مقاله پژوهشی

تشخیص داده پرت در دادگان با ابعاد بالا با استفاده از انتخاب زیرفضای مرتبط محلی مبتنی بر آنتروپی

محبوبه ریاحی مدوار، احمد اکبری ازیرانی و بابک ناصرشریف

چکیده: یکی از چالش‌های مسئله تشخیص داده پرت با ابعاد بالا، طلسم بعد است که در آن برخی ابعاد (ویژگی‌ها) منجر به پنهان‌شدن داده‌های پرت می‌گردند. برای حل این مسئله، ابعادی که حاوی اطلاعات ارزشمندی در دادگان با ابعاد بالا جهت تشخیص داده پرت هستند، جستجو می‌شوند تا با نگاشت دادگان به زیرفضای متشکل از این ابعاد مرتبط، داده‌های پرت برجسته‌تر و قابل شناسایی شوند. این مقاله با معرفی یک روش جدید انتخاب زیرفضای مرتبط محلی و توسعه یک رویکرد امتیازدهی داده پرت مبتنی بر چگالی محلی، امکان تشخیص داده پرت در دادگان با ابعاد بالا را فراهم می‌نماید. در ابتدا، یک الگوریتم برای انتخاب زیرفضای مرتبط محلی بر اساس آنتروپی محلی ارائه می‌شود تا بتواند برای هر نقطه داده با توجه به داده‌های همسایه‌اش یک زیرفضای مرتبط انتخاب کند. سپس هر نقطه داده در زیرفضای انتخابی متناظرش با یک روش امتیازدهی پرت محلی مبتنی بر چگالی امتیازدهی می‌شود، به طوری که با در نظر گرفتن یک پهنای باند تطبیقی جهت تخمین چگالی هسته سعی می‌شود که اختلاف جزئی بین چگالی یک نقطه داده نرمال با همسایه‌هایش از بین رفته و به اشتباه به عنوان داده پرت تشخیص داده نشود و در عین حال، تخمین کمتر از مقدار واقعی چگالی در نقاط داده پرت، منجر به برجسته‌شدن این نقاط داده گردد. در پایان با آزمایش‌های تجربی روی چندین دادگان دنیای واقعی، الگوریتم پیشنهادی تشخیص داده پرت زیرفضای مبتنی بر آنتروپی محلی با چند تکنیک تشخیص داده پرت بر حسب دقت تشخیص مقایسه شده است. نتایج تجربی نشان می‌دهد که الگوریتم پیشنهادی مبتنی بر معیار آنتروپی محلی و روش پیشنهادی امتیازدهی داده پرت توانسته است به دقت بالایی جهت تشخیص داده پرت دست یابند.

کلیدواژه: **تشخیص داده پرت، داده‌های با ابعاد بالا، انتخاب زیرفضای مرتبط محلی، آنتروپی محلی.**

۱- مقدمه

تشخیص داده پرت یکی از مهم‌ترین و چالش‌برانگیزترین وظایف داده‌کاوی است که اشاره به مسئله یافتن نقاط داده‌ای دارد که مشخصات آنها با اکثریت داده‌ها به طور قابل توجهی متفاوت است. تشخیص داده پرت، کاربردهای گسترده‌ای در زمینه‌های تشخیص تقلب[1]، تشخیص خطا[2]، تشخیص نفوذ[3]، تجزیه و تحلیل بیولوژیکی[4] و بازاریابی[5] [۱] دارد. در کاربردهای دنیای واقعی، ماهیت مسئله تشخیص داده پرت، یادگیری بدون نظارت است، زیرا داده‌های پرت به ندرت رخ می‌دهند و هزینه برچسب‌گذاری آنها سنگین است.

تشخیص داده پرت در زمینه‌های پژوهشی زیادی مورد توجه قرار گرفته است اما بسیاری از تکنیک‌های تشخیص داده پرت موجود، روی داده‌های با ابعاد بالا با توجه به مسئله طلسم بعد[6]، چالش‌های قابل توجهی دارند. برخی پژوهش‌های اخیر، روش‌هایی مبتنی بر زیرفضا ارائه کرده‌اند تا با انتخاب زیرفضایی با ابعاد کمتر بتوانند داده‌های پرتی را که مخفی شدند شناسایی کنند [۲] و [۳]. به دلیل ماهیت پیچیده داده‌های پرت، ممکن است یک داده در یک زیرفضا قابل تشخیص باشد، اما داده پرت دیگر در آن زیرفضا قابل تشخیص نباشد. بنابراین هیچ زیرفضایی وجود ندارد که به طور کامل مشخصات تمام نقاط داده پرت را در بر داشته باشد. بنابراین در روش‌هایی که با دیدگاه سراسری، تنها یک زیرفضای مرتبط برای تشخیص داده پرت انتخاب می‌شود، ممکن است تمام داده‌های پرت در آن زیرفضا قابل تشخیص نباشند.

در این مقاله برای مقابله با مشکلات بالا یک روش تشخیص داده پرت زیرفضای مبتنی بر آنتروپی محلی (LESOD)[7] پیشنهاد داده می‌شود که شامل دو گام مجزای انتخاب زیرفضای مرتبط و سپس، امتیازدهی و تشخیص داده‌های پرت است. این جداسازی این امکان را فراهم می‌کند که بتوان به طور مجزا روی هر یک از این دو حوزه تحقیق کرد. در این مقاله، ابتدا با پیشنهاد یک روش جدید انتخاب زیرفضای مرتبط مبتنی بر آنتروپی و سپس، توسعه یک روش امتیازدهی داده پرت، سعی بر بهبود دقت تشخیص داده پرت می‌شود. انگیزه روش پیشنهادی انتخاب زیرفضای مرتبط محلی، انتخاب یک زیرفضای مرتبط برای هر نقطه داده است. هدف از انتخاب زیرفضای مرتبط برای یک نقطه داده، حذف ویژگی‌های بی‌ربط است به طوری که نقطه داده مفروض در راستای این ویژگی‌های بی‌ربط دارای اطلاعات کمی است. بنابراین با حذف ویژگی‌های بی‌ربط برای هر نقطه داده و تنها نگه‌داشتن ویژگی‌های مرتبط که حاوی اطلاعات بالایی جهت تشخیص داده پرت هستند،

این مقاله در تاریخ ۱۳ اردیبهشت ماه ۱۴۰۰ دریافت و در تاریخ ۲۸ شهریور ماه ۱۴۰۰ بازنگری شد.

محبوبه ریاحی مدوار، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران، (email: m_riahi@comp.iust.ac.ir).

احمد اکبری ازیرانی (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران، (email: akbari@iust.ac.ir).

بابک ناصرشریف، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران، (email: bnasersharif@kntu.ac.ir).

1. Fraud Detection
2. Fault Detection
3. Intrusion Detection
4. Biological Analysis
5. Marketing
6. Curse of Dimensionality
7. Local Entropy-Based Subspace Outlier Detection

می‌توان با کارایی بالایی، نقاط داده پرت را در دادگان با ابعاد بالا شناسایی کرد. برای رتبه‌بندی نقاط داده پرت، یک روش جدید امتیازدهی مبتنی بر چگالی تطبیقی با الهام از روش [۴] پیشنهاد داده می‌شود که برای هر نقطه داده، یک امتیاز پرت با مقایسه چگالی‌اش با چگالی نقاط همسایگی‌اش در زیرفضای انتخابی متناظر با نقطه داده مفروض تخصیص می‌دهد.

به طور کلی، نوآوری‌های اصلی این مقاله می‌توانند به صورت زیر خلاصه‌سازی شوند:

۱) به کارگیری مفهوم آنتروپی محلی به همراه مفهوم اطلاعات برای تعیین باربط/ بی‌ربط‌بودن یک ویژگی که منجر به افزایش دقت تشخیص داده پرت می‌گردد.

۲) یک پهنای باند تطبیقی برای تخمین چگالی هسته پیشنهاد داده می‌شود. هنگام محاسبه و مقایسه چگالی یک نقطه داده با همسایه‌هایش، یک پهنای باند تطبیقی برای نقطه داده مفروض، محاسبه و از این همین پهنای باند برای تخمین چگالی نقاط داده همسایگی استفاده می‌شود. بدین ترتیب انتظار می‌رود که در یک ناحیه چگال، با در نظر گرفتن یک پهنای باند یکسان برای تخمین چگالی نقاط همسایگی، اختلاف‌های ناچیز بین مقادیر چگالی از بین رفته و نرخ مثبت کاذب کاهش یابد.

ساختار مقاله به صورت زیر است: ابتدا یک مرور خلاصه‌ای از کارهای مرتبط در بخش ۲ معرفی شده است. بخش ۳، روش پیشنهادی تشخیص داده پرت زیرفضا، شامل دو الگوریتم جدید انتخاب زیرفضای مرتبط مبتنی بر آنتروپی و امتیازدهی داده پرت مبتنی بر چگالی را تشریح می‌کند. آزمایش‌های تجربی و تحلیل نتایج آنها در بخش ۴ بیان شده و بخش ۵، این مقاله را با چند نتیجه‌گیری و کارهای آتی خاتمه می‌دهد.

۲- پژوهش‌های مرتبط

روش‌های تشخیص داده پرت را می‌توان با توجه به معیارهای مختلف، تقسیم‌بندی کرد:

با توجه به موجودبودن برچسب داده‌ها، تکنیک‌های تشخیص داده پرت را می‌توان به سه طبقه بانظارت، نیمه‌نظارتی و بدون نظارت تقسیم‌بندی کرد [۵]. در روش‌های بانظارت، برچسب داده‌ها به صورت نرمال و پرت وجود دارند و در روش‌های نیمه‌نظارتی، تمام نمونه‌ها متعلق به کلاس نرمال هستند. در تکنیک‌های بدون نظارت، هیچ داده برچسب‌داری موجود نیست. در این تکنیک‌ها فرض می‌شود که نمونه‌های پرت از سایر نمونه‌ها دور هستند. تکنیک‌های بدون نظارت در مسئله تشخیص داده پرت عملی‌تر هستند، زیرا جمع‌آوری یک مجموعه داده برچسب‌دار که کل رفتار نمونه‌های پرت را پوشش دهد دشوار است.

