INTERNATIONAL CONFERENCE ON
RESEARCH IN SCIENCE AND TECHNOLOGY
14 December 2015, Kualalumpur –Malaysia

مروری بر روش های تشخیص داده پرت

رسول کیانی
گروه مهندسی کامپیوتر، پردیس علوم و تحقیقات فارس، دانشگاه آزاد اسلامی، مرودشت، ایران
rasoul.kiani87@yahoo.com

مجتبی منتظری
گروه مهندسی کامپیوتر، پردیس علوم و تحقیقات فارس، دانشگاه آزاد اسلامی، مرودشت، ایران
mojtabamontazeri2007@gmail.com

چکیده

هدف از مرحله ی پیش پردازش در داده کاوی، برطرف کردن مشکلات داده های مسئله مورد بررسی می باشد. شناسایی و حذف داده های پرت[۱] در فرایند پاک سازی داده ها[۲] از جمله عملیاتی است که کیفیت داده را بهبود می دهد. روش های متعدد تشخیص داده پرت در زمینه های تشخیص نفوذ در شبکه، تشخیص تقلب کارت اعتباری، حوزه سلامت عمومی و پزشکی، تشخیص خرابی در واحد مکانیکی، پردازش تصویر و... نیاز به بررسی و مطالعه روش های موجود را آشکارتر می نماید. در این مقاله، پس از مرور مفاهیم داده پرت، شامل مؤلفه های کلیدی و انواع داده پرت، الگوریتم های اصلی بر اساس دامنه کاربرد و نوع ناهنجاری[۳] مورد بررسی و مقایسه قرار گرفته و روش های پرکاربرد و جدید نیز با جزئیات بررسی شده است. هدف از این تحقیق ارائه یک مطالعه منسجم و کاربردی بر اساس روش های رایج در تشخیص داده پرت می باشد.

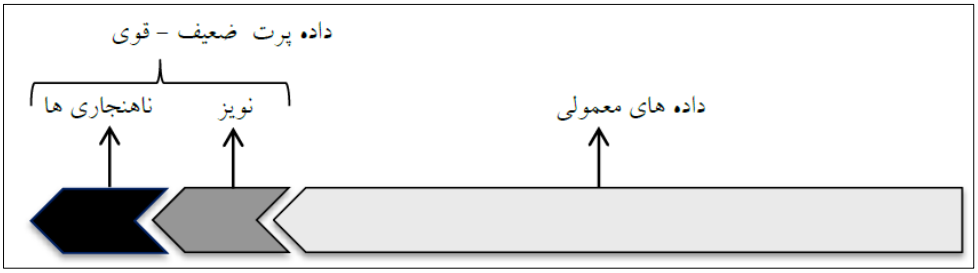
واژگان کلیدی: تشخیص داده پرت، تشخیص ناهنجاری

[۱] Outlier
[۲] Data Cleansing
[۳] Anomaly

مقدمه

داده کاوی روشی معتبر، جدید و مفید جهت شناسایی الگوهای قابل درک می باشد(Aggarwal and Yu, 2005) و در حال تبدیل شدن به یک ابزار مهم جهت تبدیل داده به دانش است. اگر نمونه ها نماینده خوبی از بدنه بزرگ داده نباشند، این فرایند شکست خواهد خورد(Arning et al, 1996). یکی از دلایل اصلی استفاده از داده کاوی تجزیه و تحلیل مؤثر و کارآمد مجموعه ای از مشاهدات مختلف با توجه به رفتار آن ها می باشد(Vijayarani, 2011). تشخیص داده پرت یا ناهنجاری در داده ها در اوایل قرن ۱۹ در جامعه آماری مطالعه شده است و به مرور تکنیک های متنوع تشخیص ناهنجاری توسعه داده شده اند، برخی از این تکنیک ها مخصوص استفاده در یک دامنه ی کاربردی خاص می باشند و برخی دیگر عمومی هستند. پرت کاوی به مشکل پیدا کردن الگوها در مجموعه داده های بزرگ که مطابق با رفتار مورد انتظار نمی باشند، اشاره دارد(Kantardzic, 2003). در واقع پایگاه داده ممکن است شامل اشیاء داده ای باشد که با رفتار عمومی یا مدل داده منطبق نیستند، چنین اشیاء داده ها که به طور برجسته متفاوت یا در تضاد با مجموعه داده باقیمانده هستند نقاط دورافتاده یا پرت نامیده می شوند(Han and Kamber, 2006). به طور کلی روش های تشخیص داده پرت به دو دسته تقسیم می شوند: دسته اول، بر اساس برچسب داده ها شامل روش های باناظر[۴]، نیمه ناظر[۵]، و بدون ناظر[۶] و دسته دوم، شامل روش های مبتنی بر توزیع[۷]، مبتنی بر خوشه بندی[۸]، مبتنی بر فاصله[۹] و مبتنی بر چگالی[۱۰].

هر نقطه داده در طیف داده های معمولی[۱۱] تا نویز[۱۲] و سرانجام ناهنجاری قرار می گیرد. در شکل ۱ تفکیک ناحیه های متفاوت این طیف دقیقاً تعریف نشده است، در واقع تفکیک بین نویز و ناهنجاری محض نیست و خیلی از نقاط داده ای که به وسیله فرایند تولید نویز ایجاد شده اند، ممکن است دارای انحراف کافی باشند برای اینکه به عنوان ناهنجاری در روش امتیاز دهی تفسیر شوند. بنابراین ناهنجاری ها امتیاز بیشتری از نویز خواهند داشت، اما این یک فاکتور تشخیص بین دو تعریف نیست و این علاقه تحلیل گر است که تفاوت بین ناهنجاری و نویز را مشخص و تنظیم کند. برخی محققان از اصطلاح پرت ضعیف[۱۳] و پرت قوی[۱۴] برای تشخیص بین نویز و ناهنجاری استفاده می کنند.



