

Introducing a Data Mining Hybrid Model Based on Autoencoder Deep Learning, and Support Vector Machine Method for Fraud Detection in Imbalanced Data

ارائه یک مدل ترکیبی داده کاوی مبتنی بر یادگیری عمیق اتوانکودر و روش ماشین بردار پشتیبان
برای تشخیص تقلب در داده‌های نامتوازن

۲- اهمیت و ضرورت تحقیق

اهمیت و ضرورت تحقیق از دو منظر مثبت و منفی بررسی خواهد شد. از منظر مثبت، تعریف تحقیق حاضر می‌تواند زمینه‌ساز شفافیت بیشتر در داده‌های تراکنش‌های بانکی یا تراکنش‌های دوجانبه (سازمان-مشتری) شود و در نتیجه منجر به افزایش سودآوری سازمان مربوطه گردد. همچنین، تعریف این پروژه و ارزیابی میزان دستیابی به نتایج مورد انتظار می‌تواند زمینه‌ساز جهت‌دهی بهتر سایر تحقیقات در زمینه تشخیص تقلب و مسائل عمده مرتبط با آن باشد. از منظر منفی، فقدان این تحقیق هزینه‌های آشکار و پنهان زیادی را بر سازمان‌ها تحمیل خواهد کرد. در حال حاضر، بسیاری از سازمان‌ها از فقدان داده‌های برجسب‌گذاری شده کافی برای مدل‌سازی مسئله برای خود شکایت دارند. بنابراین، فقدان تعریف تحقیقاتی مانند مطالعه حاضر، مسئله موجود را باز نگه می‌دارد، در حالی که ارائه مدلی برای بازسازی داده‌های تخلف از محدود تخلفات شناسایی شده می‌تواند عملکرد بالایی را برای موارد جدید و نادیده گرفته شده فراهم کند. می‌توان گفت که هم داده کاوی و هم یادگیری ماشین به عنوان زیرمجموعه‌های علم داده شناخته می‌شوند و یک هدف مشترک دارند: استخراج اطلاعات مفید از داده‌ها. هر دو روش به داده به عنوان ورودی نیاز دارند و از الگوریتم‌ها و ابزارهای مشابه زیادی استفاده می‌کنند. پیش‌پردازش داده‌ها، ارزیابی مدل و انتخاب ویژگی‌های مناسب نیز در هر دو روش بسیار مهم است.

۳- اهداف تحقیق

در این تحقیق، ما یک مدل داده کاوی ترکیبی برای یافتن تخلفات در پایگاه داده ارائه می‌دهیم، زمانی که تعداد داده‌های تخلف کم و فراوانی برجسب‌های داده‌ها نامتعادل است.

علاوه بر این، ما سعی خواهیم کرد به موارد زیر دست یابیم:

- یافتن یک رویکرد مناسب برای بازتولید داده‌های تخلف موجود برای متعادل کردن تعداد برچسب‌ها

- استفاده از یک رویکرد مناسب برای انتخاب مناسب‌ترین ویژگی‌ها برای جاسازی در مدل طبقه‌بندی.

- پیشنهاد یک مدل داده‌کاوی مناسب برای جداسازی کلاس‌های داده‌های سالم و تخلف.

- طراحی یک مدل تشخیص تخلف ترکیبی به گونه‌ای که عملکرد بهتری نسبت به مدل‌های مشابه در ادبیات داشته باشد.

۴- روش تحقیق

در ابتدا، بر اساس مطالعه و تحقیق علمی گزارش‌های فنی، کتاب‌ها، مقالات، پایان‌نامه‌ها، پروژه‌های تحقیقاتی و مطالب موجود در وبسایت‌های علمی و فنی در زمینه‌های داده‌کاوی و به‌ویژه بررسی مقالات تحقیقاتی انجام‌شده در این زمینه، شکاف تحقیقاتی بیشتر بررسی و نوآوری پیشنهادی بیشتر گسترش داده خواهد شد.