با توجه به تعداد ابعاد مورد استفاده در تشخیص داده پرت، الگوریتم‌های تشخیص داده پرت به دو دسته فضای کامل[1] و زیرفضا تقسیم می‌گردند. روش‌های بدون نظارت تشخیص داده پرت فضای کامل را بر اساس تکنیک‌های به کار گرفته شده می‌توان به روش‌های مبتنی بر خوشه‌بندی، مبتنی بر مدل‌های آماری و مبتنی بر نزدیک‌ترین همسایه‌ها تقسیم‌بندی کرد. تکنیک‌های مبتنی بر خوشه‌بندی، داده‌هایی را که متعلق به هیچ خوشه‌ای نیستند یا متعلق به خوشه‌های بسیار کوچک هستند، به عنوان داده پرت در نظر می‌گیرند. تکنیک‌های مبتنی بر مدل‌های آماری [۶] فرض می‌کنند که توزیع داده‌ها از یک مدل آماری پیروی می‌کند و سپس یک تست آماری برای تعیین نقاط داده‌ای که با مدل مفروض، سازگار نیستند به کار گرفته می‌شود. در روش‌های مبتنی بر نزدیک‌ترین همسایه‌ها، هر نقطه داده با توجه به همسایگی محلی‌اش تحلیل می‌شود. این روش‌ها با توجه به این که داده‌های پرت بر اساس چه معیاری با نزدیک‌ترین همسایه‌ها مقایسه و شناسایی می‌شوند، به دو دسته روش‌های مبتنی بر فاصله [۷] تا [۹] و مبتنی بر چگالی [۱۰] تقسیم می‌گردند. در روش‌های مبتنی بر فاصله، زمانی که چگالی نواحی مختلف داده متفاوت باشند، تکنیک‌های مبتنی بر فاصله، ضعیف عمل می‌کنند. روش‌های مبتنی بر چگالی نه تنها چگالی هر نقطه، بلکه چگالی نقاط همسایگی را نیز محاسبه می‌کنند. اگر چگالی یک نقطه داده نسبت به همسایگانش بسیار کمتر باشد، به عنوان داده پرت شناسایی می‌شود. همچنین در دسته‌بندی دیگری بر اساس اندازه مجموعه مرجع (کل/ بخشی از نقاط دادگان) برای محاسبه امتیاز پرت یک نقطه، روش‌های تشخیص داده پرت را می‌توان به دو دسته محلی و سراسری تقسیم کرد. الگوریتم شناخته‌شده عامل پرت محلی (LOF)[2] [۱۰]، اولین روش محلی مبتنی بر چگالی است که به هر نقطه داده، یک امتیاز پرت محلی تخصیص می‌دهد. امتیاز LOF برای هر نقطه داده با توجه به نرخ میانگین چگالی محلی $k-$نزدیک‌ترین همسایه‌اش (kNN)[3] و چگالی محلی آن نقطه داده تعیین می‌شود. در روش تشخیص ناهنجاری مبتنی بر چگالی تطبیقی (Adaptive-KD)[4] [۴]، یک روش امتیازدهی پیشنهاد شده که برای بهبود قدرت تفکیک بین داده‌های نرمال و پرت، یک پهنای باند تطبیقی به کار گرفتند: در مناطق با چگالی بالا، پهنای باند عریض برای برطرف‌نمودن اختلاف چگالی بین نمونه‌های نرمال و در نواحی با چگالی پایین، از پهنای باند باریک برای تشدید ناهنجاری نمونه‌های پرت بالقوه استفاده می‌کنند.

اکثر روش‌های رایج تشخیص داده پرت، کل فضای ویژگی را برای یافتن داده‌های پرت استفاده می‌کنند. در داده‌های با ابعاد بالا به دلیل مشکل طلسم بعد، حضور تعداد بالای ابعاد بی‌ربط در دادگان منجر به پنهان‌شدن اطلاعات مرتبط می‌گردد. در نتیجه، در روش‌های مبتنی بر کل ابعاد، داده پرت به راحتی مخفی می‌شود [۱۱]. روش‌های مبتنی بر زیرفضا برای کاهش اثر طلسم بعد، داده‌های پرت را در یک/ چند زیرفضا با ابعاد کمتر جستجو می‌کنند. در این روش‌ها، یافتن یک/ چند زیرفضا با ابعاد کمتر که بتواند داده‌های پرت را از سایر داده‌ها تفکیک نماید، به دلیل اندازه‌گیری میزان ارتباط یک ویژگی با مسئله بدون نظارت تشخیص داده پرت، چالش‌برانگیز است. همچنین این روش‌ها به دلیل نیاز به جستجوی زیرفضاهایی در فضای داده با ابعاد بالا، پیچیدگی محاسباتی بالایی دارند. روش‌های مبتنی بر زیرفضا در دو نوع روش‌های زیرفضای تنک [۱۲] تا [۱۴] و روش‌های زیرفضای مرتبط [۱۵] تا [۱۹] توسعه داده شدند:

۱) روش‌های تشخیص داده پرت مبتنی بر زیرفضای تنک، داده‌های پرت را در زیرفضاهایی با ابعاد پایین‌تر جستجو می‌کنند. در این روش‌ها، تمام نقاط دادگان به زیرفضای با ابعاد پایین‌تر، نگاشت می‌شوند و نمونه‌هایی که در زیرفضای تنک دارای چگالی کمتر از میانگین هستند به عنوان نمونه پرت شناسایی می‌شوند. آگاروال در اولین روش تشخیص داده پرت زیرفضا [۱۲]، از نمایش گرید برای محاسبه ضرایب تنکی زیرفضاهای مختلف استفاده کرده و با

1. Full-Dimensional

2. Local Outlier Factor

3. k-Nearest Neighbors

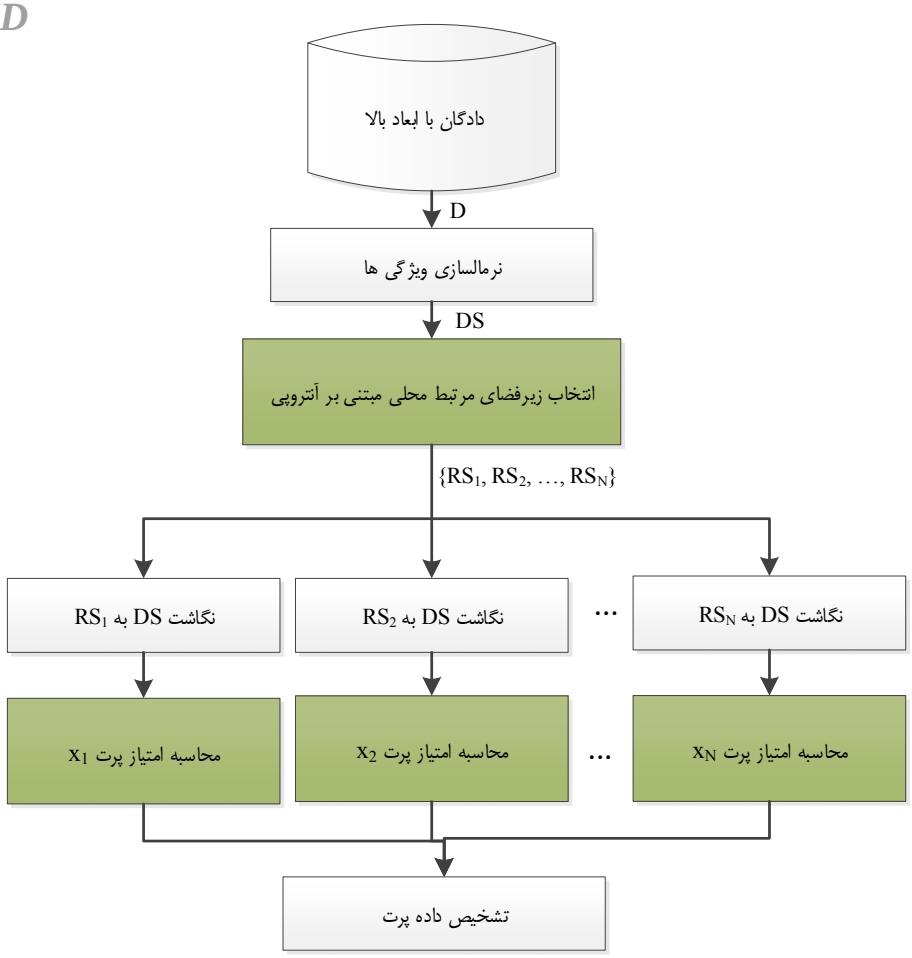
4. Adaptive Kernel Density-Based Anomaly Detection

بهره‌گیری از الگوریتم ژنتیک برای جستجوی کاراتر زیرفضاهای تنک، توانست عملکرد جستجو را بهبود دهد. اما کارایی الگوریتم ژنتیک، وابسته به پارامترهای مختلفی از قبیل جمعیت اولیه، تابع برازش و ... است و بنابراین الگوریتم ژنتیک نمی‌تواند کامل‌بودن نتایج جستجو را تضمین کند. در یک الگوریتم تشخیص داده پرت دیگر [۱۳] با بهره‌گیری از شبکه مفهومی[1] به عنوان ابزاری برای نمایش زیرفضاها و با معرفی مفهوم چگالی، کامل‌بودن نتایج جستجو تضمین شده است. اما ساخت شبکه مفهومی از زیرفضاها، پیچیده است و منجر به کارایی پایین می‌گردد. برخلاف روش‌های قبلی [۱۲]، [۱۳] و [۲۰] که زیرفضای تنک را با توجه به کل دادگان جستجو می‌کردند، یک الگوریتم تشخیص داده پرت محلی (LOMA)[2] [۱۴] مبتنی بر تحلیل ارتباط ویژگی با نمونه‌های پرت پیشنهاد داده است که ابعاد و نقاط داده بی‌ربط در دادگان را با به کارگیری تحلیل ارتباط ویژگی همراه با یک آستانه عامل تنکی هرس می‌کند. LOMA با به کارگیری روش بهینه‌سازی ازدحام ذرات[3] روی دادگان کاهش‌یافته، زیرفضای تنک را جستجو می‌کند. روش LOMA نقاط داده بی‌ربط را هرس می‌کند و تعداد ویژگی‌ها را کاهش می‌دهد. بنابراین در دادگان با ابعاد بالا کارامدتر است اما توسط دقت تشخیص محدود می‌شود.