شکل ۱. افزایش امتیاز پرت بودن از ضعیف به قوی(Chandol et al, 2009)

[٤] Supervised-Based
[٥] Semi-Supervised-Based
[٦] Unsupervised-Based
[٧] Distribution-Based
[٨] Clustering-Based
[٩] Distance-Based
[١٠] Density-Based
[١١] Normal Data
[١٢] Noise
[١٣] Weak-Outlier
[١٤] Strong-Outlier

چالش های تشخیص داده پرت سبب شده است تا الگوریتم های مختلفی به منظور شناسایی و حذف این داده ها مورد استفاده قرار گیرند. مهم ترین چالش ها عبارت است از(Charu et al, 2005) و (Singh and Upadhyay, 2012):

- تعریف یک ناحیه طبیعی که شامل هر رفتار طبیعی امکان پذیر باشد دشوار است چرا که اغلب مرز بین رفتار طبیعی و غیر طبیعی دقیق نیست، بنابراین مشاهده غیر عادی که نزدیک مرز است می تواند طبیعی و بالعکس باشد.
- وقتی ناهنجاری نتیجه اقدامات مخرب است، دشمنان مخرب اغلب خود را به گونه ای وفق می دهند تا مشاهدات غیر عادی مانند مشاهدات معمولی به نظر رسند.
- مفهوم دقیق ناهنجاری در کاربردهای مختلف، متفاوت می باشد. به عنوان مثال در پزشکی یک انحراف کوچک از حد معمول ممکن است ناهنجاری باشد(مثلا نوسانات دمای بدن) ولی انحرافی مشابه این در دامنه بازار سهام (مثلاً نوسانات ارزش سهام) ممکن است معمولی باشد. بنابراین استفاده از یک تکنیک توسعه یافته در یک دامنه در دامنه دیگر ساده نمی باشد.
- در دسترس بودن برچسب داده برای آموزش / اعتبار مدل مورد استفاده توسط روش های تشخیص ناهنجاری معمولاً یک مسئله عمده است.
- اغلب داده شامل نویز هستند که تمایل دارند شبیه به ناهنجاری واقعی باشند از این رو تشخیص و حذف آن ها دشوار است.

در ادامه این مقاله، مؤلفه های کلیدی در تشخیص روش ناهنجاری و دسته بندی روش ها را در بخش مفاهیم مورد مطالعه قرار می دهیم. در بخش روش های جدید و پرکاربرد، جزئیات روش ها ذکر می شوند و نهایتاً در بخش نتیجه گیری به مقایسه آن ها می پردازیم.

مفاهیم

مؤلفه های کلیدی در تشخیص روش ناهنجاری

همانطور که در شکل ۲ نشان داده شده است، مؤلفه های کلیدی در تشخیص روش ناهنجاری عبارتند از: ماهیت داده[۱۵]، برچسب ها[۱۶]، نوع ناهنجاری و خروجی[۱۷].

ماهیت داده

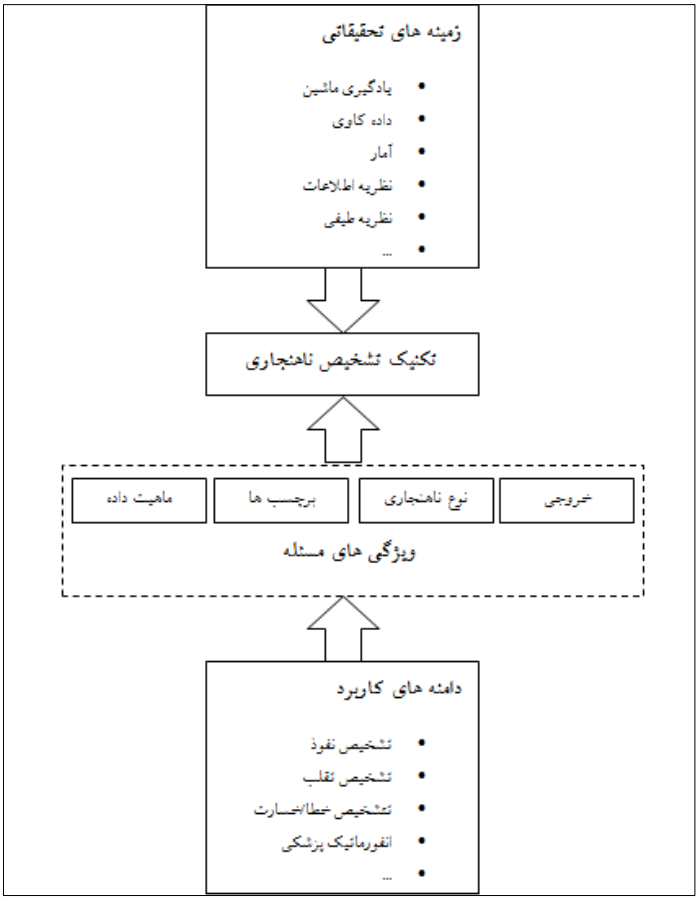
یکی از جنبه های کلیدی تکنیک های تشخیص ناهنجاری ذات داده های ورودی می باشد. ورودی به طور معمول مجموعه ای از نمونه های داده می باشد. هر نمونه داده می تواند به وسیله مجموعه ای از صفات توصیف شود که این صفات می توانند انواع مختلف نظیر دودویی، دسته ای و پیوسته باشند و هر نمونه داده ممکن است شامل یک صفت یا چندین صفت باشد. در نمونه داده های چند متغیره تمام صفات ممکن است از یک نوع یا ترکیبی از انواع مختلف باشند. در واقع ذات یا طبیعت داده، قابل اجرا بودن تکنیک تشخیص ناهنجاری را مشخص می کند.

