ششمین همایش بین‌المللی افق های نویـن در
مهندسی برق، کامپیوتـر و مکانیک
6th International Conference on the New Horizons in
Electrical Engineering, Computer and Mechanical
www.mhconf.ir

مروری بر روشهای تشخیص داده‌های پرت در داده کاوی

مهسان فریدونیان[۱]، حسن مطلبی[۲]

[۱]دانشگاه تحصیلات تکمیلی و فناوریهای پیشرفته کرمان m.fereydoonian@student.kgut.ac.ir

[۲] دانشگاه تحصیلات تکمیلی و فناوریهای پیشرفته کرمان h.motallebi@kgut.ac.ir

چکیده

یکی از مراحل مهم در آنالیز و تحلیل داده‌ها، شناسایی داده‌های پرت است. داده‌های پرت نمونه داده‌هایی هستند که با رفتار عمومی یا مدل داده منطبق نباشند. این تفاوت به حدی است که به نظر می‌رسد این داده‌ها توسط توزیع یا مکانیزم متفاوتی از سایر داده‌ها تولید شده‌اند. داده‌های پرت به سه دسته‌ی، نقطه‌ای، مشروط و گروهی تقسیم بندی می‌شوند. در این‌مقاله به کاربردهای شناسایی داده‌های پرت در زمینه های مختلفی همچون تشخیص نفوذ، تشخیص تقلب، تشخیص آسیب صنعتی، پردازش تصویر و غیره اشاره شده است. روش‌های شناسایی داده‌های پرت نیز به چند دسته تقسیم بندی می‌شوند که انواع این‌روش ها همچون روش‌های مبتنی بر دسته بندی، روش‌های مبتنی بر نزدیکترین همسایه و روش‌های مبتنی بر خوشه بندی در این مقاله ارائه شده است.

واژه‌های کلیدی

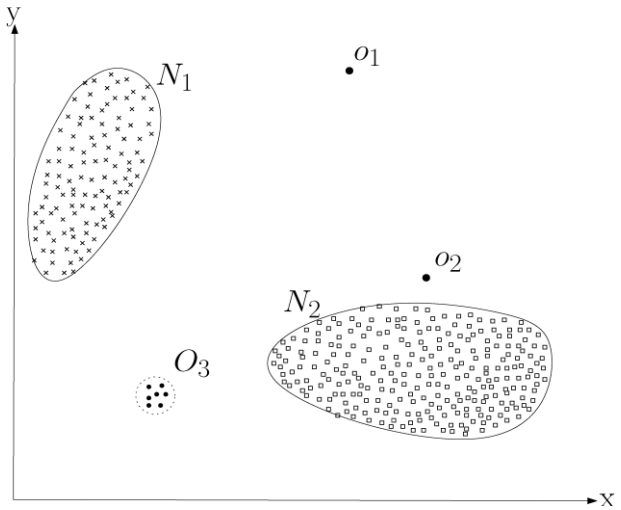
داده پرت، داده نرمال، ناهنجاری

۱. معرفی

تشخیص داده‌ی پرت به معنای جستجوی الگوهایی در داده‌ها است که این الگوها تفاوت چشمگیری با سایر داده‌ها دارند، طوری که انگار توسط توزیع یا مکانیزم دیگری تولید شده‌اند [۱]. روش‌های متعددی برای تشخیص داده‌های پرت در زمینه‌های تشخیص نفوذ در شبکه، تشخیص تقلب کارت اعتباری، حوزه سلامت عمومی و پزشکی، تشخیص خرابی در واحد مکانیکی، پردازش تصویر و غیره وجود دارد که نیاز به بررسی و مطالعه در این‌زمینه را آشکارتر می‌نماید. ظهور فناوری‌های مختلف، فرصت‌هایی برای کشف بهتر داده‌های پرت فراهم ساخته است. در این بین، نگرش داده کاوی به عنوان روشی مؤثر وکارآمد مطرح می‌شود. تکنیک‌های داده کاوی، امکان شناسایی الگوهای رفتاری را در بین حجم زیادی از داده‌ها فراهم می‌آورد که این امر امکان شناسایی و کشف الگوهای مشکوک و یا حتی تقلب را میسر می‌سازد. یکی از دلایل اصلی استفاده از داده کاوی، تجزیه و تحلیل مؤثر و کارآمد مجموعه ای از مشاهدات مختلف با توجه به رفتار آنها است. داده کاوی برای تشخیص نقاط پرت از تکنیک‌ها و روش‌ها و الگوریتم‌های گوناگونی مثل روش‌های مبتنی برتوزیع، روش‌های مبتنی برفاصله، روش‌های مبتنی برخوشه بندی و روش‌های مبتنی بر چگالی و غیره استفاده می‌کند. این نکته را باید در نظر گرفت که نقاط پرت، نقاطی هستندکه متفاوت‌اند از نویزها. نویز در حقیقت تصادفی در داده‌هاست که باید قبل از تشخیص داده پرت حذف شود [۲].