۲) روش‌های تشخیص داده پرت مبتنی بر زیرفضای مرتبط، یک/ چند زیرفضا را شامل تعدادی بعد معنادار که داده‌های پرت را برجسته می‌نمایند انتخاب می‌کنند. در روش تشخیص داده پرت زیرفضا (SOD)[4] [۱۵]، برای هر نقطه داده، یک مجموعه مرجع با استفاده از نزدیک‌ترین همسایه مشترک[5] می‌گردد. زیرفضای مرتبط با توجه به ابعادی تعیین می‌شود که در نقاط مجموعه مرجع دارای واریانس پایینی هستند. درجه پرت زیرفضا برای یک نقطه داده با توجه به فاصله‌اش تا صفحه‌ای شامل ابعاد مرتبط، تعریف می‌شود. روش SOD در موارد بسیاری به دلیل نادیده‌گرفتن تعامل بین ابعاد، دچار مشکل می‌شود [۲۱]. در روش رتبه‌بندی داده پرت زیرفضا (OUTRES)[6] [۱۶]، مجموعه زیرفضاهای مرتبط شامل زیرفضاهایی که دارای توزیع غیر یکنواخت هستند با استفاده از جستجوی پایین به بالا[7] شناسایی می‌شوند. در ارزیابی یکنواختی توزیع یک زیرفضا از تست آماری کولموگروف- اسمیرنف[8] استفاده می‌شود. روش OUTRES [۱۶] برای محاسبه امتیاز پرت، نیاز به تعداد زیادی دادگان محلی دارد و همچنین دارای پیچیدگی محاسباتی نمایی بر حسب تعداد ابعاد می‌باشد. بدین ترتیب روش OUTRES [۱۶] روی دادگان با ابعاد بالا و حجیم، دارای مقیاس‌پذیری ضعیفی می‌باشد. روش احتمال پرت همبستگی (COP)[9] [۱۹]، از تحلیل مؤلفه اصلی[10] در دادگان محلی برای جستجوی زیرفضاهای جهت‌دار دلخواه با کمترین واریانس استفاده می‌کند و داده‌های پرت را بر اساس میزان انحراف از توزیع دادگان محلی تعیین می‌نماید. بنابراین، این روش برای بازتاب مشخصات توزیع دادگان و نهایتاً تعیین میزان انحراف، به دادگان محلی به اندازه کافی بزرگ نیاز دارد. در روش زیرفضاهای با کنتراست بالا برای رتبه‌بندی داده پرت مبتنی بر چگالی (HiCS)[11] [۱۷]، با فرض استقلال ویژگی‌ها، کنتراست یک زیرفضا را با میزان همبستگی بین ویژگی‌های آن زیرفضا، اندازه‌گیری می‌کنند و زیرفضاهای با کنتراست بالا جستجو می‌شوند. معیار کنتراست زیرفضا با استفاده از محاسبه اطلاعات متقابل به صورت تخمین مونت‌کارلو اندازه‌گیری می‌شود. سپس میزان پرت‌بودن نقاط داده بر اساس نگاشت نقاط دادگان محلی به زیرفضای مرتبط، اندازه‌گیری می‌شوند. در HiCS برای محاسبه اطلاعات متقابل، اختلاف بین توزیع شرطی و حاشیه‌ای در یک بعد تصادفی از زیرفضای مورد نظر اندازه‌گیری می‌شود و بدین ترتیب با این انتخاب تصادفی، ممکن است برخی زیرفضاهای مرتبط از دست بروند [۲۲]. همچنین فرض استقلال ویژگی‌ها و انتخاب یک زیرمجموعه از ویژگی‌های وابسته ممکن است معتبر نباشد، زیرا برخی ویژگی‌ها ممکن است با تشخیص داده پرت قویاً مرتبط باشند اما هیچ وابستگی با سایر ویژگی‌ها نداشته باشند. تعیین زیرفضای مرتبط در HiCS مبتنی بر یک رویکرد سراسری است، در حالی که در الگوریتم تشخیص داده پرت ضمنی در زیرفضای مرتبط دلخواه (COAS)[12] [۱۸]، یک الگوریتم محلی برای تشخیص داده پرت ضمنی مبتنی بر زیرفضای مرتبط روی دادگان حجیم و با ابعاد بالا پیشنهاد داده شده است. زیرفضای مرتبط در COAS با استفاده از ابعادی که تنکی محلی از یک حد آستانه بزرگ‌تر باشد، ساخته می‌شود. سپس با تعریف یک فرمول محاسبه عامل پرت محلی احتمالی[13] در زیرفضای مرتبط، درجه پرت‌بودن نقاط داده‌ای که از توزیع دادگان محلی پیروی نمی‌کنند محاسبه می‌گردد. اخیراً یک روش تشخیص داده پرت زیرفضای مبتنی بر چگالی [۲] با معرفی یک نمایش مبتنی بر چگالی، امکان ارائه دو معیار جدید مبتنی بر چگالی برای انتخاب زیرفضای مرتبط فراهم شده است. در معیار اول، ماکسیمم- ارتباط- با- چگالی[14] با استفاده از اطلاعات متقابل ویژگی‌هایی که بیشترین ارتباط با چگالی داده‌ها را دارند محاسبه می‌شود. در معیار دوم، مینیمم- افزونگی- ماکسیمم- ارتباط- با- چگالی[15] با به کارگیری مفهوم افزونگی بین ویژگی‌ها سعی می‌شود یک زیرفضای فشرده شامل ویژگی‌هایی که ماکسیمم ارتباط با چگالی و در عین حال، کمترین افزونگی بین آنهاست انتخاب گردد. نقاط داده پرت در زیرفضای انتخابی با استفاده از الگوریتم امتیازدهی داده پرت LOF امتیازدهی و شناسایی می‌شوند. این روش [۲] تنها یک زیرفضا برای تشخیص داده‌های پرت در کل دادگان انتخاب می‌کند و این در حالی است که در برخی دادگان ممکن است یک زیرفضا نتواند تمام داده‌های پرت را نمایش دهد.

به طور خلاصه، مطالعه کارهای پیشین تشخیص داده پرت بیان می‌کند که یک روش دقیق و کارامد تشخیص داده پرت در دادگان با ابعاد

1. Concept Lattice
2. Local Outlier Mining Algorithm
3. Particle Swarm Optimization
4. Subspace Outlier Detection
5. Shared Nearest Neighbors
6. Subspace Outlier Ranking
7. Bottom-up
8. Kolmogorov-Smirnov
9. Correlation Outlier Probability
10. Principal Component Analysis
11. High Contrast Subspaces for Density-Based Outlier Ranking
12. Contextual Outlier Detection in Arbitrarily Relevant Subspaces
13. Probability Local Outlier Factor
14. Maximum-Relevance-to-Density
15. Minimum-Redundancy-Maximum-Relevance-to-Density



شکل ۱: روندنمای روش پیشنهادی تشخیص داده پرت زیرفضای مبتنی بر آنتروپی.

بالا، باید دارای مزایای زیر باشد: انتخاب ابعاد مرتبط بدون در نظر گرفتن هیچ گونه فرضی روی ابعاد و توزیع داده‌ها، امکان تشخیص چندین نوع داده پرت با انتخاب زیرفضاهای مرتبط به صورت محلی، بهبود قدرت تفکیک معیار امتیازدهی پرت محلی با استفاده از پهنای تطبیقی و در عین حال، کاهش اختلاف‌های ناچیز بین چگالی یک نقطه و همسایگانش به منظور کاهش تعداد داده‌های نرمالی که به اشتباه به عنوان داده پرت، تشخیص داده می‌شوند.

۳- روش پیشنهادی

دست‌یافتن به دقت و کارایی بالا در مسئله تشخیص داده‌های پرت روی دادگان با ابعاد بالا، یک چالش است. اکثر روش‌های موجود در تشخیص داده‌های پرت با ابعاد بالا ناکارامد هستند. در این بخش، یک روش انتخاب زیرفضای مرتبط محلی برای کاهش تعداد ابعاد پیشنهاد داده می‌شود که بر اساس میزان اطلاعات یک نقطه داده در راستای ابعاد مختلف، ویژگی‌های مرتبط برای تعیین پرت‌بودن آن نقطه داده شناسایی می‌شوند. سپس با نگاشت کل دادگان به زیرفضای به دست آمده برای نقطه داده مفروض، درجه پرت‌بودن آن نقطه داده محاسبه می‌گردد.

برای توصیف روش پیشنهادی، ابتدا برخی نمادگذاری‌های مورد استفاده معرفی می‌شود. فرض کنید که D دادگانی شامل N نقطه داده در فضای ویژگی M بعدی می‌باشد که مجموعه ویژگی‌ها به صورت $F=\{f_1,f_2,\ldots,f_M\}$ است.

روند کلی الگوریتم پیشنهادی در شکل ۱ آمده است. ابتدا نرمال‌سازی داده‌ها (یا مقیاس‌بندی ویژگی‌ها) ضروری است، زیرا ویژگی‌های با مقادیر بزرگ‌تر در محاسبه فاصله و همچنین در ساختار همسایگی بر سایر ویژگی‌ها تسلط دارند. معمولاً در تشخیص داده‌های پرت، روش نرمال‌سازی حداقل- حداکثر[1] (min-max) که هر ویژگی را به بازه [۰,۱] [۰,۱] نرمال‌سازی می‌کند استفاده می‌شود [۲۳]. نرمال‌سازی min-max، تبدیل خطی زیر را روی مقادیر ویژگی f انجام می‌دهد

$$f^* = \frac{f - \min(f)}{\max(f) - \min(f)} \qquad (۱)$$

که $\min(f)$ و $\max(f)$ به ترتیب برابر مقادیر مینیمم و ماکسیمم ویژگی f هستند و بدین ترتیب با نرمال‌سازی تمام ویژگی‌ها (ستون‌ها)، ماتریس داده $D_{N\times M}$ تبدیل به $DS_{N\times M}$ می‌شود.

