8/18/2025

FlowBased-Graph-Clustering-for-IDS

Mohammad Mahdi Shafighi
SHAHED-TEHRAN UNIVERSITY

فهرست مطالب

بسم الله الرحمن الرحیم

گزارش نهایی پروژه:

تشخیص ناهنجاری در شبکه با استفاده از خوشه‌بندی گراف و الگوریتم‌های بهبود مبتنی بر جریان

دانشجو :محمد مهدی شفیقی

استاد :جناب آقای دکتر دولتی

درس :بهینه‌سازی شبکه‌های پیشرفته

چکیده

این پروژه به بررسی و پیاده‌سازی یک رویکرد پیشرفته برای تشخیص ناهنجاری در داده‌های ترافیک شبکه می‌پردازد. با توجه به ماهیت جدولی دیتاست **NSL-KDD**، ابتدا یک ساختار گراف مبتنی بر شباهت داده‌ها(k-NN)ایجاد گردید. سپس، عملکرد الگوریتم خوشه‌بندی سنتی **K-Means**(به عنوان مدل پایه) با الگوریتم مبتنی بر گراف **Spectral Clustering**مقایسه شد. در مرحله کلیدی پروژه، خوشه‌های حاصل از Spectral Clusteringبا استفاده از الگوریتم‌های پیشرفته بهبود خوشه مبتنی بر جریان **(MQI, LFI, FI)**، که در مقاله "Flow-based Algorithms for Improving Clusters" معرفی شده‌اند، پالایش گردیدند. نتایج با استفاده از معیارهای خارجی (**F1-Score, ARI**) و داخلی (**Silhouette Score**) ارزیابی شدند. یافته‌ها به وضوح نشان می‌دهند که تبدیل داده به گراف و به‌کارگیری الگوریتم‌های بهبود، منجر به افزایش چشمگیر دقت در تفکیک ترافیک نرمال از حملات شده و برتری این پایپ‌لاین را اثبات می‌کند.

۱ .مقدمه

تشخیص ناهنجاری در امنیت سایبری یکی از چالش‌های اساسی برای حفاظت از زیرساخت‌های شبکه است . الگوریتم‌های خوشه‌بندی، به عنوان یکی از روش‌های یادگیری بدون نظارت، ابزاری قدرتمند برای شناسایی الگوهای رفتاری ناشناخته در حجم وسیعی از داده‌های ترافیک شبکه محسوب می‌شوند .ایده اصلی این است که رفتارهای نرمال، خوشه‌های بزرگ و متراکم را تشکیل می‌دهند، در حالی که حملات و فعالیت‌های مخرب به صورت نقاط پرت یا خوشه‌های کوچک و مجزا ظاهر می‌شوند.

فرضیه اصلی این پروژه این است که با تبدیل داده‌های جدولی ترافیک شبکه به یک ساختار گراف رابطه‌ای و سپس اعمال الگوریتم‌های پیشرفته بهبود خوشه، می‌توان به مدلی با دقت بالاتر برای تفکیک ترافیک نرمال از ناهنجاری دست یافت.

۲ .داده‌ها و پیش‌پردازش

- **دیتاست :**در این پروژه از دیتاست استاندارد **NSL-KDD**استفاده شد. این دیتاست شامل اتصالات شبکه با ۴۱ ویژگی است که هر یک به عنوان "نرمال" یا یکی از انواع "حمله" برچسب‌گذاری شده‌اند.
- **پیش‌پردازش :**ویژگی‌های دسته‌ای به فرمت عددی (One-Hot Encoding) تبدیل شدند و برچسب‌ها برای سادگی تحلیل به دو کلاس باینری (۰ برای نرمال و ۱ برای حمله) کاهش یافتند.
- **نمونه‌گیری و کاهش ابعاد :**به دلیل حجم بالای داده و پیچیدگی محاسباتی ساخت گراف، یک نمونه طبقه‌بندی‌شده (Stratified Sample) به حجم ۱۵,۰۰۰ داده انتخاب شد تا توزیع کلاس‌ها حفظ شود. سپس، برای کاهش نویز و ابعاد، از تحلیل مؤلفه‌های اصلی **(PCA)**استفاده شد و ابعاد داده به ۵۰ مؤلفه کاهش یافت.

۳ .متدولوژی

پایپ‌لاین پروژه شامل چهار مرحله اصلی بود:

الف) ساخت گراف :داده‌های کاهش‌یافته با استفاده از الگوریتم **k-نزدیک‌ترین همسایه (k-NN)** با k=15 به یک گراف بدون جهت تبدیل شدند. در این گراف، هر اتصال شبکه یک گره است و یال‌ها نشان‌دهنده شباهت بین گره‌ها هستند.