۱.۱ تعریف داده پرت

داده‌های پرت در حقیقت الگوهایی در مجموعه داده هستند که با رفتار عادی و نرمال مورد انتظار از داده‌ها مطابق نیستند. همان طور که در شکل نشان داده شده است بیشتر داده‌ها در دو ناحیه نرمال N2،N1 قرار دارند. نقاطی مثل O2،O1 که به اندازه کافی از این دو ناحیه دور هستند، داده پرت نامیده می‌شوند.



شکل۱.مجموعه ای از دادههای نرمال و پرت

۲.۱ چالش‌ها در کشف داده‌های پرت

مدل سازی نمونه داده‌های نرمال و پرت باید به صورت کارآمد باشد. کیفیت تشخیص نمونه داده‌های پرت به شدت بستگی به مدل سازی نمونه داده‌های نرمال و پرت دارد. ساخت یک مدل کامل برای نمونه داده‌های نرمال اگر ممکن باشد با چالش‌های بسیاری روبرو است، چرا که شمارش کلیه رفتارهای نرمال ممکن در یک مجموعه داده کار دشواری است. اغلب مرز میان نرمال بودن و غیر نرمال بودن داده‌ها مشخص نیست. در عوض ممکن است یک محدوده وسیعی از ناحیه خاکستری وجود داشته باشد. بدین ترتیب در حالی که برخی

از روش‌های تشخیص داده‌های پرت تنها یکی از دو برچسب نرمال و پرت را برای هر نمونه داده در مجموعه داده‌ها در نظر می‌گیرند برخی دیگر انتساب درجه پرت بودن به نمونه داده را استفاده می‌کنند.

تشخیص داده‌های پرت برای کاربردهای خاص. انتخاب مقیاس‌های تشابه یا فاصله و مدلی که داده‌ها را توصیف میکند برای تشخیص داده‌های پرت امری مهم و حساس است. اغلب چنین انتخابی بستگی به ماهیت داده‌ها دارد. وابستگی شدید تشخیص داده‌های پرت به نوع کاربرد آن باعث شده است که توسعه یک روش کاربردی جامع امکان پذیر نباشد. در عوض باید روش‌هایی را توسعه داد که هریک به کاربردهای خاص تخصیص داده شود. برای مثال در تحلیل داده‌های بیمارستانی یک انحراف کوچک برای توجیه پرت بودن یک داده کافی است اما در بورس برای توجیه پرت بودن یک داده نیاز به انحراف چشمگیری داریم.

کنترل نویز در فرآیند تشخیص داده‌های پرت. همان طور که پیش‌تر اشاره شد داده‌های پرت متفاوت از نویزها هستند. اکثر داده-هایی که با آنها سر و کار داریم از کیفیت خوبی برخوردار نیستند. اغلب وجود نویزها در مجموعه داده‌ها اجتناب‌ناپذیر است. نویزها ممکن است به عنوان انحرافاتی در مقادیر صفت‌خاصه یا به عنوان مقادیر ناموجود ظاهر شوند. کیفیت پایین داده‌ها و حضور نویزها چالش بزرگی در فرآیند تشخیص داده‌های پرت ایجاد می‌کنند. این دو موضوع باعث می‌شود داده‌ها از شکل واقعی خود خارج شده و تمایز میان داده-های نرمال و داده‌های پرت نامشخص به نظر آید. به علاوه ممکن است نویز و داده‌های ناموجود باعث پنهان ماندن داده‌های پرت شوند و کارآیی فرآیند تشخیص داده پرت را کاهش دهند. گاهی اوقات ممکن است یک داده پرت به عنوان نویز شناسایی شود و بالعکس یک نویز به عنوان داده پرت شناسایی و تلقی شود.

قابل فهم بودن. در برخی از کاربردها ممکن است کاربر نه تنها به دنبال تشخیص داده‌های پرت باشد بلکه به دنبال این باشد که اشیایی که به عنوان داده پرت شناسایی شده اند چرا داده‌ی پرت هستند؟ به این منظور روش تشخیص داده‌های پرت باید توجیهی برای شناسایی داده‌های پرت خود داشته باشد.

مسأله تشخیص داده پرت به واسطه این چالش‌ها به سادگی قابل حل نیست. در حقیقت بیشتر تکنیک‌های تشخیص داده پرت در یک زمینه خاصی کاربرد دارند و حالت عمومی ندارند. برای تشخیص اینکه بفهمیم چه تکنیکی را باید استفاده کنیم به فاکتورهای گوناگونی مثل ماهیت داده‌ها، در دسترس بودن یا نبودن برچسب هر داده، نوع داده پرتی که باید تشخیص داده شود، محدودیت‌ها و نیازمندی‌های دامنه برنامه کاربردی مورد نظر و غیره نیاز داریم.

۳.۱ مروری بر کارهای موجود

تشخیص ناهنجاری موضوع تعدادی از بررسی‌ها و مقالات مروری و همچنین کتاب‌ها بوده است. در [۳] بررسی گسترده ای از تکنیک‌های تشخیص ناهنجاری توسعه یافته در یادگیری ماشین و حوزه‌های آماری ارائه می‌شود. در [۴] بررسی گسترده ای از تکنیک‌های تشخیص ناهنجاری برای داده‌های عددی و همچنین نمادین ارائه شده است. در [۵] بررسی گسترده ای از تکنیک‌های تشخیص ناهنجاری با استفاده از شبکه‌های عصبی و رویکردهای آماری ارایه شده است. [۶,۷] بررسی تکنیک‌های تشخیص ناهنجاری که به طور خاص برای تشخیص نفوذ سایبری استفاده می‌شود را ارائه می‌دهند.