در این مقاله مشابه با HiCS [۱۷] و گروه‌بندی ویژگی[2] [۲۴]، گام انتخاب زیرفضای مرتبط و امتیازدهی پرت به صورت مسئله‌های مجزا جداسازی شده که با این جداسازی می‌توان به طور عمیق روی هر یک از این مسئله‌ها پژوهش کرد. گام‌های روش پیشنهادی آمده در شکل ۱، به طور کامل در زیربخش‌های بعدی شرح داده می‌شوند.

۳-۱ تخمین چگالی هسته تطبیقی برای تشخیص داده پرت

تخمین چگالی هسته (KDE)[3] یک روش غیر پارامتری برای تخمین تابع چگالی احتمال متغیرهای تصادفی است. پهنای باند نقش مهمی در

1. Minimum-Maximum
2. Feature Bagging
3. Kernel Density Estimation

کیفیت KDE دارد. روش تخمین چگالی هسته $k-$نزدیک‌ترین همسایگی (kNN-KDE)[1]، گونه مهمی از روش تخمین چگالی هسته می‌باشد که در آن پهنای باند به صورت محلی انتخاب می‌گردد. در مسئله تشخیص داده پرت، استفاده از روش تخمین چگالی kNN-KDE دارای مزیت‌های امکان به کارگیری یک پهنای باند هسته تطبیقی، کاهش پیچیدگی محاسباتی و از دست نرفتن اختلاف‌های محلی در چگالی به دلیل استفاده از $k-$نزدیک‌ترین همسایگی به جای کل نقاط دادگان می‌باشد.

در این مقاله از روش kNN-KDE برای تخمین چگالی محلی استفاده می‌شود. در این تخمین‌گر چگالی هسته، مقدار چگالی برای نقطه داده x_i با توجه به $k-$نزدیک‌ترین همسایه‌اش به صورت زیر تعریف می‌گردد

$$KDE(x_i)=\frac{1}{k}\sum_{x_j\in kNN(x_i)}\frac{1}{h^d}K(\frac{dist(x_i,x_j)}{h}) \qquad (۲)$$

که پارامتر h پهنای باند، d تعداد ابعاد نقاط داده، K تابع هسته و $dist(x_i,x_j)$ فاصله اقلیدسی بین نقاط داده x_i و x_j می‌باشد. چگالی نقطه داده x_i با استفاده از میانگین تخمین چگالی با توجه به کل نقاط همسایگی‌اش تخمین زده می‌شود.

در داده‌های ناااریب[2]، توزیع نقاط داده در بخش‌های مختلف، متفاوت است و به کارگیری پهنای باند ثابت در تشخیص داده پرت در دادگانی شامل چندین خوشه با چگالی‌های متفاوت، منجر به عملکرد مناسبی نمی‌گردد. بنابراین منطقی نیست که یک پهنای باند ثابت برای کل نقاط دادگان به کار گرفته شود. یک روش رایج برای تعیین پهنای تطبیقی در مسایل تخمین چگالی و تشخیص داده پرت [۲۵]، به کارگیری مقادیر کوچک پهنای باند h در نواحی با چگالی بالا و مقادیر بزگ h در نواحی با چگالی پایین است. اما در مسئله تشخیص داده پرت، نواحی با چگالی بالا مورد توجه نیستند، بنابراین با در نظر گرفتن یک پهنای باند بزرگ می‌توان چگالی را هموار[3] تخمین زد و بدین ترتیب واریانس امتیاز پرت بین نقاط نرمال را کاهش داد. در نواحی با چگالی پایین، پهنای باند باریک منجر به تخمین چگالی‌های کوچک‌تر می‌شود که این می‌تواند باعث برجسته‌نمودن نقاط داده پرت گردد. بنابراین در مسئله تشخیص داده پرت بر عکس مسئله تخمین چگالی، انتخاب یک پهنای باند بزرگ در نواحی با چگالی بالا و یک پهنای باند کوچک در نواحی با چگالی پایین ترجیح داده می‌شود.

در این مقاله از یک پهنای باند متفاوت برای هر نقطه داده $x_i\in DS$ استفاده خواهد شد. فرض کنید که برای iامین نقطه داده، $d_k(x_i)$ برابر با میانگین فاصله تا $k-$نزدیک‌ترین همسایه‌اش یعنی $d_k(x_i)=(1/k)\sum_{x_j\in KNN(x_i)}d(x_i,x_j)$ است و $d_{k-\min}$ و $d_{k-\max}$ برابر کوچک‌ترین و بزرگ‌ترین مقادیر در مجموعه $\{d_k(x_i)|i=1,2,\ldots,N\}$ باشند. با الهام از پهنای باند تطبیقی ارائه‌شده در روش Adaptive-KD [۴]، از $h_i=c\times[d_{k-\max}+d_{k-\min}+\varepsilon-d_k(x_i)]$ به عنوان پهنای باند نقطه داده iام استفاده می‌شود که $c(c>0)$ عامل مقیاس‌بندی برای کنترل اثر هموارسازی سراسری و ε یک مقدار مثبت بسیار کوچک (برای مثال، 10^{-5}) برای اطمینان از غیر صفر شدن پهنای باند می‌باشد. در روش Adaptive-KD [۴]، برای هر نقطه داده به صورت تطبیقی یک پهنای باند محاسبه و بر اساس آن، چگالی نقطه داده مفروض تعیین می‌شود. در حالی که روش پیشنهادی پس از محاسبه پهنای باند یک نقطه داده بر اساس فاصله تا $k-$نزدیک‌ترین همسایگی و تعداد ابعاد زیرفضا، از این پهنای باند برای محاسبه چگالی نقاط همسایگی نیز استفاده می‌کند تا اختلاف‌های جزئی بین چگالی نقاط داده نرمال را کاهش دهد.

در اینجا تابع هسته گاوسی به کار گرفته شده است. بدین ترتیب مدل KDE مورد استفاده در این مقاله به صورت زیر می‌باشد

$$KDE(x_i)=\frac{1}{k}\sum_{x_j\in kNN(x_i)}\frac{1}{(2\pi)^{\frac{d}{2}}\times(h_i)^d}e^{\frac{-dist(x_i,x_j)^2}{2(h_i)^2}} \qquad (۳)$$

که $h_i=c\times[d_{k-\max}+d_{k-\min}+\varepsilon-d_k(x_i)]$ است.

در گام انتخاب ابعاد مرتبط، نیاز به محاسبه چگالی در راستای یک ویژگی است. همچنین در مرحله امتیازدهی به منظور محاسبه امتیاز پرت نقاط داده مختلف در زیرفضاهای محلی متناظر با تعداد ابعاد متفاوت، نیاز به محاسبه چگالی در این زیرفضاها است. فرض کنید که زیرفضای مد نظر (ویژگی lام یا زیرفضای انتخابی متناظر با نقطه داده iام) با SS نمایش داده شود. مدل KDE در (۳) برای محاسبه چگالی در زیرفضای SS به صورت زیر درمی‌آید

$$KDE(x_i^{SS})=\frac{1}{k}\sum_{x_j^{SS}\in kNN^{SS}(x_i^{SS})}\frac{1}{h_i^{|SS|}}K(\frac{dist(x_i^{SS},x_j^{SS})}{h_i^{SS}})$$
$$\text{where } K(\frac{dist(x_i^{SS},x_j^{SS})}{h_i^{SS}})=\frac{1}{(2\pi)^{\frac{|SS|}{2}}}e^{\frac{-1}{2}(\frac{dist(x_i^{SS},x_j^{SS})}{h_i^{SS}})^2} \qquad (۴)$$

که x_i^{SS} نگاشت نقطه داده iام به زیرفضای SS، $|SS|$ تعداد ابعاد زیرفضای SS، $kNN^{SS}(x_i^{SS})$ شامل $k-$نزدیک‌ترین همسایه نقطه داده x_i^{SS} در زیرفضای SS و h_i^{SS} برابر پهنای باند نقطه داده iام در زیرفضای SS می‌باشد. نحوه محاسبه پارامتر h_i^{SS} جهت تخمین چگالی نقطه داده iام در زیرفضای SS در گام‌های انتخاب ابعاد و امتیازدهی داده پرت، در زیربخش‌های ۳-۲ و ۳-۳ آورده شده است.

۳-۲ انتخاب زیرفضای مرتبط مبتنی بر آنتروپی

در تئوری اطلاعات، آنتروپی توسعه‌یافته توسط شانون یک معیار مؤثر برای اندازه‌گیری اطلاعات یا عدم قطعیت برای یک متغیر است [۲۶]. مفهوم آنتروپی برای تشخیص داده پرت بسیار شهودی است زیرا حضور داده پرت، آنتروپی دادگان را افزایش می‌دهد. همچنین نقطه داده پرت در مقایسه با همسایه‌هایش دارای چگالی کمتر و اطلاعات بیشتری است.

فرض کنید که y یک متغیر تصادفی و $S(y)$ برابر مجموعه مقادیری که متغیر y می‌تواند اختیار کند، باشد و $p(y)$ تابع احتمال متغیر تصادفی y را نمایش دهد. آنتروپی $E(Y)$ می‌تواند به صورت (۵) تعریف گردد

$$E(Y)=-\sum_{y\in S(y)}p(y)\log p(y) \qquad (۵)$$

در این مقاله، پیرو تعریف شانون از آنتروپی، آنتروپی محلی[4] تعریف و به منظور انتخاب زیرفضای مرتبط به کار گرفته می‌شود. هدف از انتخاب زیرفضای مرتبط، هرس‌کردن ابعاد بی‌ربط از طریق شناسایی ابعادی است که حاوی اطلاعات کمی هستند.