برچسب ها

بر اساس در دسترس بودن برچسب برای داده های معمولی و پرت، سه روش وجود دارد(Lin et al, 2014) و (Zhou et al, 2013):

۱۵ Nature of Data
۱۶ Labels
۱۷ Output

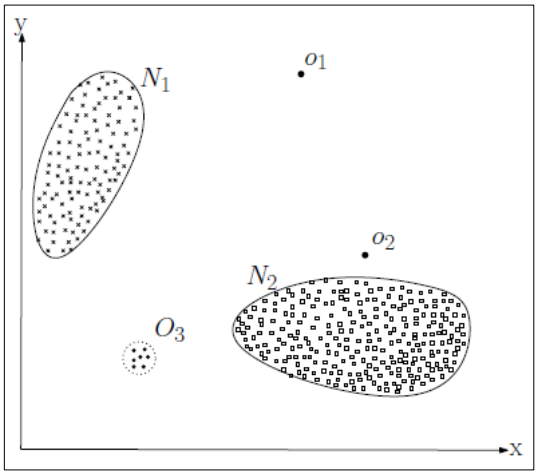
- روش با ناظر: تکنیک های آموزش در روش با ناظر فرض می کنند در مجموعه داده آموزش، داده های در دسترس، برچسب معمول یا پرت بودن را دارند. در روش با ناظر دو چالش پیش رو داریم. چالش اول رده های نامتوازن است چرا که داده های پرت معمولاً نادر هستند. راه حل این مشکل ایجاد داده های پرت مصنوعی است.
چالش دوم تشخیص هرچه بیشتر داده های پرت است. آن چه در این جا مهم است تشخیص درست داده های پرت است.
- روش نیمه ناظر: تکنیک هایی که بر روی حالت نیمه ناظر اعمال می شوند فرض می کنند که نمونه های داده فقط برچسب معمولی را دارند و بنابراین آن ها احتیاجی به برچسب پرت بودن ندارند. این روش ها نسبت به روش با ناظر بیشتر مورد استفاده قرار می گیرند.
- روش بدون ناظر: تکنیک های بدون ناظر که جهت تشخیص داه پرت استفاده می شوند نیازی به داده های آموزشی ندارند، بنابراین قابلیت استفاده بیشتری دارند. در این روش داده های معمولی به چند گروه متمایز تقسیم می شوند. انتظار می رود داده های پرت با فاصله از هر یک از گروه های داده های معمولی قرار گیرند. اما ضعفی که موجود است معمولاً داده های معمولی از الگوی خاصی تبعیت نمی کنند و بالعکس داده های پرت دسته جمعی دارای شباهت زیادی در یک محدوده کوچک هستند.



شکل ۲. مؤلفه های کلیدی مرتبط با روش تشخیص ناهنجاری (Chandol et al, 2009)

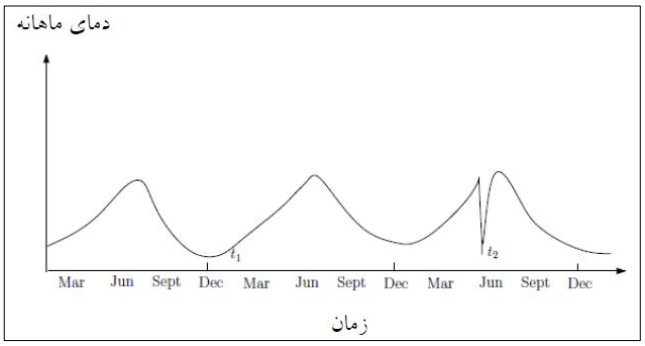
نوع ناهنجاری

- داده های پرت عمومی(ناهنجاری نقطه ای[۱۸]): این داده ها با فاصله قابل توجهی از سایر داده ها قرار دارند. ناهنجاری نقطه ای می تواند به عنوان یک نمونه داده منحصر بفرد که از الگوی معمول داده ها واگرا شده است، تعریف شود(Shukla et al, 2014). به عنوان مثال در شکل ۳ نقاط o_1 و o_2 از این نوع می باشند.



شکل ۳. ناهنجاری نقطه ای و جمعی (Chandol et al, 2009)

- داده های پرت زمینه ای(ناهنجاری مشروط[۱۹]): این داده ها بسته به شرایط می توانند نسبت به سایر داده ها پرت باشند یا نباشند. انتخاب تکنیک مناسب جهت تشخیص ناهنجاری مشروط به وسیله معنی دار بودن آن در دامنه کاربرد هدف تعیین می شود. به عنوان مثال در شکل ۴ سری زمانی دما در ماه های مختلف در یک ناحیه خاص نشان داده شده است. دمای هوای ۳۵ درجه ممکن است در فصل زمستان(در زمان t_1) در آن مکان معمولی باشد، اما همان دما در فصل تابستان(در زمان t_2) ممکن است ناهنجاری محسوب گردد.