سپس، بر اساس روش‌شناسی Crisp-DM، فرآیند داده‌کاوی و پیاده‌سازی مدل انجام خواهد شد.

مراحل این روش‌شناسی شامل مراحل زیر خواهد بود:

مرحله ۱:

درک ماهیت کسب و کار: درک ماهیت متغیرها و کسب و کار حاکم بر داده‌های مورد مطالعه می‌تواند دیدگاه‌های کلی در مورد نحوه مهندسی ویژگی‌ها در آینده ارائه دهد. بنابراین، این مرحله، گامی اساسی در پیاده‌سازی داده‌کاوی است.

مرحله ۲: شناخت داده‌ها:

داده‌های موجود در پایگاه داده باید بررسی و ویژگی‌های مناسب برای پیش‌بینی متغیرها انتخاب شوند. در این مرحله، با استفاده از رویکردهای تحلیل اکتشافی و مصورسازی، داده‌ها از دیدگاه‌های مختلف بررسی می‌شوند.

مرحله ۳: آماده‌سازی داده‌ها:

داده‌های خام انتخاب‌شده در بیشتر موارد برای تجزیه و تحلیل مناسب نیستند. اگر داده‌های از دست رفته‌ای وجود دارد، باید با جایگزین مناسب پر شود. در صورت نیاز به تبدیل و نرمال‌سازی، باید انجام شود، در صورت لزوم، ویژگی‌های جدیدی از ویژگی‌های موجود ایجاد شود یا ویژگی‌های جدیدی مانند تعطیلات و غیره به داده‌ها اضافه شود و غیره. این مرحله گامی مهم و کلیدی در دستیابی به نتایج با دقت بالا است.

مرحله ۴: مدل‌سازی:

در این مرحله، ابتدا رویکرد پایه کاغذی پیشنهادی بر روی داده‌های تخلف نامتوازن پیاده‌سازی می‌شود و سپس رویکرد پیشنهادی پیاده‌سازی می‌شود و در نهایت مقایسه عملکرد مدل‌ها بر اساس شاخص‌ها و موارد مختلف گزارش می‌شود.

مرحله ۵: ارزیابی:

مدل پیشنهادی باید از نظر شاخص‌هایی مانند دقت و تعمیم‌پذیری مقایسه شود تا بهترین مدل انتخاب شود. در این مرحله از شاخص‌های مختلفی استفاده خواهد شد.

مرحله ۶: پیاده‌سازی:

پس از انتخاب بهترین مدل، مدل باید در قالب مفهومی توضیح داده شود تا بتوان از آن در کاربردهای مختلف استفاده کرد. همچنین لازم به ذکر است که روش جمع‌آوری اطلاعات در این پژوهش کتابخانه‌ای بوده و مطالب به‌دست‌آمده از مطالعه و بررسی کتاب‌ها، مقالات، طرح‌های تحقیقاتی، پایان‌نامه‌ها و سایت‌های علمی و فنی به‌دست آمده است. از آنجا که هدف پژوهش حاضر بررسی یک مسئله فنی و مهندسی در یک زمینه خاص است، نه یک مطالعه تحلیلی و آماری، در این پژوهش نمونه‌گیری یا جامعه آماری خاصی وجود نداشته است، مگر در بخش پایانی و نتیجه‌گیری که به بررسی و مقایسه نتایج به‌دست‌آمده با نتایج قبلی می‌پردازد.

در نهایت، از نظر دامنه پژوهش، در ابعاد موضوعی، کاربردهای الگوریتم‌های داده‌کاوی در حوزه

تشخیص تخلف بررسی خواهد شد.

از نظر زمانی، پژوهش‌های بررسی‌شده مربوط به ۸ سال گذشته است، با این حال، بیشترین تأثیر از پژوهش‌های جدید خواهد بود. از نظر ابعاد مکانی، پژوهش هیچ محدودیت یا قلمرویی نخواهد داشت.