1. k-Nearest Neighbor Kernel Density Estimation
2. Skewed Data
3. Smooth
4. Local Entropy

برای اندازه‌گیری میزان اطلاعات محلی نقطه داده x_i در راستای ویژگی lام (x_i^l)، ابتدا یک دادگان محلی به صورت kNN در راستای lامین ویژگی تعریف می‌شود $(kNN^l(x_i^l))$ که آنتروپی‌اش به عنوان آنتروپی محلی اشاره می‌شود. همچنین بر اساس این همسایگی، مقدار اطلاعات محلی نقطه داده x_i در راستای ویژگی lام به صورت زیر محاسبه می‌گردد

$$I^l(x_i^l) = -\log_2(p(x_i^l)) \qquad (۶)$$

که $p(x_i^l)$ احتمال نقطه داده x_i در راستای ویژگی f_l است. برای محاسبه مقدار احتمال $p(x_i^l)$، لازم است مقدار چگالی نقطه داده x_i و نقاط همسایگی‌اش در راستای ویژگی f_l محاسبه و سپس نرمال‌سازی گردد. مقادیر چگالی هر یک از نقاط $x_r^l \in \{x_i^l \cup kNN^l(x_i^l)\}$ در راستای ویژگی lام با استفاده از محاسبه پهنای باند h_i^l طبق (۷) و سپس با جایگذاری در (۴) تعیین می‌شوند

$$h_i^l = c \times [d_{k-\min}^l + d_{k-\max}^l + \varepsilon - d_k^l(x_i^l)] \qquad (۷)$$

که $d_k^l(x_i^l)$ برابر میانگین فاصله x_i^l تا $k-$نزدیک‌ترین همسایه‌اش در راستای ویژگی lام و $d_{k-\min}^l$ و $d_{k-\max}^l$ به ترتیب برابر کوچک‌ترین و بزرگ‌ترین مقادیر در مجموعه $\{d_k^l(x_i^l) | i = 1,2,\ldots,N\}$ هستند. بنابراین با نرمال‌سازی چگالی $KDE(x_i^l)$ نسبت به مجموع چگالی نقاط همسایگی‌اش، احتمال $p(x_i^l)$ به نقطه داده x_i^l تخصیص داده می‌شود

$$p(x_i^l) = \frac{KDE(x_i^l)}{\sum_{x_j^l \in kNN^l(x_i^l)} KDE(x_j^l)} \qquad (۸)$$

آنتروپی محلی نقطه داده x_i^l در راستای ویژگی f_l می‌تواند به صورت زیر تعریف گردد

$$\begin{aligned} LE^l(x_i^l) &= \sum_{x_j^l \in kNN^l(x_i^l)} p(x_j^l) I^l(p(x_j^l)) \\ &= -\sum_{x_j^l \in kNN^l(x_i^l)} p(x_j^l) \log_2(p(x_j^l)) \end{aligned} \qquad (۹)$$

که $kNN^l(x_i^l)$ بیانگر $k-$نزدیک‌ترین همسایه به نقطه داده x_i^l در راستای lامین ویژگی است.

بر طبق این شهود که نقطه داده پرت در مقایسه با همسایه‌هایش دارای مقدار احتمال کوچک‌تر و حاوی اطلاعات بیشتری است، می‌توان ابعاد مرتبط را انتخاب کرد. ویژگی f_l برای تشخیص پرت‌بودن/ نبودن نقطه داده x_i مرتبط است، اگر مقدار اطلاعات نقطه داده x_i^l نسبت به مقدار میانگین اطلاعات نقاط همسایگی‌اش (آنتروپی محلی) در راستای ویژگی f_l بیشتر باشد

$$I^l(x_i^l) > LE^l(x_i^l) \qquad (۱۰)$$

بعد از محاسبه مقدار اطلاعات و آنتروپی محلی x_i در طول تمام ویژگی‌ها، می‌توان زیرفضای مرتبط با تشخیص داده پرت در این نقطه داده را به صورت زیر انتخاب کرد

$$RS_i = \{f_l \,|\, I^l(x_i^l) > LE^l(x_i^l)\} \qquad (۱۱)$$

در الگوریتم ۱، فرایند انتخاب زیرفضای مرتبط مبتنی بر آنتروپی آورده شده است (شکل ۲).

۳-۳ نگاشت داده‌ها و امتیازدهی داده‌های پرت

در داده‌های با ابعاد بالا ممکن است با توجه به تعداد زیاد ویژگی‌های بی‌ربط، برخی داده‌های پرت گم شوند. در این مقاله با پیشنهاد یک روش انتخاب زیرفضای مرتبط محلی برای تشخیص داده‌های پرت سعی شده که اکثر داده‌های پرت شناسایی شوند. پس از این که برای هر نقطه داده، یک زیرفضای مرتبط استخراج شد، لازم است که کارایی این زیرفضاهای محلی در تشخیص داده‌های پرت ارزیابی گردد.

برای امتیازدهی داده‌های پرت، هر نقطه داده $x_i \in DS$ در فضای اولیه با ابعاد بالا به زیرفضای مرتبط متناظرش با ابعاد پایین‌تر نگاشت می‌شود $(x_i^{RS_i})$. با به کارگیری تعریف ویژگی مرتبط در (۱۰)، می‌توان نقاط داده‌ای را که در راستای کل ابعاد، حاوی اطلاعات کمتری در مقایسه با همسایه‌ها هستند (یعنی در طول هیچ یک از ویژگی‌ها در ناحیه تنک قرار ندارند) نرمال در نظر گرفت و امتیاز پرت صفر به آنها تخصیص داد. در اینجا برای تعیین میزان پرت‌بودن نقاط دادگان $x_i \in DS$، یک معیار نسبی به کار گرفته می‌شود تا میزان انحراف چگالی نقطه داده $x_i^{RS_i}$ از نقاط همسایگی‌اش در زیرفضای RS_i را اندازه‌گیری کند. برای تعیین نقاط همسایگی $x_i^{RS_i}$ در زیرفضای مرتبط متناظرش $(kNN^{RS_i}(x_i^{RS_i}))$ لازم است کل دادگان اولیه DS به زیرفضای RS_i نگاشت شود تا نقاط همسایگی $x_i^{RS_i}$ در فضای نگاشت‌یافته تعیین گردند.

وابستگی شدیدی بین چگالی یک نقطه داده و تعداد ابعاد زیرفضای مرتبط محلی‌اش وجود دارد. با افزایش تعداد ابعاد زیرفضای مرتبط، چگالی نقاط داده کاهش می‌یابند و یک پهنای باند هسته تطبیقی پیشنهاد داده می‌شود که تعداد ابعاد زیرفضا را نیز در نظر می‌گیرد. همچنین به منظور کاهش واریانس چگالی نقاط داده نرمال و در نتیجه، کاهش نرخ مثبت کاذب، در مرحله امتیازدهی نقطه داده iام در زیرفضای RS_i از پهنای باند تطبیقی استفاده شده است. بدین صورت که پهنای باند نقطه داده iام $(h_i^{RS_i})$ به صورت تطبیقی بر حسب فاصله تا نقاط همسایگی‌اش و تعداد ابعاد زیرفضای RS_i با توجه به نسخه چندبعدی قانون اسکات[1] [۲۷] تعیین می‌گردد. سپس از همین پهنای باند $h_i^{RS_i}$ برای محاسبه چگالی نقاط همسایگی استفاده می‌شود. با در نظر گرفتن یک پهنای باند یکسان برای نقاط همسایگی، انتظار می‌رود اختلاف‌های جزئی بین مقدار چگالی نقطه $x_i^{RS_i}$ و چگالی نقاط همسایگی‌اش از بین رود. بنابراین چگالی نقطه داده $x_i^{RS_i}$ و نقاط داده همسایگی‌اش در زیرفضای RS_i $(KDE(x_r^{RS_i}) | x_r^{RS_i} \in \{x_i^{RS_i} \cup kNN^{RS_i}(x_i^{RS_i})\})$ با استفاده از پهنای باند تطبیقی $h_i^{RS_i}$ به دست آمده از (۱۲) و سپس با جایگذاری در (۴)، محاسبه می‌شوند

$$\begin{aligned} h &= h_i^{RS_i} \\ &= c \times [d_{k-\max}^{RS_i} + d_{k-\min}^{RS_i} + \varepsilon - d_k^{RS_i}(x_i^{RS_i})] \times N^{\frac{-1}{|RS_i|+4}} \end{aligned} \qquad (۱۲)$$

که $|RS_i|$ تعداد ابعاد زیرفضای مرتبط RS_i، $d_k^{RS_i}$ برابر میانگین فاصله نقطه داده $x_i^{RS_i}$ تا $k-$نزدیک‌ترین همسایه‌اش در زیرفضای RS_i و $d_{k-\min}^{RS_i}$ و $d_{k-\max}^{RS_i}$ به ترتیب برابر کوچک‌ترین و بزرگ‌ترین مقادیر در مجموعه $\{d_k^{RS_i}(x_i^{RS_i}) | i = 1,2,\ldots,N\}$ می‌باشند.

با فرض این که اندازه دادگان N ثابت باشد، پهنای باند تطبیقی $h_i^{RS_i}$ تابعی افزایشی بر حسب تعداد ابعاد است. بدین ترتیب با افزایش تعداد ابعاد، پهنای باند نیز افزایش می‌یابد و چگالی نقاط در زیرفضاهای با ابعاد مختلف قابل مقایسه خواهند بود.