شکل ۴. ناهنجاری مشروط (Chandol et al, 2009)

۱۸ Point Anomaly
۱۹ Contextual Anomaly

- داده های پرت جمعی(ناهنجاری دسته ای[۲۰]): اگر دسته ای از داده های مرتبط با یکدیگر نسبت به سایر داده ها دارای انحراف باشند به عنوان ناهنجاری دسته ای در نظر گرفته می شوند. یکی از مشکلات تشخیص این نوع از داده ها این است که ما باید از ارتباطات بین داده ها اطلاع داشته باشیم. در شکل ۳ O_3 ناهنجاری دسته ای است(امیری، ۱۳۸۹) و (Chandula et al, 2009).

خروجی

یکی دیگر از مهم ترین جنبه های تکنیک های تشخیص ناهنجاری، رفتاری است که برای ناهنجاری ها گزارش می شود. معمولاً خروجی های تولید شده یک از دو نوع زیر می باشند:

- امتیازها[۲۱]: تکنیک های امتیازی وابسته به درجه ای که هر نمونه به عنوان ناهنجاری در نظر گرفته می شود، امتیاز ناهنجاری را برای آن در نظر می گیرند. بنابراین خروجی این تکنیک ها لیستی از ناهنجاری های رتبه بندی شده هستند. تحلیل گر ممکن است تعدادی از ناهنجاری های رتبه بالای این لیست یا یک حد آستانه برش[۲۲] را برای انتخاب ناهنجاری ها استفاده نماید.
- برچسب ها: در این تکنیک به هر نمونه آزمایش برچسب معمولی یا ناهنجاری تخصیص داده می شود. تکنیک های تشخیص ناهنجاری مبتنی بر امتیاز، به تحلیل گر اجازه استفاده از حدآستانه برای انتخاب ناهنجاری های مرتبط را می دهند ولی تکنیک های مبتنی بر برچسب که از برچسب های دودویی برای نمونه های آزمایش استفاده می کنند به طور مستقیم چنین اجازه ای را نمی دهند. بنابراین این ها می توانند به طور غیر مستقیم با استفاده از پارامترهای انتخاب کنترل شوند.

پس از توضیحات لازم در مورد مؤلفه های کلیدی تأثیر گذار در انتخاب تکنیک مناسب جهت تشخیص نوع ناهنجاری، در جدول ۱ روش های ارائه شده بر اساس دامنه کاربرد، روش مورد استفاده و نوع ناهنجاری مشخص شده است.

جدول ۱. بررسی روش های تشخیص داده پرت بر اساس دامنه کاربرد و نوع ناهنجاری

دامنه کاربرد	روش مورد استفاده			نوع ناهنجاری		
	با ناظر	نیمه ناظر	بدون ناظر	ناهنجاری نقطه	ناهنجاری مشروط	ناهنجاری جمعی
تشخیص نفوذ	*	*	*	*		*
تشخیص تقلب		*		*	*	
خسارت صنعتی		*			*	*
پزشکی		*		*		*
پردازش تصویر	*		*		*	*

دسته بندی رایج تشخیص داده پرت

روش های مبتنی بر توزیع

روش های مبتنی بر توزیع، مدل های آماری را برای مجموعه داده ها توسعه می دهند و سپس برای تعیین اینکه آیا یک داده متعلق به این مدل می باشد یا خیر، آزمون آماری اعمال می نمایند. داده هایی که با احتمال کم متعلق به مدل آماری هستند به عنوان داده پرت اعلام می شوند.کیفیت این روش معمولاً وابستگی شدیدی به مفروضات ما در مدل واقعی دارد. از دیگر معایب این روش می توان به اینکه معمولاً آزمایش برای یک صفت انجام می پذیرد، در بسیاری از موارد توزیع داده برای ما ناشناخته

۲۰ Collective Anomaly
۲۱ Scores
۲۲ Cut-Off Threshold

است و اینکه توانایی تشخیص تمامی داده های پرت را ندارند، اشاره کرد. این روش به دو دسته ی پارامتری و غیر پارامتری تقسیم می شود.

روش های مبتنی بر خوشه بندی

روش های مبتنی بر خوشه بندی، خوشه های کوچک را به عنوان داده های پرت در نظر می گیرند. منظور از خوشه های کوچک، خوشه هایی هستند که میزان قابل توجهی نقاط داده کمتری نسبت به سایر خوشه ها دارند. روش فوق، بدون ناظر می باشد و می توان پس از خوشه بندی نقاط جدید را وارد و پرت بودن آن ها را مورد آزمایش قرار داد.

روش های مبتنی بر فاصله

در روش های مبتنی بر فاصله یک داده، داده ی پرت است اگر در همسایگی آن به اندازه کافی نقاط دیگری موجود نباشند. نگرانی اصلی در روش مبتنی بر فاصله حجم بالای محاسبات است. چرا که برای هر نقطه باید فاصله ی آن نقطه تا سایر داده های موجود تعیین گردد. برای بهبود این روش از دو مکانیزم استفاده می شود. اولاً هر داده را تنها با همسایگی های خود آزمایش کرده و دوم به جای بررسی داده ها به صورت یکی با یکی، به صورت گروهی بررسی انجام می گیرد.

روش های مبتنی بر چگالی

در روش های مبتنی بر تراکم، داده ای پرت است که تراکم آن نسبت به تراکم همسایگان خود به میزان قابل توجهی کمتر باشد برای معرفی روش های مبتنی بر تراکم ابتدا داده های پرت محلی را تعریف می کنیم. داده های پرت محلی به جای توزیع داده های عمومی به نسبت همسایگان خود تعریف می شوند.