۲. جنبه‌های مختلف تکنیک‌های تشخیص ناهنجاری

این بخش جنبه‌های مختلف تکنیک‌های تشخیص ناهنجاری را شناسایی و مورد بحث قرار می‌دهد.

۱.۲ ماهیت داده‌ها

یکی از کلیدی ترین جنبه‌های تکنیک‌های تشخیص داده پرت ماهیت داده‌های ورودی است. داده‌ی ورودی ما به طور کلی مجموعه از نمونه داده‌ها است. که این نمونه داده‌ها می‌توانند شی، رکورد، نقطه، بردار، الگو، رویداد، مشاهدات و یا موجودیت باشند [۸]. هر نمونه داده به وسیله یک سری ویژگی‌ها توصیف می‌شود که ممکن است یک ویژگی باشد که به آن تک متغیره گویند یا چند ویژگی که به آن چند متغیره گویند. ماهیت این ویژگی‌ها باید با تکنیکی که برای تشخیص داده پرت استفاده می‌کنیم منطبق باشد. برای مثال برای تکنیک‌هایی که بر پایه نزدیکترین همسایه هستند ماهیت ویژگی نمونه داده‌های ورودی باید به گونه‌ای باشد که فاصله مورد نیاز این تکنیک‌ها را تعیین کند.

۲.۲ انواع داده پرت

۱.۲.۲ داده‌های پرت عمومی

در یک مجموعه داده یک نمونه داده را داده پرت عمومی یا سراسری می‌گوییم اگر به طور چشمگیری نسبت به سایر نقاط محرف شده باشد. داده‌های پرت عمومی ساده ترین نوع داده‌های پرت می‌باشند. چالشی که در این نوع داده‌ها وجود دارد، انتخاب بهترین معیاراندازه گیری ممکن برای تشخیص انحراف می‌باشد. به عنوان مثال شبکه‌های کامپیوتری را در نظر می‌گیریم که فرض می‌کنیم مجموعه ای از کامپیوترها باهم در ارتباط بوده و برای هم بسته ارسال می‌کنند، حال یکی از این کامپیوترها به طور غیرعادی و خیلی متفاوت از الگوی نرمال ارسال بسته، شروع به ارسال بسته‌ها می‌کند، این کامپیوتر به عنوان یک داده پرت عمومی شناخته می‌شود.

۲.۲.۲ داده‌های پرت مشروط

در مجموعه داده یک نمونه داده را داده پرت مشروط می‌گوییم اگر با توجه به شرایط تعریف شده در مسأله این نمونه داده به طور چشگیری نسبت به سایر نمونه‌ها منحرف شده باشد. برای نمونه دمای ۲۵درجه را در نظر می‌گیریم. آیا این دما یک داده پرت است؟ جواب این‌است که بستگی به مکان و زمان دارد. اگر مکان ماکشور آفریقا و فصل تابستان باشد این دما داده پرت محسوب می‌شود در غیر این‌صورت اگر در همان کشور و فصل زمستان باشد این دما یک نقطه پرت محسوب نمی‌شود. به طور کلی در تشخیص داده‌های پرت مشروط، ویژگی و صفت داده‌های مورد نظر به دو گروه تقسیم می‌شود:

۱-ویژگی‌های متنی: این ویژگی‌ها زمینه داده مورد نظر را مشخص می‌کنند. مثلا در مثال بالا ویژگی‌های متنی ممکن است تاریخ و یا موقعیت باشد.

۲-ویژگی‌های رفتاری: این ویژگی‌ها برای ارزیابی کردن این موضوع است که آیا داده مورد نظر با توجه به زمینه‌ای که دارد نقطه پرت است یا نه. (منظور با توجه به ویژگی متنی‌ای که دارد.) مثلا در مثال ذکر شده می‌توان ویژگی‌های رفتاری را دما، فشار، رطوبت در نظر گرفت وگفت که آیا این دما باتوجه به ویژگی متنی آن، که کشور آفریقا است و زمان آن فصل تابستان است آیا نقطه پرت محسوب می-شود یا خیر.

۳.۲.۲ داده‌های پرت گروهی

در یک مجموعه داده، هنگامی یک زیرمجموعه از مجموعه داده‌ها، داده پرت گروهی یا جمعی را تشکیل می‌دهند که رفتارهمه اشیاء موجود در این‌زیر مجموعه به صورت چشمگیری با کل مجموعه داده متفاوت باشد. نکته‌ای که در اینجا وجود دارد این‌که ممکن است هر یک از داده‌های این زیرمجموعه به صورت مجزا داده پرت نباشد. برای مثال ارسال محموله را در نظر بگیریم، اگر در ارسال یک محموله

تاخیر داشته باشیم نمی‌توان این‌تاخیر را به عنوان داده پرت در نظر گرفت، حال اگر تعداد تاخیرها به صد تا رسید اینجاست که صدتا تاخیر یک داده پرت گروهی را تشکیل می‌دهند.