پس از تخمین چگالی نقطه داده $x_i^{RS_i}$ و همسایگانش در زیرفضای

1. Scott's Rule

Algorithm 1: Local Entropy-based Subspace Selection for Outlier Detection(LESS)

Input: $DS,\ k, FS$

Output: $RS = \{RS_i\}_{i=1}^{N}$

Initialization: Set $RS_i = \{\}, i = 1, 2, ..., N$

for each f^l in FS:

- $DS^l = l^{th}$ column of dataset matrix DS

for each x_i^l in DS^l:

- $kNN^l(x_i^l)$ = Compute kNN for data point x_i^l in DS^l
- $d_k^l(x_i^l)$ =Compute the average Euclidean distance x_i^l to its k nearest neighbors

end

- Compute $d_{k-\min}^l$ and $d_{k-\max}^l$ from all the quantities $d_k^l(x_i^l),\ i = 1, 2, ..., N$

end

for each x_i in DS:

for each f_l in FS:

- $DS^l = l^{th}$ column of dataset matrix DS
- Compute the kernel width of the ith data point: $h_i^l = c * [d_{k-\min}^l + d_{k-\max}^l + \varepsilon - d_k^l(x_i^l)]$
- Compute the local density of the ith data point: $KDE(x_i^l) = \frac{1}{k} \sum_{x_j^l \in kNN^l(x_i^l)} \frac{1}{2\pi * h_i^l} e^{\frac{-dist(x_i^l, x_j^l)^2}{2(h_i^l)^2}}$
- Compute the sum of local density for k nearest neighbors of the ith data point using the kernel width of the ith data point:

$$\sum_{x_j^l \in kNN^l(x_i^l)} KDE(x_j^l) = \sum_{x_j^l \in kNN^l(x_i^l)} \frac{1}{k} \sum_{x_t^l \in kNN^l(x_j^l)} \frac{1}{2\pi * h_i^l} e^{\frac{-dist(x_j^l, x_t^l)^2}{2(h_i^l)^2}}$$

- Compute the probability of the ith data point and its kNN on feature f^l by means of the normalization of their local density:

$$\forall x_r^l \in \{x_i^l \cup x_j^l \in kNN^l(x_i^l)\} : p(x_r^l) = \frac{KDE(x_r^l)}{\sum_{x_u^l \in kNN^l(x_r^l)} KDE(x_u^l)}$$

- Compute the information provided by the ith data on feature f^l: $I^l(x_i^l) = -\log_2(p(x_i^l))$
- Compute the local entropy of the ith data point on feature f^l: $LE^l(x_i^l) = -\sum_{x_j^l \in kNN^l(x_i^l)} p(x_j^l) \log_2(p(x_j^l))$

if $I^l(x_i^l) > LE^l(x_i^l)$:

- Add f_l to RS_i:

end

end

end

شکل ۲: الگوریتم ۱.

RS_i، معیار عامل پرت مبتنی بر چگالی نسبی نقطه داده iام به صورت زیر تعریف می‌شود

$$RDOF_i = RDOF(x_i^{RS_i}) = \frac{\overline{KDE(kNN^{RS_i}(x_i^{RS_i}))}}{KDE(x_i^{RS_i})}$$

$$\text{where } \overline{KDE(kNN^{RS_i}(x_i^{RS_i}))} = \frac{\sum_{x_j^{RS_i} \in kNN^{RS_i}(x_i^{RS_i})} KDE(x_j^{RS_i})}{k} \quad (۱۳)$$

که $RDOF$ برابر نرخ میانگین چگالی نقاط همسایگی $x_i^{RS_i}$ به چگالی نقطه داده $x_i^{RS_i}$ است. اگر $RDOF(x_i^{RS_i})$ بزرگ‌تر از ۱ باشد، آن گاه نقطه داده $x_i^{RS_i}$ خارج از یک ناحیه چگال قرار دارد و بنابراین یک داده پرت می‌باشد. اگر $RDOF(x_i^{RS_i})$ کوچک‌تر یا مساوی با ۱ باشد، آن گاه نقطه داده $x_i^{RS_i}$ توسط یک ناحیه همسایگی با چگالی یکسان یا توسط ناحیه تنک‌تر احاطه شده که یک داده پرت نیست. الگوریتم نگاشت و امتیازدهی داده‌های پرت در الگوریتم ۲ توصیف شده است (شکل ۳).

در زیربخش‌های بالا، گام‌های الگوریتم با تمرکز بر انتخاب زیرفضای مرتبط و امتیازدهی داده پرت شرح داده شد. حال در الگوریتم ۳، این مراحل در کنار هم قرار گرفته و شبه‌کد کلی روش پیشنهادی برای تشخیص داده‌های پرت با ابعاد بالا آورده شده است (شکل ۴).

سرانجام به طور خلاصه روی پیچیدگی محاسباتی روش پیشنهادی بحث می‌شود. در الگوریتم پیشنهادی، زیرفضای مرتبط نقطه داده x_i بر اساس آنتروپی و اطلاعات محلی محاسبه شده و سپس میزان پرت‌بودن این نقطه داده در زیرفضای مرتبط بر اساس مقایسه چگالی محلی‌اش با چگالی نقاط داده همسایگی‌اش تعیین می‌گردد. در الگوریتم ۱ بر اساس نقاط داده یک‌بعدی در راستای یک ویژگی، مرتبط‌بودن آن ویژگی به نقاط داده مختلف بررسی می‌گردد که محاسبات عمده در الگوریتم ۱ مربوط به تعیین نقاط داده همسایگی است، به طوری که پیچیدگی محاسباتی یافتن kNN برای یک نقطه داده در راستای یک ویژگی برابر

Algorithm 2: Local Density-based Outlier Scoring (LDOS)

Input: DS, RS, k, c, ε

Output: $RDOF = \{RDOF_i\}_{i=1}^{N}$

---

for each x_i in DS :

- DS^{RS_i} =project data matrix DS into RS_i
- $kNN^{RS_i}(x_i^{RS_i})$ =compute kNN for data point x^{RS_i} in the subspace RS_i
- Compute the local density of the data point $x_i^{RS_i}$ and its neighbors $kNN^{RS_i}(x_i^{RS_i})$ as follows:

$\forall x_r^{RS_i} \in \{x_i^{RS_i} \cup kNN^{RS_i}(x_i^{RS_i})\}$:

$$KDE(x_r^{RS_i}) = \frac{1}{k} \sum_{x_t^{RS_i} \in kNN^{RS_i}(x_r^{RS_i})} \frac{1}{h^{|RS_i|}} K\left(\frac{dist(x_r^{RS_i}, x_t^{RS_i})}{h_i^{RS_i}}\right)$$

where $K\left(\frac{dist(x_r^{RS_i}, x_t^{RS_i})}{h_i^{RS_i}}\right) = \frac{1}{(2\pi)^{|RS_i|/2}} e^{\frac{-1}{2}\left(\frac{dist(x_r^{RS_i}, x_t^{RS_i})}{h_i^{RS_i}}\right)^2}$ and $h = h_i^{RS_i} = c * [d_{k-\max}^{RS_i} + d_{k-\min}^{RS_i} + \varepsilon - d_k^{RS_i}(x_i^{RS_i})] * N^{\frac{-1}{|RS_i|+4}}$

- Compute the Relative Density-based Outlier Factor(RDOF) for data point $x_i^{RS_i}$:

if $|RS_i| = 0$:

$RDOF_i = 0$

else

$$RDOF_i = RDOF(x_i^{RS_i}) = \frac{\overline{KDE(kNN^{RS_i}(x_i^{RS_i}))}}{KDE(x_i^{RS_i})}, \text{ where } \overline{KDE(kNN^{RS_i}(x_i^{RS_i}))} = \frac{\sum_{x_j^{RS_i} \in kNN^{RS_i}(x_i^{RS_i})} KDE(x_j^{RS_i})}{k}$$

end

end

شکل ۳: الگوریتم ۲.

Algorithm 3: Local Entropy-based Subspace Outlier Detection (LESOD)
Input: DS, k
Output: Outlier data points

- DS = Applying feature normalization/scaling on D
- RS =Select the local relevant subspaces for outlier detection using Algorithm 1 : $LESS(DS, FS, k)$
- $RDOF$ =Compute outlier scores for all data points in DS using Algorithm 2: $LDOS(RS,\ RS,\ k,\ c,\ \varepsilon)$

Detect outlier data points given as $RDOF$ score values in the previous step

شکل ۴: الگوریتم ۳.

جدول ۱: مشخصات دادگان مورد استفاده از مخزن UCI.

دادگان	تعداد نمونه‌ها	تعداد نمونه‌های پرت	تعداد ویژگی‌ها
Arrhythmia	۴۲۰	۱۸	۱۲۹
Ionosphere	۳۵۱	۱۲۶	۳۲
Breast_diagnostic	۵۶۹	۲۱۲	۳۰
Diabetes	۷۶۸	۲۶۸	۸
Glass	۲۱۴	۹	۷

$O(N)$ است، البته با استفاده از ساختار درختی می‌تواند برای کل نقاط دادگان به $O(N \times \log N)$ کاهش یابد. بدین ترتیب پیچیدگی زمانی الگوریتم ۱، $O(M \times N \times \log N)$ است. در الگوریتم ۲ برای امتیازدهی یک نقطه داده، نیاز به نگاشت دادگان به زیرفضای مرتبط متناظر با آن نقطه داده، یافتن نقاط همسایگی‌اش و سپس امتیازدهی نقطه داده مفروض است که پیچیدگی محاسباتی این عملیات از مرتبه $O(N \times N)$ است. بنابراین پیچیدگی محاسباتی کلی روش تشخیص داده پرت آمده در الگوریتم ۳ برابر $O(M \times N \times \log N + N^2)$ است. پیچیدگی محاسباتی روش پیشنهادی با برخی تکنیک‌های محلی تشخیص داده پرت با ابعاد بالا SOD [۱۵]، COP [۱۹] و OUTRES [۱۶] با پیچیدگی به ترتیب برابر با $O(M.N^2)$، $O(N.\log N.M^2)$ و $O(N.(2^M.M))$ قابل مقایسه است. پیچیدگی محاسباتی روش پیشنهادی بر حسب تعداد ابعاد، خطی است در حالی که پیچیدگی الگوریتم‌های COP و OUTRES به ترتیب چندجمله‌ای و نمایی است.

٤- ارزیابی روش پیشنهادی

در این بخش، عملکرد روش پیشنهادی از طریق آزمایش‌هایی روی چندین دادگان واقعی مخزن یادگیری ماشین UCI [۲۸] ارزیابی و با روش‌های LOF [۱۰]، HiCS [۱۷]، COP [۱۹] و Adaptive-KD [۴] مقایسه می‌شود.

٤-۱ روش راه‌اندازی

الگوریتم پیشنهادی تشخیص داده پرت در داده‌های با ابعاد بالا به زبان پایتون روی سیستمی با مشخصات Intel® Core™ i۷ CPU و حافظه ۸ GB پیاده‌سازی شده است.

دادگان: آزمایش‌ها روی چندین دادگان واقعی متعلق به مخزن یادگیری ماشین UCI [۲۸] شامل دادگان Arrhythmia، Breast_diagnostic، Ionosphere، Diabetes و Glass که مشخصات آنها در جدول ۱ آمده، انجام شده است.

معیار ارزیابی: سطح زیر منحنی ROC (AUC)[1] یک معیار شناخته‌شده برای ارزیابی عملکرد روش‌های تشخیص داده پرت است. نمودار ROC نرخ مثبت صحیح (TPR)[2] در مقابل نرخ مثبت کاذب

1. Area Under Curve
2. True Positive Rate

جدول ۲: مقایسه AUC (بر حسب ٪) روی دادگان واقعی UCI.