در جدول ۲ روش های رایج تشخیص داده پرت بر اساس پیچیدگی محاسباتی، کارایی، فضای ویژگی و کاربرد عملی مقایسه شده است.

جدول ۲. بررسی روش های رایج تشخیص داده پرت

نام روش	پیچیدگی محاسباتی	کارایی	فضای ویژگی	کاربرد عملی
مبتنی بر توزیع	خیلی پیچیده	کم	تک متغیره	اطلاعات آماری
مبتنی بر فاصله	ساده	کارآمد	چند متغیره	بر اساس نزدیکی نقاط فردی
مبتنی بر چگالی	خیلی پیچیده	کارآمدتر	چند متغیره	بر اساس نزدیکی نقاط فردی و نزدیک ترین همسایه
مبتنی بر خوشه بندی	پیچیدگی کم	بسیار کارآمد	تک متغیره/ چند متغیره	بر اساس خوشه بندی داده های مشابه

روش های جدید و پرکاربرد

در این بخش به معرفی روش های جدید و پرکاربرد در زمینه تشخیص داده های پرت می پردازیم.

روش مبتنی بر RNN[۲۳]

این روش مناسب مجموعه داده های چند متغیری بزرگ می باشد. در واقع RNN یک نوع MLP است که تعداد نرون های ورودی و خروجی برابر هستند و سه لایه پنهان دارد. هدف این روش باز تـولید نـقاط ورودی در لایه خروجی با حداقل خطای

۲۳ Replicator Neural Network

بازسازی، بعد از فشرده سازی از طریق لایه های پنهان می باشد. (که شامل تعداد گره های کمتری از لایه ورودی و خروجی است.) این نقاط می توانند به عنوان داده پرت در نظر گرفته شوند. معیار پرت بودن مبتنی بر میانگین خطای بازسازی، به عنوان علامت گذاری جدید استفاده می شود. روش بهبود یافته RNN مدل شبکه عصبی auto-associative می باشد که از میانگین مربع خطا به منظور معیار پرت بودن استفاده می کند. این روش به منظور تشخیص داده پرت در داده های ارتعاش[۲۴] برای تشخیص تقلب در واحد مکانیکی همزمان، مورد استفاده قرار گرفته است(Marco et al, 2014).

روش CISO[۲۵]

در این روش از نمونه های برچسب دار مشکوک که توسط روش های بدون ناظر تشخیص داده شده است استفاده می شود. در واقع مجموعه برچسب دار تعریف می شود در حالی که قبلاً مجموعه برچسب دار معادل مجموعه آموزش بوده است. بر اساس این ایده الگوریتم CISO ، (بر مبنای پاسخ به این سؤال که: چگونه تشخیص دهیم مجموعه برچسب دار به طوری که تقریباً شامل تمام داده های پرت و حداقل درون هشته[۲۶] ها باشد؟) ارائه شده است. ابتدا نمونه ها در مخزن[۲۷] به وسیله ی الگوریتم بدون ناظر تشخیص داده پرت دسته بندی می شوند، سپس نمونه های مشکوک انتخاب و به صورت دستی برچسب زده می شوند و سایر نمونه های باقیمانده برچسب درون هشته دریافت می کنند. در نتیجه تمام نمونه ها در مخزن برچسب دارند و در مجموعه آزمایش استفاده می شوند. یک نقطه برش خوب باید مجموعه مشکوک شامل تمام داده های پرت و تعداد کمی درون هشته را تولید کند، بنابراین تمام داده های پرت می توانند توسط متخصص با حداقل تلاش ممکن تأیید شوند. ممکن است کاربر بخواهد بر اعتبار برچسب زدن کنترل داشته باشد، به همین دلیل روش Budgeted CISO پیشنهاد می شود. در این روش داده بدون برچسب، ارائه شده و برای ساخت مجموعه آزمایش با توزیع یکسان به عنوان مجموعه داده اصلی تلاش می شود. هدف به حداقل رساندن برچسب زدن دستی می باشد(Liu and Fern, 2012).

روش HAC[۲۸]

در روش پیشنهادی ازدحام ذرات از طریق مراحل مختلف برای شناسایی نقاط دورافتاده و خوشه های طبیعی مورد استنتاج قرار می گیرد و از یک روش کاربردی برای تولید یک سلسله مراتب از خوشه ها استفاده می شود. هر سطح سلسله به عنوان نسلی از جمعیت عمل می شود. نسل اولیه شامل کل جمعیت می باشد. هر ذره از جمعیت به عنوان یک مرکز خوشه است. در ادامه جمعیت به وسیله ادغام دو خوشه در هر نسل موفق به سمت تبدیل شدن به یک خوشه پیش می رود. مزیت این روش این است که ذرات در بهینه محلی گرفتار نمی شوند. در این روش از رفتار جمعی ذرات برای خود سازماندهی موقعیت آن ها با توجه به پیکربندی خوشه بندی صحیح نمایش گذاشته شده توسط داده ها استفاده می شود. روش کار جهت تشخیص داده های پرت: ایجاد نسلی از خوشه به منظور تشخیص مشاهدات مشکوک و شناسایی آن ها بر مبنای فاصله آن ها از مراکز خوشه است(Alam et al,2010).