۳.۲ برچسب‌های داده

برچسب مربوط به نمونه‌داده مشخص می‌کند که آیا این‌داده نرمال یا پرت است. باید در نظر داشت که به دست آوردن برچسب مربوط به داده‌ها که هم دقیق باشد و هم نشان دهنده انواع رفتارهای داده باشد بسیار هزینه بر است. عمل برچسب گذاری اغلب به صورت دستی توسط یک نیروی انسانی متخصص انجام می‌شود. از این رو تلاش قابل توجهی برای بدست آوردن مجموعه داده‌های آموزشی برچسب گذاری شده نیاز است. معمولا به دست آوردن مجموعه ای از برچسب‌ها از نمونه داده‌های پرت که نشان دهنده همه رفتارهای غیرنرمال آن‌ها است بسیار دشوارتر از بدست آوردن مجموعه‌ای از برچسب‌ها از داده‌های نرمال است. بر اساس میزان در دسترس بودن برچسب‌ها، تکنیک‌های تشخیص داده‌های پرت می‌توانند در یکی از سه حالت زیر عمل کنند:

۱.۳.۲ روش‌های نظارت شده

روش‌های نظارت شده نرمال بودن و نرمال نبودن داده‌ها را مدل سازی می‌کنند. افراد کارشناس و خبره در حوزه کاری مربوطه، داده‌ها را بررسی و برخی از آنها را برچسب‌گذاری می‌کنند. در برخی کاربردها ممکن است تنها نمونه داده‌های نرمال برچسب‌گذاری شوند و در این صورت نمونه داده‌هایی که با مدل نمونه داده‌های نرمال مطابقت ندارند به عنوان نمونه داده‌های پرت گزارش می‌شوند. در برخی روش‌ها نمونه داده‌های پرت مدل سازی می‌شوند و نمونه داده‌هایی که با مدل نمونه داده‌های پرت مطابقت نداشته باشند به عنوان نمونه داده‌های نرمال شناخته می‌شوند.

۲.۳.۲ روش‌های غیرنظارتی

در برخی از کاربردها برچسب نمونه داده‌ها موجود نیست، یعنی نمونه داده‌ها برچسب نرمال بودن یا پرت بودن را ندارند. در این‌حالت باید از یک روش یادگیری نظارت نشده استفاده شود. روش‌های نظارت نشده‌ی تشخیص داده‌های پرت، دارای یک فرض ضمنی هستند: نمونه داده‌های نرمال تا اندازه‌ای که بشود، خوشه بندی می‌شوند. به عبارت دیگر یک روش نظارت نشده برای تشخیص داده‌های پرت انتظار دارد که نمونه داده‌های نرمال از یک الگویی پیروی کنند که تعداد تکرارهای این الگو بسیار بیشتر از داده‌های پرت است. اشیاء می‌توانند چندین گروه با ویژگی‌های متمایز را تشکیل دهند. یک داده پرت در فضایی متفاوت از گروه نمونه داده‌های نرمال قرار می‌گیرد.

۳.۳.۲ روش‌های نیمه نظارتی

در بسیاری از کاربردها اگرچه برخی از نمونه داده‌ها دارای برچسب هستند، اما تعداد این نمونه داده‌ها اغلب کم است. ممکن است با مواردی روبه رو شویم که تنها مجموعه کوچکی از نمونه داده‌های نرمال و یا نمونه داده‌های پرت دارای برچسب باشند و اکثر نمونه داده‌ها بدون برچسب باشند. روش‌های نیمه‌نظارت شده برای تشخیص داده‌های پرت در چنین شرایطی توسعه داده شده‌اند. برای نمونه وقتی بعضی نمونه داده‌های نرمال دارای برچسب هستند می‌توانیم از آنها و نمونه داده‌های بدون برچسبی که بسیار شبیهشان است برای آموزش مدل جهت نمونه داده‌های نرمال استفاده کنیم. از این مدل می‌توان برای شناسایی داده‌های پرت نیز استفاده کرد، نمونه داده‌هایی که با این مدل مشابهتی نداشته باشند به عنوان داده پرت شناسنایی می‌شوند. اگر تنها برخی از داده‌های پرت دارای برچسب باشند، به نظر می‌رسد روش‌های نیمه نظارت شده برای تشخیص داده‌های پرت، گول زننده و غیرقابل اعتماد باشد. بعید است که بتوان با تعداد کمی از داده‌های پرت برچسب دار، مدلی ساخت که موثر و کارآمد باشد و بتوان به کمک آن بقیه داده‌های پرت را تشخیص داد. در روش‌های نیمه نظارت شده بهترین کار ساخت و آموزش مدل از روی داده‌های نرمال است.