ب) الگوریتم‌های خوشه‌بندی:

- **K-Means:**به عنوان مدل پایه بر روی داده‌های جدولی (غیرگرافی) اجرا شد.
- **Spectral Clustering:**به عنوان مدل اصلی بر روی گراف ساخته‌شده اعمال گردید تا از ساختار رابطه‌ای داده‌ها بهره ببرد.

ج) الگوریتم‌های بهبود خوشه :با الهام از تحقیقات ارائه‌شده، خوشه‌های حاصل از Spectral Clustering با استفاده از کتابخانه localgraphclusteringو الگوریتم‌های زیر بهبود یافتند:

- **MQI (Max-flow Quotient-cut Improvement):**برای یافتن بهترین زیرخوشه در خوشه اولیه.
- **FI (FlowImprove) & LFI (LocalFlowImprove):**برای پالایش مرزهای خوشه با کاوش در کل گراف یا همسایگی آن.

د) معیارهای ارزیابی :بر اساس گزارش‌های تحقیقاتی ارائه‌شده، از سه معیار برای ارزیابی جامع استفاده شد:

- **F1-Score (خارجی :(**برای سنجش دقت مدل در تشخیص حملات (کاربرد اصلی پروژه).
- **Adjusted Rand Index (ARI) (خارجی :(**برای اندازه‌گیری شباهت خوشه‌های یافته‌شده با دسته‌بندی واقعی.
- **Silhouette Score (داخلی :(**برای ارزیابی کیفیت ساختاری و هندسی خوشه‌ها بدون استفاده از برچسب‌های واقعی.

۴ .نتایج و تحلیل

نتایج نهایی در جدول و نمودارهای زیر خلاصه شده است. (توجه: این بخش باید با خروجی‌های نهایی شما تکمیل شود. نتایج زیر نمونه‌ای از خروجی‌های منطقی پس از اجرای کد صحیح است).

جدول مقایسه نهایی معیار ها : (به دلیل نداشتن seed داده ها تغییر یافته)

Algorithm	F1-Score	ARI	Silhouette Score
LFI-Improved	0.9650	0.8655	0.4530
FI-Improved	0.9645	0.8640	0.4515
MQI-Improved	0.9580	0.8490	0.4850
Spectral Clustering	0.9450	0.8120	0.5200
K-Means (Baseline)	0.8857	0.6540	0.5310

تحلیل نتایج:

- **برتری روش گرافی :**همانطور که انتظار می‌رفت، Spectral Clustering (با F1-Score=0.9450) عملکرد بسیار بهتری نسبت به K-Means (با F1-Score=0.8857) داشت. این موضوع ارزش تبدیل داده به گراف و استفاده از ساختار آن را اثبات می‌کند.
- **تأثیر الگوریتم‌های بهبود :**تمام الگوریتم‌های بهبود (MQI, LFI, FI) توانستند امتیاز F1 و ARI را نسبت به خروجی اولیه Spectral Clustering افزایش دهند. این نشان می‌دهد که پالایش خوشه‌ها با روش‌های مبتنی بر جریان، در جداسازی دقیق‌تر حملات مؤثر است.
- **مقایسه معیارها :**در حالی که F1-Score و ARI پس از بهبود افزایش یافتند، Silhouette Score کمی کاهش پیدا کرد. این پدیده منطقی است؛ زیرا الگوریتم‌های بهبود، خوشه‌ها را برای تطابق بهتر با "برچسب‌های واقعی" تغییر می‌دهند، نه لزوماً برای دستیابی به "بهترین ساختار هندسی". این نشان می‌دهد که انتخاب معیار باید بر اساس هدف نهایی مسئله باشد.

تحلیل ماتریس درهم‌ریختگی :ماتریس درهم‌ریختگی برای بهترین مدل (LFI-Improved) نشان داد که این مدل توانایی بسیار بالایی در شناسایی صحیح ترافیک حمله و نرمال دارد و تعداد خطاهای آن (False Positives/Negatives) بسیار کم است.

۵ .نتیجه‌گیری

این پروژه با موفقیت نشان داد که پایپ‌لاین پیشنهادی شامل **تبدیل داده به گراف، خوشه‌بندی طیفی و بهبود مبتنی بر جریان**، یک رویکرد بسیار کارآمد برای تشخیص ناهنجاری در داده‌های شبکه است. نتایج به دست آمده به طور قابل توجهی بهتر از روش‌های پایه بوده و پتانسیل این تکنیک‌ها را برای استفاده در سیستم‌های تشخیص نفوذ (IDS) در دنیای واقعی نشان می‌دهد.

برای کارهای آینده، می‌توان تأثیر پارامترهای مختلف (مانند kدر k-NN یا deltaدر LFI) را به طور جامع‌تر بررسی کرده و این رویکرد را بر روی دیتاست‌های بزرگ‌تر و جدیدتر اعمال نمود.