کنترل کشف داده پرت با برچسب ناقص به وسیله تعمیم چارچوب یادگیری بردار پشتیبان داده

در این روش دو مدل احتمالی تعریف می شود: تک احتمالی[۲۹] و دو احتمالی[۳۰]. در مدل تک احتمالی هر داده ورودی همراه با یک مقدار احتمال است که نشان دهنده درجه عضویت به سمت برچسب کلاس خود می باشد. در مدل دو احتمالی، هر نمونه

---

۲٤ Vibration
۲٥ Constructing training set by Identification Suspected Outliers
۲٦ Inlier
۲۷ Repository
۲۸ Hierarchical Agglomerative Clustering
۲۹ Single-Likelihood
۳۰ Bi-likelihood

دو ارزش احتمال دارد که دلالت بر درجه عضویت به سمت برچسب کلاس مثبت و منفی می باشد. بر اساس مدل دو احتمالی، شبه مجموعه داده آموزش به وسیله محاسبه مقادیر احتمال بر مبنای رفتار داده های محلی در فضای ویژگی تولید می شود.
روش معرفی شده بر مبنای خوشه بندی K-means و LOF برای تولید مقادیر احتمال جدید می باشد. این الگوریتم ها را Kernel k-means clustering و Kernel LOF می نامند.
ابتدا دو شبه مجموعه آموزش برای دو مدل احتمال معرفی شده که هر نمونه آن شامل مقادیر احتمال است، تعیین می شود. در ادامه دو کلاس هدف، برای تشخیص داده پرت بوسیله مدلSVDD بر مبنای مدل دو احتمالی ایجاد می گردد. مدل مشتق شده از تک احتمالی را soft-SVDD و مدل مشتق شده از دو احتمالی را bi-soft-SVDD می نامند. برای هر دو روش ارزش احتمال تولید از هر نمونه و نمونه های منفی محدود برای ساخت کلاس های دقیق تشخیص داده پرت ترکیب می شوند. در این فرایند هر نمونه به تشخیص مرز داده پرت بر اساس ارزش احتمالی آن ها کمک می کند. این روش یک مصالحه[۳۱] بین نرخ تشخیص و نرخ هشدارِ غلط، ارائه می دهد و حساسیت کمتری نسبت به داده های نویز دارد(Liu et al, 2014).

انتخاب آزمایش هایی برای تشخیص داده پرت

در این روش با استفاده از یک رویکرد نقص گرا[۳۲] و چارچوب ریاضی مجموعه ای از ابزار هدف به منظور انجام آزمایش ها برای تشخیص داده پرت و تعیین پارامترها جهت روش های مختلف ارائه می شود. همچنین آزمایش هایی انتخاب می شوند که اگر همه آن ها با محدودیت های مشخص اجرا شوند، شکست های یکنواخت تشخیص داده خواهند شد. در نهایت کاهش عملکرد حاصل به عنوان ابزارهای شکست می تواند نشانه ای برای قابلیت اطمینان پایین باشد. بخش کلیدی مدل صفحه نمایش پرت[۳۳] نامیده می شود. صفحه نمایش پرت به عنوان ترکیبی از روش های تشخیص تعریف شده است و آزمایش و تنظیمات پارامترها بر روی آن اعمال می گردد. این بدان معناست که اگر یک روش تشخیص داده پرت بر روی آزمایش های مختلفی اعمال شود، همه این ها به عنوان صفحه نمایش مختلف در این مدل در نظر گرفته می شوند. از این رو، استفاده از روش های (اکتشافی) راه حل که قادر به مقابله با تعداد زیادی از صفحه نمایش پرت است ضروری می باشد. در این مدل فقط نتیجه روش تشخیص داده پرت در نظر گرفته شده است و از آن جا که می تواند هر روش تشخیص داده پرت را کنترل کند، بنابراین هرکاربر این امکان را دارد که از روش مورد علاقه خود استفاده نماید.
مؤلفه مهم دیگر مدل مجموعه هدف می باشد. مجموعه هدف زیر مجموعه ای از ابزار است که باید توسط روش های تشخیص داده پرت، شناسایی شود. این مجموعه هدف برای هدایت فرایند انتخاب آزمایش ضروری است. همچنین مجموعه پاداش[۳۴] معرفی می شود. اگر ابزارهایی وجود دارند که اهداف مستقیمی ندارند و نیازی نیست که در عملکرد محاسبات از دست دادن گنجانده شوند، این مجموعه می تواند استفاده شود(Bossers et al, 2013).

تشخیص داده پرت با استفاده از روش مبتنی بر فاصله بر روی داده های نامشخص

در این روش تعریف جدیدی از داده پرت ارائه می شود و نتایج بدست آمده نشان می دهد که کارایی لازم را بر روی مجموعه داده ها دارد. روش های تشخیص داده های پرت های موجود به طور عمده در پایگاه داده های سنتی قطعی استفاده می شوند، که در آن وجود و قابلیت اطمینان از داده ها قطعی می باشد اما در برخی دامنه های کاربردی مهم مثل شبکه های حسگر، عدم قطعیت در داده ها ذاتی است و به معیارهای مختلفی وابسته است. برای مثال، استفاده از شبکه های حسگر در مقیاس بزرگ محیط پیمایش، تکمیل نبودن داده ها، محدودیت تجهیزات، و تأخیر یا از دست دادن در انتقال داده ها، ممکن است به عدم اطمینان از داده ها منجر شود. روابط انحصاری متقابل، چالش بیشتری را بر روی تشخیص داده پرت مطرح می کند. برای تعیین اینکه آیا t پرت است، باید احتمال ظاهر شدن حداقل k تاپل در N(t) محاسبه شود. بدین منظور به صورت تدریجی

۳۱ Tradeoff
۳۲ Defect-Oriented
۳۳ Outlier Screens
۳٤ Bonus Set

نمونه های تمام اشیاء در N(t) بازیابی می شوند و محاسبه احتمال ظاهر شدن حداقل k تاپل که در مجموعه نمونه ها ملاقات شده است وجود دارد. اگر بعد از بازیابی تعدادی تاپل t_j ،احتمال بیشتر از λ شود الگوریتم متوقف می شود، t داده پرت نیست و اگر بعد از ملاقات تمام تاپل ها در N(t) احتمال هنوز کمتر از λ باشد، t داده پرت است. بر اساس نتایج روش ذکر شده مقیاس خوبی دارد و در پایگاه داده های بزرگ قابل استفاده می باشد(Mingke et al, 2013).