۴.۲ خروجی تکنیک‌های تشخیص داده پرت

یکی از جنبه مهم برای هر تکنیک تشخیص داده‌ی پرت، نحوه گزارش این داده‌ها است. به طور معمول، خروجی‌های تولید شده توسط تکنیک‌های تشخیص داده‌های پرت یکی از دو نوع زیر است:

۱.۴.۲ تخصیص امتیاز

تکنیک‌های امتیازدهی بسته به درجه‌ای که آن نمونه به عنوان داده پرت در نظر گرفته می‌شود، به هر نمونه در داده‌های آزمایشی، یک امتیاز پرت بودن اختصاص می‌دهند. بنابراین خروجی چنین تکنیک‌هایی یک لیست رتبه‌بندی شده از داده‌های پرت است.

۲.۴.۲ تخصیص برچسب

تکنیک‌های این دسته به نمونه داده‌های آزمایشی برچسب نرمال یا پرت اختصاص می‌دهند.

۵.۲ کاربردهای تشخیص داده پرت

۱.۵.۲ تشخیص نفوذ

تشخیص نفوذ به شناسایی فعالیت‌های مخرب در یک سیستم مرتبط با رایانه اشاره دارد. این فعالیت‌های مخرب یا نفوذها از منظر امنیت کامپیوتر جالب هستند. یک نفوذ رفتاری متفاوت با رفتار عادی سیستم دارد. از این رو تکنیک‌های تشخیص داده پرت در حوزه تشخیص نفوذ قابل اجرا هستند. چالش کلیدی برای تشخیص داده پرت در این حوزه، حجم عظیم داده است. تکنیک‌های تشخیص داده پرت باید از نظر محاسباتی کارآمد باشند تا این حجم عظیم داده‌های ورودی را مدیریت کنند. علاوه بر این، داده‌ها معمولاً به صورت جریانی ارائه می‌شوند، بنابراین نیاز به تجزیه و تحلیل آنلاین است. مسئله دیگری که به دلیل اندازه ورودی بزرگ ورودی به وجود می‌آید نرخ هشدار کاذب است. از آنجایی که حجم داده‌ها به میلیون‌ها نمونه داده می‌رسد، درصد کمی از هشدارهای کاذب می‌تواند تجزیه و تحلیل را برای یک تحلیل‌گر طاقت فرسا کند. داده‌های برچسب دار مربوط به رفتار نرمال معمولاً در دسترس هستند، در حالی که برچسب هایی برای نفوذ که همان داده‌های پرت است وجود ندارد. بنابراین، تکنیک‌های تشخیص داده پرت نیمه نظارت شده و نظارت نشده در این حوزه ترجیح داده می‌شوند.

۲.۵.۲ تشخیص تقلب (کلاهبرداری)

تشخیص کلاهبرداری به کشف فعالیت‌های مجرمانه‌ای اطلاق می‌شود که در سازمان‌های تجاری مانند بانک‌ها، شرکت‌های کارت اعتباری، آژانس‌های‌بیمه، شرکت‌های‌تلفن همراه، بازار سهام و غیره اتفاق می‌افتد. کاربران مخرب ممکن است مشتریان واقعی آن‌ها باشند، یا ممکن است خود را به عنوان مشتری معرفی کنند. تقلب زمانی رخ می‌دهد که این کاربران منابع ارائه شده توسط سازمان مربوطه را به صورت غیرمجاز مصرف کنند. سازمان‌ها علاقه مند به شناسایی فوری چنین موارد تقلب برای جلوگیری از ضررهای اقتصادی هستند. مقاله [۹] اصطلاح نظارت بر فعالیت را به عنوان یک رویکرد کلی برای تشخیص تقلب معرفی می‌کند. رویکرد معمول تکنیک‌های تشخیص داده پرت، حفظ و نگهداری مشخصات استفاده برای هر مشتری و نظارت بر پروفایل‌ها برای تشخیص هرگونه انحراف است.

۳.۵.۲ تشخیص داده‌های پرت پزشکی و بهداشت عمومی

تشخیص داده پرت در حوزه‌های پزشکی و بهداشت عمومی معمولاً با سوابق بیمار کار می‌کند. نمونه داده‌ها می‌توانند به دلایل مختلفی مانند وضعیت غیرعادی بیمار، خطاهای ابزارهای مورد استفاده، یا خطاهای ضبط، داده پرت داشته باشند. بنابراین تشخیص داده پرت یک

مشکل بسیاربحرانی در این‌حوزه است و به دقت بالایی نیاز دارد. داده‌ها معمولاً شامل رکوردهایی هستند که ممکن است دارای چندین ویژگی مختلف باشند، مانند سن بیمار، گروه خونی و وزن. هدف بیشتر تکنیک‌های تشخیص داده پرت در این حوزه، شناسایی رکوردهای غیرعادی (داده‌های پرت نقطه‌ای) است. معمولاً داده‌های برچسب‌گذاری‌شده متعلق به بیماران سالم است، از این رو بیشتر تکنیک‌ها یک رویکرد نیمه نظارتی را اتخاذ می‌کنند. در مقاله [۱۰] چندین تکنیک تشخیص داده پرت برکشف شیوع بیماری در یک منطقه خاص متمرکز شده‌اند. تکنیک‌های تشخیص داده پرت تجمعی برای تشخیص ناهنجاری‌ها در داده‌های سری زمانی (برای مثال داده‌های مربوط به ثبت ضربان قلب) استفاده می‌شود [۱۱]. چالش برانگیزترین مشکل تشخیص داده پرت در این حوزه این است که تشخیص یک داده-پرت به عنوان داده نرمال می‌تواند بسیارگران تمام شود.