دادگان/الگوریتم	LOF	HiCS	COP	Adaptive-KD	LESOD
Arrhythmia	۶۴٫۹۸	۶۰٫۳۵	۶۰٫۴۹	۶۶٫۰۶	**۷۲٫۵۳**
Ionosphere	۸۶٫۴۳	۶۹٫۹۹	۷۶٫۲۶	۹۱٫۰۴	**۹۳٫۰۳**
Breast_diagnostic	۵۳٫۱۹	۶۴٫۰۳	۵۰٫۶۱	۶۱٫۸	**۷۰٫٤**
Diabetes	۶۱٫۹۵	۵۸٫۶۹	۵۴٫۱۶	۶۱٫۸۱	**٦٧٫۰۹**
Glass	۸۱٫۶۱	۸۳٫۹۵	۷۵٫۳۳	۸۶٫۵	**۸۷٫۹٦**

[1] (FPR) با در نظر گرفتن حد آستانه‌های مختلف، ترسیم می‌گردد که TPR و FPR به صورت زیر محاسبه می‌شوند

$$TPR = \frac{TP}{P}$$
$$FPR = \frac{FP}{N} \qquad (۱۴)$$

که TP، P، FP و N به ترتیب بیانگر تعداد نقاط داده‌ای که به درستی پرت تشخیص داده شده‌اند، تعداد کل نقاط داده پرت موجود در دادگان، تعداد نقاط داده‌ای که به اشتباه پرت تشخیص داده شده‌اند و تعداد کل نقاط نرمال در دادگان است. برای محاسبه سطح زیر منحنی ROC روش‌های ذوزنقه‌ای خطی[2]، ذوزنقه‌ای لگاریتمی[3] و ذوزنقه‌ای خطی-لگاریتمی[4] وجود دارد. روش ذوزنقه‌ای خطی از درون‌یابی خطی[5] بین دو نقطه نمودار ROC برای محاسبه AUC استفاده می‌کند. برای یک بازه زمانی مشخص $(t_۱ - t_۲)$، AUC می‌تواند به صورت زیر محاسبه گردد

$$AUC = \frac{۱}{۲}(c_۱ + c_۲)(t_۲ - t_۱) \qquad (۱۵)$$

که در آن $c_۱$ و $c_۲$ به ترتیب برابر مقادیر منحنی ROC در زمان‌های $t_۱$ و $t_۲$ هستند.

در آزمایش‌ها، روش پیشنهادی با چهار الگوریتم شناخته‌شده تشخیص داده پرت LOF [۱۰]، HiCS [۱۷]، COP [۱۹] و Adaptive-KD [۴] با استفاده از AUC مقایسه گردیده است.

٤-۲ نتایج

پارامترهای روش HiCS مطابق پیشنهاد نویسندگان [۱۷] به صورت $\alpha = ۰٫۱$، $M = ۵۰$ و $candidate\_cutoff = ۱۰۰$ تنظیم شده است. پارامتر تعداد نزدیک‌ترین همسایه (k) در هر سه روش LOF، HiCS و Adaptive-KD مشابه با برخی کارهای تشخیص داده پرت قبلی [۱۴] و [۱۸] برابر $\sqrt{N}$ تنظیم شده است. در COP، پارامتر k باید به اندازه کافی بزرگ و مقدارش از تعداد ابعاد دادگان بزرگ‌تر باشد تا قادر به تخمین ساختار کواریانس محلی باشد. بنابراین از $k = ۵۰$ روی تمام دادگان استفاده شده است، به جز دادگان Arrhythmia که k برابر ۲۰۰ تنظیم گردیده است. نتایج این سه روش روی هر یک از دادگان آمده در جدول ۱ بر حسب مقدار AUC در جدول ۲ گزارش شده است. در این مقاله، از پیاده‌سازی‌های انجام‌شده الگوریتم‌های LOF [۱۰]، HiCS [۱۷] و COP [۱۹] در چارچوب ELKI [۲۹] استفاده گردیده و الگوریتم Adaptive-KD [۴] به زبان پایتون پیاده‌سازی شده است.

در این آزمایش‌ها از هسته گاوسی استفاده شده که برای تعیین پهنای باند تطبیقی، نیاز به تنظیم سه پارامتر ε، c و k دارد. پارامترهای c و ε به ترتیب برابر با ۰٫۵ و $۱۰^{-۵}$ و پارامتر تعداد نزدیک‌ترین همسایه k برای هر دو گام تعیین انتخاب زیرفضای مرتبط و امتیازدهی داده پرت برابر $\sqrt{N}$ تنظیم شده است. با تنظیم این پارامترها، مقادیر AUC به دست آمده از روش پیشنهادی روی هر یک از دادگان آمده در جدول ۱ در جدول ۲ آورده شده و همچنین بهترین مقدار AUC در بین الگوریتم‌های LOF، HiCS، COP، Adaptive-KD و روش پیشنهادی روی هر یک از دادگان، برجسته شده است.

همان طور که در جدول ۲ پیداست، روش پیشنهادی LESOD روی تمام دادگان آمده در جدول ۱، توانسته بهترین نتایج را در بین الگوریتم پیشنهادی، LOF [۱۰]، HiCS [۱۷]، COP [۱۹] و Adaptive-KD [۴] به دست آورد. روش‌های LOF و Adaptive-KD، نقاط داده پرت را در فضای کل ابعاد جستجو می‌کنند و بنابراین به دلیل مسئله طلسم بعد ممکن است برخی داده‌های پرت مخفی شوند و تشخیص داده نشوند.

در جدول ۲ مشاهده می‌شود که دقت الگوریتم پیشنهادی LESOD بالاتر از الگوریتم COP است. زیرفضای مرتبط در الگوریتم COP با استفاده از همبستگی خطی پیرسون به دست می‌آید و این روش همبستگی تنها مشخصات توزیع خطی دادگان محلی را منعکس می‌کند. بدین ترتیب، زیرفضای مرتبط در COP، تنها می‌تواند داده‌های پرتی که از توزیع خطی منحرف هستند را شناسایی کند. بنابراین توزیع دادگان و طلسم بعد روی دقت تشخیص روش COP تأثیرگذار است. این در حالی است که روش پیشنهادی ما برای تعیین زیرفضای مرتبط تفاوت بین ابعاد با مشخصات مختلف را بر حسب آنتروپی محلی و بدون در نظر گرفتن هیچ فرضی روی توزیع داده‌ها محاسبه می‌کند و در نتیجه، توزیع داده‌ها تأثیر کمتری روی دقت تشخیص داده پرت در روش پیشنهادی می‌گذارد. در روش HiCS [۱۷] بر اساس یک روش سراسری، چندین زیرفضای مرتبط انتخاب می‌گردد. همچنین زیرفضاهای مرتبط از طریق یک الگوریتم Apriori-like جستجو می‌شوند که ممکن است زیرفضاهای مرتبط انتخاب‌شده توسط این الگوریتم جستجو، ناقص باشند. علاوه بر این، روش HiCS [۱۷] با فرض استقلال بین ویژگی‌ها، کنتراست یک زیرفضا را با اندازه‌گیری همبستگی بین ابعادش اندازه‌گیری کرده است. در کل، فرض استقلال بین ابعاد صحیح نیست یعنی ممکن است انتخاب زیرفضایی با ابعاد وابسته، بیانگر ماهیت واقعی داده‌های پرت در هر دادگانی نباشد. با توجه به دلایل اشاره‌شده، روش ما با به کارگیری یک رویکرد انتخاب محلی ابعاد مرتبط و بدون در نظر گرفتن هیچ فرضی روی ویژگی‌ها و توزیع داده‌ها می‌تواند ویژگی‌های مرتبط برای تشخیص پرت‌بودن هر یک از نقاط داده را تعیین کند و نسبت به روش HiCS [۱۷] بهبود چشم‌گیری داشته باشد. البته قابل اشاره است که پهنای باند تطبیقی پیشنهادی نیز بر روی عملکرد مناسب الگوریتم امتیازدهی داده

1. False Positive Rate
2. Linear Trapezoidal Method
3. Logarithmic Trapezoidal Method
4. Linear-Log Trapezoidal Method
5. Linear Interpolation

پرت و نهایتاً روی بهبود دقت الگوریتم تشخیص داده پرت پیشنهادی تأثیرگذار است.

۵- نتیجه‌گیری و کارهای آتی

تشخیص داده پرت در فضای داده با ابعاد بالا، چالش‌برانگیز و هزینه‌بر است. در این خصوص، انتخاب زیرفضای مرتبط نقش مهمی در دقت و کارایی مسئله تشخیص داده پرت بازی می‌کند. در این مطالعه، یک روش بدون نظارت تشخیص داده پرت محلی مبتنی بر زیرفضا در داده‌های با ابعاد بالا پیشنهاد شده است. در الگوریتم پیشنهادی، زیرفضای مرتبط برای هر نقطه داده، بر اساس آنتروپی محلی و مقدار اطلاعات آن نقطه داده در دادگان محلی که با استفاده از تخمین چگالی تطبیقی محاسبه شده‌اند، تعریف می‌گردد. زیرفضای مرتبط محلی یک نقطه داده، متشکل از ابعادی است که مقدار اطلاعات آن نقطه داده در راستای آن ابعاد بزرگ‌تر از آنتروپی محلی باشد و بر این اساس، از تأثیر ویژگی‌های بی‌ربط روی تشخیص داده پرت می‌تواند به طور مؤثری جلوگیری کند و منجر به تشخیص داده‌های پرت پنهان‌شده در زیرفضاهای با ابعاد پایین‌تر گردد. همچنین در این مقاله یک روش امتیازدهی داده پرت مبتنی بر تخمین چگالی هسته پیشنهاد داده شده که برای کاهش اختلاف امتیاز پرت بین نقاط داده نرمال و برجسته‌نمودن نقاط داده پرت، از یک پهنای باند تطبیقی که رابطه معکوس با فاصله تا $k-$نزدیک‌ترین همسایگی دارد استفاده می‌کند. باید اشاره کرد که پهنای باند تطبیقی محاسبه‌شده برای یک نقطه داده، جهت تخمین چگالی نقاط داده همسایگی نیز استفاده می‌شود تا اختلاف‌های جزئی بین مقدار چگالی و امتیاز پرت نقاط داده نرمال از بین رود. با استفاده از دادگان واقعی، نتایج تجربی تأیید می‌کنند که روش پیشنهادی ما برای تشخیص داده پرت در دادگان با ابعاد بالا کارامد است.