روش ODAGS[۳۵]

در این روش فرمول جدیدی برای تشخیص درجه پرت اشیاء بر مبنای نظریه محاسبه دانه ای ارائه می گردد. ابتدا تمام اشیاء بر اساس فاکتور پرت بودن رتبه بندی شده و k شیء بالای این رتبه بندی به عنوان پرت معرفی می شوند. این الگوریتم برای تشخیص داده پرت مجموعه داده دسته ای مورد استفاده قرار می گیرد. سیستم اطلاعات دسته ای بر اساس رابطه (۱) تعریف می شود(Li, 2014):

$$IS=(U,A,V,f_a)_{a \in A} \quad (۱)$$

که در آن $U=\{x_1,x_2,\ldots,x_n\}$ و A مجموعه ای از اشیاء محدود و غیر تهی که جهان نامیده می شود، V دامنه ی مجموعه مقادیر صفات a که محدود و نا مرتب است و f تابع اطلاعاتی است که بر اساس رابطه (۲) نشان داده شده است:

$$f:U\times A\rightarrow V \quad (۲)$$

معیار پرت بودن شیء x بر صفات f بر اساس رابطه (۳) می باشد:

$$OF_A(x)=\frac{1}{|A|}\sum_{a\in A}\frac{1}{|[x]_{\{a\}}|} \quad (۳)$$

معیار پرت بودن می تواند شاخصی از درجه پرت بودن برای شیء باشد.

طرح کلی از روش های تشخیص داده پرت با تمرکز بر نقاط دورافتاده بر اساس رویداد[۳۶] و خطا[۳۷]

در شبکه های حسگر بی سیم[۳۸] تعدادی از گره های حسگر (چند ده هزار نفر) با یکدیگر کار می کنند، به طوری که یک منطقه خاص می تواند برای جمع آوری داده های مربوط به محیط، تحت نظارت قرار گیرد. تشخیص ناهنجاری یک گام لازم برای اطمینان از داده های جمع آوری شده توسط شبکه های حسگر بی سیم می باشد که کیفیت خوب داده، نظارت امن و تشخیص قابل اطمینان رویدادهای بحرانی و مجتمع را فراهم می کند. تمایز میان علل نقاط پرت اهمیت دارد و این بینش را که چگونه به آن ها رسیدگی شود ارائه می دهد. منبع پرت بودن در شبکه های حسگر بی سیم می تواند خطا یا رویداد داده باشد.

- خطا: خطاها اساسا مشاهداتی هستند که متفاوت از مقدار مناسب اندازه گیری می باشند. حسگر معیوب مسئول خطاها و داده های نویز است. داده های پرت ناشی از خطاها ممکن است بارها رخ دهد ولی داده های پرت ناشی از رویداد با احتمال کمتری نسبت به آن ها ظاهر می شود.

۳۵ Outlier Detection Algorithm Based on Granular Set
۳۶ Event
۳۷ Error
۳۸ Wireless Sensor Networks

- رویداد: رویداد ها مشاهدات تغییر در وضعیت محیط را نسبت به رفتار معمول از پیش تعریف شده نشان می دهند. یک رویداد به عنوان یک پدیده خاص که حالتی از جهان واقعی را تغییر می دهد تعریف می شود. به عنوان مثال: آتش جنگل، آلودگی هوا و غیره.

برای تشخیص داده در شبکه های حسگر بی سیم، دو مکانیزم ارائه می شود: مکانیزم متمرکز و توزیع شده. در مکانیزم متمرکز دو الگوریتم خوشه بندی و تشخیص داده پرت، وقتی که تمام داده ها از هر گره حسگر به گره مقصد[۳۹] انتقال داده شد، اجرا می شوند. در مکانیزم توزیع شده ابتدا الگوریتم خوشه بندی بر روی داده در گره حسگر اجرا می شود.

برای اجرای خوشه بندی الگوریتم خوشه بندی از پایین به سطح شبکه منتقل می شود آن جا که هر گره حسگر، الگوریتم خوشه بندی را بر روی داده خودش برای تولید خوشه ها اجرا می کند و والدین گره ها، خوشه های خودشان را با خوشه فرزندان میانی ترکیب می کنند. سرانجام الگوریتم تشخیص داده پرت برای تشخیص داده پرت اجرا می گردد. روش ارائه شده خوشه بندی جدید در ترکیب با نزدیک ترین همسایه، برای طبقه بندی داده های پرت، داده های نویز یا رویدادهای جالب و اطلاعات نادرست ارائه شده است(Shukla et al,2014).

نتیجه گیری

در این مقاله ابتدا به توضیح در مورد تعریف داده پرت پرداختیم، سپس مؤلفه های کلیدی تأثیر گذار در انتخاب تکنیک مناسب جهت شناسایی و تشخیص آن ها مورد بررسی و مقایسه قرار گرفت و دانستیم بر حسب دامنه کاربرد، ویژگی های مسئله و زمینه های تحقیقاتی می بایست روش مناسب جهت شناسایی داده های پرت انتخاب شود. همچنین دسته بندی های مختلف روش های تشخیص ذکر گردید، در ادامه الگوریتم های رایج و کاربردی تر را شرح دادیم و در این بخش بر اساس جدول ۳ به مقایسه آن ها از نظر زمینه کاربرد، تکنیک مورد استفاده، هدف، نوع تکنیک و چالش های بررسی شده می پردازیم.