۴.۵.۲ تشخیص آسیب صنعتی

واحدهای صنعتی به دلیل استفاده مداوم و فرسودگی معمولا آسیب می‌بینند. برای جلوگیری از تشدید و تلفات بیشتر، چنین آسیبی باید زود تشخیص داده شود. داده‌های این حوزه معمولاً به عنوان داده‌های حسگر شناخته می‌شوند، زیرا با استفاده از حسگرهای مختلف ثبت می‌شوند و برای تجزیه و تحلیل جمع‌آوری می‌شوند [۱۲,۱۳].

۵.۵.۲ پردازش تصویر

تکنیک‌های تشخیص داده پرت که با تصاویر سروکار دارند به هرگونه تغییر در یک تصویر در طول زمان و یا به نقاطی که در تصویر ثابت، غیرعادی به نظر می‌رسند علاقه دارند. این حوزه شامل تجزیه و تحلیل تصاویر ماهواره‌ای [۱۴]، تشخیص عدد [۱۵]، طیف سنجی [۱۶] و تجزیه و تحلیل تصاویر ماموگرافی [۱۷] است. داده‌های پرت در اثر حرکت یا درج یک شی خارجی یا خطاهای ابزار دقیق ایجاد می‌شوند. یکی از چالش‌های کلیدی در این حوزه، اندازه بزرگ ورودی‌ها است. هنگام برخورد با داده‌های ویدئویی، تکنیک‌های تشخیص داده پرت آنلاین مورد نیاز است.

۶.۵.۲ تشخیص داده پرت در داده‌های متنی

تکنیک‌های تشخیص داده پرت در این‌حوزه عمدتاً موضوعات یا رویدادها یا اخبار جدید را در مجموعه‌ای از اسناد یا مقالات خبری شناسایی می‌کنند. این داده‌های پرت به دلیل یک رویداد جدید یا یک موضوع غیرنرمال ایجاد می‌شوند. داده‌های این حوزه معمولاً ابعادی بالا دارند و بسیار پراکنده هستند. داده‌ها همچنین جنبه زمانی دارند زیرا اسناد در طول زمان جمع آوری می‌شوند. یک چالش مهم برای تکنیک‌های تشخیص داده پرت در این‌حوزه، رسیدگی و مدیریت کردن تغییرات زیاد در اسناد متعلق به یک دسته یا موضوع است.

۶.۲ تکنیک‌های تشخیص داده پرت

۱.۶.۲ تکنیک‌های تشخیص داده پرت مبتنی بردسته بندی

تکنیک دسته بندی برای یادگیری یک مدل (دسته بند) از مجموعه‌ای از نمونه داده‌های برچسب گذاری‌شده (داده‌های آموزشی) استفاده می‌کند و سپس برای دسته بندی یک نمونه داده آزمایشی و اختصاص دادن آن به یکی از کلاس‌های نرمال یا غیرنرمال (داده پرت)، از مدل آموخته شده استفاده می‌کند. بر اساس برچسب‌های موجود برای فاز آموزش، تکنیک‌های تشخیص داده پرت مبتنی بر دسته بندی را می‌توان به دو دسته کلی دسته‌بندی کرد: تکنیک‌های تشخیص داده پرت چندکلاسه و تک کلاسه. تکنیک‌های تشخیص داده پرت مبتنی بر دسته‌بندی چندکلاسه فرض می‌کنند که داده‌های آموزشیه حاوی نمونه‌های برچسب‌گذاری شده متعلق به چندین کلاس نرمال است. چنین تکنیک‌های تشخیص داده پرت، به یک دسته بندی کننده می‌آموزد که بین هر کلاس نرمال و بقیه کلاس‌ها تمایز قائل شود. تکنیک‌های تشخیص داده پرت مبتنی بر دسته بندی تک کلاسه فرض می‌کنند که همه نمونه‌های آموزشی فقط یک برچسب کلاس

دارند. چنین تکنیک‌هایی، یک مرز متمایزکننده در اطراف نمونه‌های نرمال را با استفاده از یک الگوریتم دسته بندی تک کلاسه،ایجاد می‌کنند. [۱۸,۱۹]. انواع تکنیک‌های تشخیص داده پرت که از الگوریتم‌های دسته بندی مختلف برای ساخت دسته بندها استفاده می‌کنند:

۱- تکنیک‌های تشخیص داده پرت مبتنی بر شبکه‌های عصبی [۲۰].