در الگوریتم پیشنهادی برای انتخاب ابعاد مرتبط از معیار کلی آنتروپی استفاده شده که در کارهای آتی می‌توان از فاصله کولبک- لیبلر[1] برای اندازه‌گیری اختلاف توزیع داده‌های پرت با توزیع داده‌های نرمال استفاده کرد. در الگوریتم ما، زیرفضای مرتبط برای یک نقطه داده بر اساس مشخصات توزیع دادگان محلی‌اش تعیین می‌شود و بنابراین انتخاب مناسب نقاط داده در دادگان محلی روی کارایی انتخاب زیرفضای مرتبط تأثیرگذار است. در پژوهش‌های آتی می‌توان به تحلیل هم‌زمان دو زیرمسئله انتخاب دادگان محلی و انتخاب زیرفضای مرتبط پرداخت. به علاوه، با توجه به رشد روزافزون داده‌ها می‌توان الگوریتم پیشنهادی تشخیص داده پرت در زیرفضای مرتبط را در محیط‌های محاسباتی توزیع‌شده و موازی بسط داد تا امکان به کارگیری این روش روی داده‌های عظیم و با ابعاد بالا فراهم گردد.

مراجع

[1] C. C. Aggarwal and S. Y. Philip, "An effective and efficient algorithm for high-dimensional outlier detection," *The VLDB J.*, vol. 14, no. 2, pp. 211-221, Apr. 2005.

[2] M. Riahi-Madvar, B. Nasersharif, and A. Akbari Azirani, "A new density-based subspace selection method using mutual information for high dimensional outlier detection," *Knowledge-Based Systems*, vol. 216, Article ID: 106733, 16 Mar. 2021.

[3] M. Riahi-Madvar, B. Nasershari, and A. Akbari Azirani, "Subspace outlier detection in high dimensional data using ensemble of PCA-based subspaces," in *Proc. 26th Int. Computer Conf., Computer Society of Iran, CSICC'21*, 5 pp., Tehran, Iran, 3-4 Mar. 2021.

[4] J. Zhang, J. Liu, and R. Karim, "Adaptive kernel density-based anomaly detection for nonlinear systems," *Knowledge-Based Systems*, vol. 139, pp. 50-63, 1 Jan. 2018.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Outlier detection: a survey," *ACM Computing Surveys*, vol. 14, p. 15, Aug. 2007.

[6] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd Ed., Wiely, 1994.

[7] H. V. Nguyen, V. Gopalkrishnan, and I. Assent, "An unbiased distance-based outlier detection approach for high-dimensional data," in *Proc. of the Int. Conf. on Database Systems for Advanced Applications,* pp. 138-152, Taipei, Taiwan, 11-14 Apr. 2011.

[8] E. M. Knox and R. T. Ng, "Algorithms for mining distancebased outliers in large datasets," in *Proc. of the Int. Conf. on Very Large Data Bases,* pp. 392-403, New York, NY, USA, 24-27 Aug. 1998.

[9] F. Angiulli and C. Pizzuti, "Outlier mining in large high-dimensional data sets," *IEEE Trans. on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 203-215, Jan. 2005.

[10] M. M. Breunig, *et al.*, LOF: identifying density-based local outliers, ACM sigmod record, ACM, 2000.

[11] A. Zimek, E. Schubert, and H. P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining: the ASA Data Science J.*, vol. 5, no. 5, pp. 363-387, Oct. 2012.

[12] C. C. Aggarwal and P. S. Yu, Outlier detection for high dimensional data, in ACM Sigmod Record, ACM, 2001.

[13] J. Zhang, *et al.*, "A concept lattice based outlier mining method in low-dimensional subspaces," *Pattern Recognition Letters*, vol. 30, no. 15, pp. 1434-1439, Nov. 2009.

[14] X. Zhao, J. Zhang, and X. Qin, "LOMA: a local outlier mining algorithm based on attribute relevance analysis," *Expert Systems with Applications*, vol. 84, pp. 272-280, Oct. 2017.

[15] H. P. Kriegel, *et al.*, "Outlier detection in axis-parallel subspaces of high dimensional data," in *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining,* pp. 831-838, Bangkok, Thailand, 27-30 Apr. 2009.

[16] E. Muller, M. Schiffer, and T. Seidl, "Statistical selection of relevant subspace projections for outlier ranking," in *Proc. IEEE 27th Int. Conf. on Data Engineering*, pp. 434-445, Hannover, Germany, 11-16 Apr. 2011.

[17] F. Keller, E. Muller, and K. Bohm, "HiCS: high contrast subspaces for density-based outlier ranking," in *Proc. IEEE 28th Int. Conf. on Data Engineering*, pp. 1037-1048, Arlington, VA, USA, 1-5 Apr. 2012.

[18] J. Zhang, *et al.*, "A relevant subspace based contextual outlier mining algorithm," *Knowledge-Based Systems*, vol. 99, pp. 1-9, May 2016.

[19] H. P. Kriegel, *et al.*, "Outlier detection in arbitrarily oriented subspaces," in *Proc. IEEE 12th Int. Conf. on Data Mining*, pp. 379-388, Brussels, Belgium, 10-13 Dec. 2012.

[20] F. Cheraghchi, A. Iranzad, and B. Raahemi, "Subspace selection in high-dimensional big data using genetic algorithm in apache spark," in *Proc. of the 2nd Int. Conf. on Internet of Things, Data and Cloud Computing, ACM*, Article ID: 54, 7 pp., Cambridge, United Kingdom, 22-23 Mar. 2017.

[21] C. C. Aggarwal, *Outlier Analysis, in Data Mining*, Springer, 2015.

[22] H. V. Nguyen, *et al.*, "CMI: an information-theoretic contrast measure for enhancing subspace cluster and outlier detection," in *Proc. of the SIAM Int. Conf. on Data Mining*, 9 pp., Austin, TX,USA, 2-4 May 2013.

[23] S. Kandanaarachchi, *et al.*, "On normalization and algorithm selection for unsupervised outlier detection," *Data Mining and Knowledge Discovery*, vol. 34, no. 2, pp. 309-354, Mar. 2020.

[24] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in *Proc. of the 11th ACM SIGKDD Int. Conf. on Knowledge Discovery in Data Mining*, pp. 157-166, Chicago, IL, USA, 21-24 Aug. 2005.

[25] F. Liu, *et al.*, "Scalable KDE-based top-n local outlier detection over large-scale data streams," *Knowledge-Based Systems*, vol. 204, Article ID: 106186, Sept. 2020.

[26] C. E. Shannon, "A mathematical theory of communication," *the Bell System Technical J.*, vol. 27, no. 3, pp. 379-423, Jul. 1948.

[27] D. W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley & Sons, 2015.

[28] K. Bache and M. Lichman, UCI machine learning repository, 2013.

[29] E. Achtert, H. P. Kriegel, and A. Zimek, "ELKI: a software system for evaluation of subspace clustering algorithms," in *Proc. Int. Conf. on Scientific and Statistical Database Management,* pp. 580-585, Hong Kong, 9-11 Jul. 2008.

1. Kullback-Leibler

محبوبه ریاحی مدوار تحصیلات خود را در مقطع کارشناسی مهندسی کامپیوتر در سال ۱۳۹۰ از دانشگاه صنعتی شریف و کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی در سال ۱۳۹۲ از دانشگاه صنعتی امیرکبیر به پایان رسانده است. از سال ۱۳۹۲ تا ۱۳۹۴ نام‌برده در دانشگاه شهید باهنر کرمان و ولیعصر (عج) رفسنجان مشغول تدریس بود. پس از آن، در سال ۱۳۹۴ به دوره دکتری مهندسی کامپیوتر گرایش هوش مصنوعی در دانشگاه علم و صنعت ایران وارد گردید و هم‌اکنون نیز در حال تحصیل است. زمینه‌های علمی مورد علاقه ایشان عبارتند از: داده‌کاوی، بازشناسی الگو، تشخیص داده‌های پرت، داده‌های با ابعاد بالا.

احمد اکبری ازیرانی دانشیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران هستند. ایشان ۲۵ سال سابقه تدریس و پژوهش در زمینه های مختلف رشته مهندسی کامپیوتر را در این دانشکده را دارند و دبیر قطب علمی شبکه‌های ارتباطی و اطلاعاتی نسل جدید می‌باشند و بیش از ۵۰ مقاله در ژورنال های معتبر بین المللی انتشار داده‌اند. ایشان مسئولیت‌های علمی و اجرایی مختلفی از جمله ریاست دانشکده به مدت ۷ سال و عضویت در هیات مدیره انجمن کامپیوتر ایران به مدت ۱۶ سال را در کارنامه خود دارند. زمینه‌های پژوهشی مورد علاقه ایشان پردازش داده‌ها، شبکه‌های کامپیوتری و امنیت شبکه است.

بابک ناصر شریف درجه کارشناسی را در رشته مهندسی کامپیوتر گرایش سخت‌افزار از دانشگاه صنعتی امیرکبیر در سال ۱۳۷۶ دریافت نمود و موفق به اخذ درجه کارشناسی ارشد و دکتری در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه علم و صنعت ایران به‌ترتیب در سال‌های ۱۳۷۹ و ۱۳۸۶ گردید. نام‌برده از سال ۱۳۸۶ تا ۱۳۹۰ عضو هیأت علمی گروه مهندسی کامپیوتر در دانشکده فنی گیلان و ازسال ۱۳۹۰ تاکنون عضو هیأت علمی دانشکده مهندسی کامپیوتر در دانشگاه صنعتی خواجه نصیر طوسی است. زمینه تحقیقاتی ایشان پردازش گفتار، یادگیری عمیق برای پردازش گفتار و بازشناسی الگو است.