[۳۹] Sink

جدول ۳. مقایسه الگوریتم های جدید و کاربردی تشخیص داده پرت

نام روش	زمینه کاربرد	تکنیک مورد استفاده	هدف	نوع تکنیک	نوع ناهنجاری	چالش های بررسی شده
مبتنی بر RNN	تشخیص داده های ارتعاش در واحد مکانیکی	شبکه عصبی MLP با ۳ لایه پنهان	بازسازی نقاط ورودی در خروجی با حداقل خطای بازسازی	نیمه ناظر	مشروط- دسته ای	درجه پرت بودن و تأثیر آن
CISO	ساخت مجموعه آموزش شامل داده های معمولی و پرت	خوشه بندی	بررسی اعتبار مجموعه داده پرت	نیمه ناظر	نقطه	در درسترس بودن برچسب داده- درجه برت بودن و تأثیر آن
HAC	ایجاد نسلی از خوشه به منظور تشخیص مشاهدات مشکوک	خوشه بندی	عدم گیر افتادن در بهینه محلی بر مبنای فاصله از مرکز به منظور تشخیص نویز یا پرت	خوشه بندی	دسته ای	تأثیر گذاری نویز
کنترل کشف داده پرت با برچسب ناقص	ساخت مجموعه آموزش	K-means بر مینای رفتار داده های محلی پرت	ساخت مجموعه آموزش	خوشه یندی	نقطه-دسته ای	مرز بین داده های معمولی و پرت- تأثیر گذاری نویز
انتخاب آزمایش های مختلف	استفاده کاربر از روش مورد علاقه	ترکیبی از روش های تشخیص با اعمال محدودیت های مشخص	تشخیص داده پرت	صفحه نمایش پرت	نقطه	مرز بین داده های معمولی و پرت
روش مبتنی بر فاصله	شبکه های حسگر	روش مبتنی بر فاصله	قابل استفاده در پایگاه داده های بزرگ	نیمه ناظر	دسته ای- مشروط	زمینه کاربرد
ODGAS	تشخیص داده پرت	ارائه فرمول جدید بر مبنای نظریه محاسبات دانه ای	رتبه بندی اشیاء بر اساس فاکتور پرت بودن	بدون ناظر	دسته ای	مرز بین داده های معمولی و پرت- درجه برت بودن
تشخیص داده پرت براساس رویداد و خطا	شبکه های حسگر	خوشه بندی- الگوریتم تشخیص داده پرت	تشخیص داده پرت	خوشه یندی	مشروط- دسته ای	تأثیر گذاری نویز

منابع

امیری، محمد جواد،۱۳۸۹، روش های تشخیص داده پرت، دانشگاه علم و صنعت ایران.

Aggarwal and Yu. (2005). An effective and efficient algorithm for high-dimensional outlier detection, The VLDB Journal, (vol. 14), pp. 211-221.

Alam, Dobbie, Riddle and Naeem. (2010). A swarm intelligence based clustering approach for outlier detection, IEEE Press, 978-1-4244-8126.

Arning, Agrawal and Raghavan. (1996). A linear method for deviation detection in large databases, ACM SIGKDD.

Bossers, Hurink and Smit. (2013). Selection of tests for outlier detection, IEEE 31st VLSI Test Symposium (VTS).

Chandola, Banerjee and Kumar. (2009). Anomaly detection: A survey, ACM Comput, Surv, 41(3).

Charu, Aggarwal, and Phillip. (2005). An effective and efficient algorithm for high dimensional outlier detection, The VLDB Journal.

Han and Kamber. (2006). Data mining concepts and techniques, Morgan kaufmann publishers.

Kantardzic. (2003). Data mining concepts, models, methods, and algorithms, Wiley – Interscience Publications, IEEE Press.

Li. (2014). Outlier detection algorithm for categortical data using a granular computing theory, IEEE Workshop on Electronics, Computer and Applications.

Lin, Feng and Ying. (2014). A new outlier detection algorithm and its application in intelligent transportation system, IEEE Press 978-1-4799-4419.

Liu, Xiao, Yu, Hao and Cao. (2014). An efficient approach for outlier detection with imperfect data labels, IEEE Transactions on knowledge and data engineering, (VOL. 26), NO. 7.

Liu and Fern. (2012). Constructing training sets for outlier detection. SDM conference.

Marco, Pimental, Clifton.D, Clifton.L, and Tarassenko. (2014). A review of novelty detection, page 215-249.

Mingke, Zheyuan and Ni. (2013). Distance based outlier detection on uncertain data of mutually exclusive relation, IEEE Press 978-1-4799-1027.

Shukla, Pandey and Kulhari. (2014). Outlier detection: A survey on techniques of WSNs involving event and error based outliers, International conference on innovative applications of computational intelligence on power, energy and controls with their impact on humanity.

Singh and Upadhyay. (2012). Outlier detection: Applications and techniques,IJCSI international journal of computer science issues, (Vol. 9), Issue 1, No 3.

Vijayarani. (2011). An efficient clustering algorithm, for outlier detection, IJCA,(vol 32).

Zhou, Zhao, Liu and Cui. (2013). Semi-supervised based training set construction For outlier detection, International conference on cloud computing and big data.