۲- تکنیک‌های تشخیص داده پرت مبتنی بر شبکه‌های بیزی،

۳- تکنیک‌های تشخیص داده پرت مبتنی بر ماشین‌های بردار پشتیبانی [۲۱]

۴- تکنیک‌های تشخیص داده پرت مبتنی بر قانون

۲.۶.۲ تکنیک‌های تشخیص داده پرت مبتنی بر نزدیکترین همسایه

مفهوم تجزیه و تحلیل نزدیکترین همسایه در چندین تکنیک تشخیص داده پرت استفاده شده است. چنین تکنیک‌هایی بر اساس این فرض کلیدی است: نمونه داده‌های نرمال در محله‌های متراکم رخ می‌دهند، در حالی که داده‌های پرت دور از نزدیک‌ترین همسایه‌ها رخ می‌دهند. تکنیک‌های تشخیص داده پرت مبتنی بر نزدیک‌ترین همسایه نیاز به اندازه‌گیری فاصله یا مقیاس تشابه تعریف شده بین دو نمونه داده را دارد. فاصله (یا شباهت) بین دو نمونه داده را می‌توان به روش‌های مختلف محاسبه کرد. برای صفات پیوسته، فاصله اقلیدسی یک انتخاب رایج است، اما می‌توان از معیارهای دیگری هم استفاده کرد. تکنیک‌های تشخیص داده پرت مبتنی بر نزدیکترین همسایه را می‌توان به طور کلی به دو دسته تقسیم کرد [۲۲,۲۳,۲۴].

۱-تکنیک‌هایی که از فاصله یک نمونه داده تا k امین نزدیکترین همسایه خود به عنوان امتیاز پرت بودن استفاده می‌کنند.

۲-تکنیک‌هایی که چگالی نسبی هر نمونه داده را برای بدست آوردن امتیاز پرت بودن آن محاسبه می‌کنند.

در این تکنیک‌ها یک امتیاز پرت بودن به هر نمونه داده به عنوان فاکتور پرت بودن محلی یا به اختصار(LOF) اختصاص داده می‌شود. برای هر نمونه داده معین، امتیاز LOF برابر با نسبت میانگین چگالی محلی k نزدیکترین همسایه نمونه داده و چگالی محلی خود نمونه داده است. برای یافتن چگالی محلی یک نمونه داده، ابتدا شعاع کوچکترین هایپرکره را در مرکز نمونه داده، که حاوی k نزدیکترین همسایه آن است پیدا می‌کنند. سپس چگالی محلی با تقسیم k بر حجم این‌هایپرکره محاسبه می‌شود. برای نمونه داده نرمال که در یک منطقه متراکم قرار دارد، چگالی محلی(نسبی) آن شبیه به همسایگانش خواهدبود، در حالی که برای یک نمونه غیرنرمال، چگالی محلی آن کمتر از نزدیکترین همسایگانش خواهدبود. از این رو نمونه غیرنرمال امتیاز LOF بالاتری خواهدداشت. در زیر نحوه محاسبه LOF نشان داده شده‌است.

$$\mathrm{RD}(X_i, X_j) = \max\left(K-\ \text{distance}\ (X_j)\,,\ \text{distance}\ (X_i, X_j)\right) \quad (۱)$$

$$LRD_k(A) = \frac{1}{\sum_{Xj\ \in N_k(A)} \frac{\mathrm{RD}(A,Xj)}{\|N_k(A)\|}} \quad (۲)$$

$$LOF_k(A) = \frac{\sum_{Xj \in N_k(A)} LRD_k(Xj)}{||N_k(A)||} \times \frac{1}{LRD_k(A)}$$

(۳)

۳.۶.۲ تکنیک‌های تشخیص داده پرت مبتنی بر خوشه بندی

خوشه بندی برای گروه بندی نمونه داده‌های مشابه استفاده می‌شود. خوشه بندی در درجه اول یک تکنیک بدون نظارت است، اگرچه خوشه بندی نیمه نظارت شده نیز اخیرا مورد بررسی قرارگرفته است. آن‌طور که به نظر می‌رسد خوشه بندی و تشخیص داده پرت اساساً با یکدیگر متفاوت هستند، اما چندین تکنیک تشخیص داده پرت مبتنی بر خوشه ایجاد شده است. تکنیک‌های تشخیص داده پرت مبتنی بر خوشه بندی را می‌توان به سه دسته، دسته‌بندی کرد [۲۵,۲۶]. دسته اول تکنیک‌های تشخیص داده پرت مبتنی بر خوشه بندی بر این فرض متکی است: نمونه داده‌های نرمال متعلق به یک خوشه در داده‌ها هستند، در حالی که داده‌های پرت به هیچ خوشه‌ای تعلق ندارند. دسته دوم تکنیک‌های تشخیص داده پرت مبتنی بر خوشه بندی بر این فرض متکی است: نمونه داده‌های نرمال اطراف نزدیک ترین مرکز خوشه‌ی نزدیک به خود قرار دارند، در حالی که داده‌های پرت از نزدیک ترین مرکز خوشه خود دور هستند. دسته سوم تکنیک‌های تشخیص داده پرت مبتنی بر خوشه بندی بر این فرض متکی است: نمونه داده‌های نرمال، متعلق به خوشه‌های بزرگ و متراکم هستند، در حالی که داده‌های پرت به خوشه‌های کوچک یا پراکنده تعلق دارند.

