

[۳] سیدحسین غفاریان، هادی صدوقی یزدی و یونس الله‌یاری، «دسته‌بند تک‌کلاسه گرانش‌گرای مبتنی بر ماشین بردار پشتیبان»، نشریه مهندسی برق و مهندسی کامپیوتر ایران، سال ۱۰، شماره ۲، ۱۳۹۱.

[۴] وحیده منعمی‌زاده و جواد حمیدزاده، «جستجوی k نزدیک‌ترین همسایه تقریبی به روش ترکیب خطی»، نشریه مهندسی برق و مهندسی کامپیوتر ایران، آماده انتشار.

[5] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," Artificial Intelligence and Cognitive Science, vol. 6206, pp. 188-197, 2010.
[6] A. Wenjuan, M. Liang and H. Liu, "An improved one-class support vector machine classifier for outlier detection," Mechanical Engineering Science, vol. 229, pp. 580-588, 2015.
[7] S. S. Khan and M. G. Madden, "One-Class Classification: Taxonomy of Study and Review of Techniques," The Knowledge Engineering Review, vol. 29, pp. 1-30, 2014.
[8] S. Kang, S. Cho and P. Kang, "Multi-class classification via heterogeneous ensemble of one-class classifiers," Engineering Applications of Artificial Intelligence, vol. 43,pp. 35–43, 2015.
[9] L. Zhang, L. Xingning, W. Bangjun and H. Shuping, "Similarity learning based on multiple support vector data description," Neural Networks (IJCNN), pp. 1-7, 2015.
[10] D. M. Tax and R.P. Duin, "Uniform object generation for optimizing one-class classifiers," The Journal of Machine Learning Research, vol. 2, pp. 155-173, 2002.
[11] R. Sadeghi and J. Hamidzadeh, "Automatic Support Vector Data Description," Soft Computing, 2016, DOI: 10.1007/s00500-016-2317-5.
[12] V. H. Moghaddam, and J. Hamidzadeh, "New Hermite orthogonal polynomial kernel and combined kernels in Support Vector Machine classifier," Pattern Recognition, vol. 60, pp. 921-935, 2016.
[13] D. M. Tax and R.P. Duin, "Support vector data description," Machine Learning, vol. 54, pp. 45–66, 2004.
[14] J. Bootkrajang, "A generalised label noise model for classification in the presence of annotation errors," Neurocomputing, vol. 192, pp. 61–71, 2016.
[15] J. Hamidzadeh, R. Monsefi and H. SadoghiYazdi, "IRAHC: Instance Reduction Algorithm using Hyperrectangle Clustering," Pattern Recognition, vol. 48, pp.1878-1889, 2015.
[16] J. Hamidzadeh, R. Monsefi and H. SadoghiYazdi, "LMIRA: Large Margin Instance Reduction Algorithm," Neurocomputing, vol. 145, pp. 477-487, 2014.
[17] S. Y. Xia, Z. Xiong, Y. He, K. Li, L. M. Dong and M. Zhang, "Relative density-based classification noise detection," Optik International Journal for Light and Electron Optics, vol. 125, pp. 6829–6834, 2014.
[18] K. Lee, D. Kim, K. H. Lee and D. Lee, "Density-induced support vector data description," Neural Networks, IEEE Transactions on, vol. 18, pp. 284–289, 2007.
[19] C. K. Wang, Y. Ting, Y. H. Liu and G. Hariyanto, "A Novel Approach to Generate Artificial Outliers for support Vector Data Description," IEEE International Symposium on Industrial Electronics (ISIE), Korea, pp. 2202-2207, 2009.
[20] H. W. Cho, "Data description and noise filtering based detection with its application and performance comparison," Expert systems with Applications, vol. 36, no. 1, pp. 434-441, 2009.
[21] S. M. Guo, L. C. Chen and J. S. Tsai, "A boundary method for outlier detection based on support vector domain description," Pattern Recognition, vol. 42, pp. 77-83, 2009.
[22] G. X. Huang, H. F. Chen and F. Yin, "Improved support vector data description," International Conference on Machine Learning and Cybernetics, vol. 3, pp. 1459-1463, 2010.
[23] B. Liu, Y. Xiao, L. Cao, Z. Hao and F. Deng, "SVDD-based outlier detection on uncertain data," Knowledge and Information Systems, vol. 34, pp. 597-618, 2013.
[24] M. Cha, J. Kim and J. Baek, "Density weighted support vector data description," Expert Systems with Applications, vol. 41, pp. 3343–3350, 2014.
[25] G. Chen, X. Zhang, Z. Wang and F. Lia, "Robust support vector data description for outlier detection with noise or uncertain data," Knowledge-Based Systems, vol. 90, pp. 129–137, 2015.
[26] S. Kim, Y. Choi and M. Lee, "Deep learning with support vector data description," Neurocomputing, vol. 165, pp. 111–117, 2015.
[27] G. Wang, L. Guo, A. Gandomi, G. Hao and H. Wang, "Chaotic Krill Herd algorithm," Information Sciences, vol. 274, pp. 17–34, 2014.
[28] A. Asuncion and D. Newman, UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA, 2013.
[29] S.S. Khan, J. Hoey and D. Lizotte, "Bayesian multiple imputation approaches for one-class classification," Advances in Artificial Intelligence, pp. 331–336, 2012.
[30] B. Liu, Y. Xiao and Z. Hao, "An efficient approach for outlier detection with imperfect data labels," IEEE Trans Knowl. Data Eng, vol. 26, pp. 1602-1616, 2014.

زیرنویس‌ها

[1] Outlier detection
[2] One-class classification
[3] Classification
[4] Krill herd support vector data description
[5] Support vector data description
[6] Kernel trick
[7] Parzen
[8] Post-processing
[9] Likelihood
[10] Overfitting
[11] 10-fold cross validation
[12] Confusion