جمع بندی

در این مقاله مروری بر ادبیات داده‌های پرت انجام شد، و انواع داده‌های پرت و کاربردهای شناسایی داده‌های پرت در زمینه‌های مختلف و انواع تکنیک تشخیص داده‌های پرت ارائه شد.

منابع

[1] Edgeworth, F. Y. 1887. On discordant observations. Philosoph. Mag. 23, 5, 364–375.

[2] Kumar, V. 2005. Parallel and distributed computing for cybersecurity. IEEE Distrib. Syst. Online 6,10.

[3] Teng, H, Chen, K, and Lu, S. 1990. Adaptive real-time anomaly detection using inductively generated sequential patterns. In Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy. IEEE Computer Society Press, 278–284.

[4] Agyemang, M, Barker, K, and Alhajj, R. 2006. A comprehensive survey of numeric and symbolic outlier mining techniques. Intel. Data Anal. 10, 6, 521–538.

[5] Marko, M. and Singh, S. 2003a. Novelty detection: A review-part 1: Statistical approaches. Sig. Proc. 83, 12, 2481–2497.

[6] Patcha, A. and Park, J.-M. 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Comput. Netw. 51, 12, 3448–3470.

[7] Snyder, D. 2001. Online intrusion detection using sequences of system calls. M.S. thesis, Department of Computer Science, Florida State University.

[8] Tan, P.-N., Steinbach, M., and Kumar, V. 2005. Introduction to Data Mining. Addison-Wesley.

[9] Fawcett, T.and Provost, F. 1999. Activity monitoring: noticing interesting changes in behavior. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 53–62.

[10] Wong, W.-K., Moore, A., Cooper, G., and Wagner, M. 2003. Bayesian network anomaly pattern detection for disease outbreaks. In Proceedings of the 20th International Conference on Machine Learning. AAAI Press, 808–815.

[11] Lin, J., Keogh, E., Fu, A., and Herle, H. V. 2005. Approximations to magic: Finding unusual medical time series. In Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society, 329–334.

[12] Keogh, E., Lin, J., Lee, S.-H., and Herle, H. V. 2006. Finding the most unusual time series subsequence: Algorithms and applications. Knowl. Inform. Syst. 11, 1, 1–27.

[13] Basu, S., Bilenko, M., and Mooney, R. J. 2004. A probabilistic framework for semi-supervised clustering. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 59–68.

[14] Augusteijn, M. and Folkert, B. 2002. Neural network classification and novelty detection. Int. J. Rem. Sens. 23, 14, 2891–2902.

[15] Cun, Y. L., Boser, B., Denker, J. S., Howard, R. E., Hubbard, W., Jackel, L. D., and Henderson, D. 1990.Handwritten Digit Recognition with A Back-Propagation Network. In Advances in Neural Information Processing Systems. 396–404. Morgan Koufamann.

[16] Chen, D., Shao, X., Hu, B., and Su, Q. 2005. Simultaneous Wavelength Selection and Outlier Detection in Multivariate Regression of Near-Infrared Spectra. Anal. Sci. 21, 2, 161–167.

[17] Spence, C., Parra, L., and Sajda, P. 2001. Detection, Synthesis and Compression in Mammographic Image Analysis with A Hierarchical Image Probability Model. In Proceedings of The IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. IEEE Computer Society, 3.

[18] Tandon, G. and Chan, P. 2007. Weighting Versus Pruning in Rule Validation for Detecting Network and Host Anomalies. In Proceedings of the 13th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. Acm Press.

[19] Lin, J., Keogh, E., Fu, A., and Herle, H. V. 2005. Approximations to Magic: Finding Unusual Medical Time Series. In Proceedings of the 18th Ieee Symposium on Computer-Based Medical Systems. Ieee Computer Society, 329–334.

[20] Stefano, C., Sansone, C., and Vento, M. 2000. To Reject or Not to Reject: That Is the Question: An Answer in The Case of Neural Classifiers. Ieee Trans. Syst. Manag. Cyber. 30, 1, 84–94.

[21] Odin, T. and Addison, D. 2000. Novelty Detection Using Neural Network Technology. In Proceedings of The Comaden Conference.

[22] SYKACEK, P. 1997. Equivalent error bars for neural network classifiers trained by Bayesian inference. In Proceedings of the European Symposium on Artificial Neural Networks. 121–126.

[23] BORIAH, S., CHANDOLA, V., AND KUMAR, V. 2008. Similarity measures for categorical data: A comparative evaluation. In Proceedings of the 8th SIAM International Conference on Data Mining. 243–254.

[24] Chandola, V., Boriah, S., and Kumar, V. 2008. Understanding Categorical Similarity Measures for Outlier Detection. Tech. Rep. 08-008, University of Minnesota.

[25] Jain, A. K. and Dubes, R. C. 1988. Algorithms for Clustering Data. Prentice-Hall, Inc.

[26] Basu, S., Bilenko, M., and Mooney, R. J. 2004. A Probabilistic Framework for Semi-Supervised Clustering. In Proceedings of the 10th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. Acm Press, 59–68