Detection approach

Isolation Forest

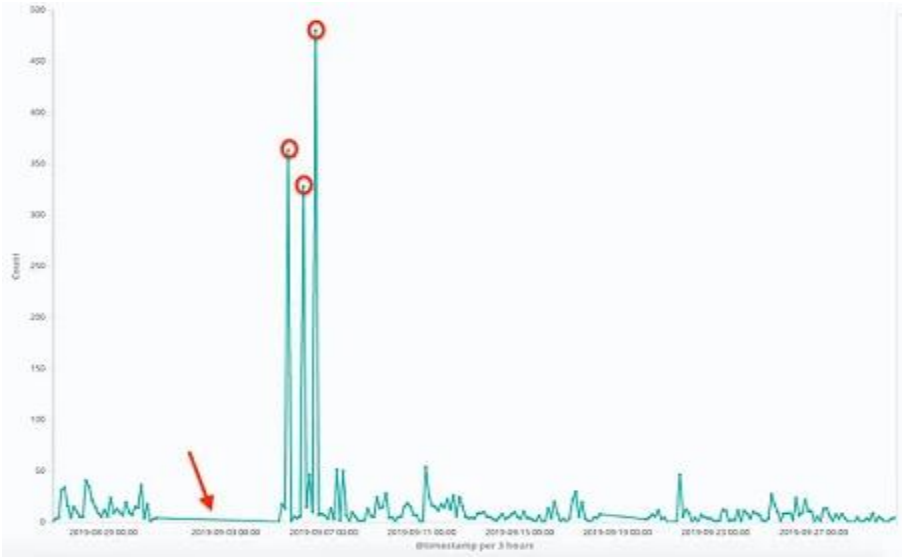
مقدمه

یافتن نقاط پرت (outliers) ،کاری است که باید تقریبا در هر گونه تحلیل‌های آماری صورت گیرد. وجود نقاط نامتعارف یا ناهنجار، باعث ایجاد خطا در مدل‌های آماری شده و پیش‌بینی را با مشکل مواجه می‌کند. به همین دلیل شناسایی آن ها در علم داده بسیار مهم است که این کار را با detection approach انجام می دهند. درادامه به معرفی یکی از انواع آن ها به نام isolation forest می‌پردازیم.

توضیحات کلی isolation forest

الگوریتم جنگل ایزوله (Isolation Forest) یا جنگل جداسازی، یک الگوریتم «یادگیری بدون نظارت» (Unsupervised Learning Algorithm) برای تشخیص ناهنجاری (Anomaly) است که برای جداسازی نقاط پرت (Outlier) به کار می‌رود. البته در اغلب روش‌های شناسایی نقاط پرت، بقیه نقاط که رفتار عادی دارند مورد ارزیابی قرار گرفته و براساس رفتار آن‌ها، نقاط پرت مشخص می‌شوند در حالیکه در الگوریتم جنگل ایزوله از ابتدا این گونه نقاط مورد بررسی قرار می‌گیرند. به ویژه در مجموعه‌های داده بزرگ و با ابعاد بالا، بسیار مؤثر عمل می‌کند. برخلاف بسیاری از الگوریتم‌های دیگر که سعی می‌کنند الگوی داده‌های نرمال را یاد بگیرند، Isolation Forest مستقیماً به دنبال جداسازی داده‌های ناهنجار است.

مثال



تاریخچه

در سال 2008 (Zhi-Hua Zhou) «و «ژوی هوآ ژو (Kai Ming Ting) «کای مینگ تینگ»ا لگوریتم جنگل ایزوله، ارائه کردند. آنان از دو ویژگی عددی داده‌های نامتعارف یا ناهنجار استفاده کردند.

-کم بودن تعداد این نقاط

-تفاوت زیاد مقادیر آن‌ها نسبت به مشاهدات هنجار

IForest از آنجا که ناهنجاری‌ها کم و متفاوت هستند، در مقایسه با نقاط عادی جداسازی آن‌ها ساده‌تر است. الگوریتم را برای مجموعه داده ایجاد می‌کند و ناهنجاری‌ها نقاطی هستند که بطور (iTrees) «مجموعه‌ای از «درختان جداسازی متوسط مسیر کوتاه‌تری نسبت به بقیه نقاط دارند.

در مقاله بعدی، که توسط همان نویسندگان در سال 2012 منتشر شد، مجموعه‌ای از آزمایشات را معرفی کردند تا نشان دهند که iForest دارای ویژگی‌های زیر است:

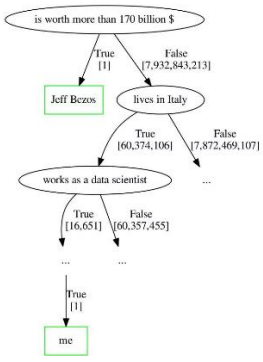
1.از پیچیدگی زمانی خطی کم و نیاز به حافظه محدود برخوردار است

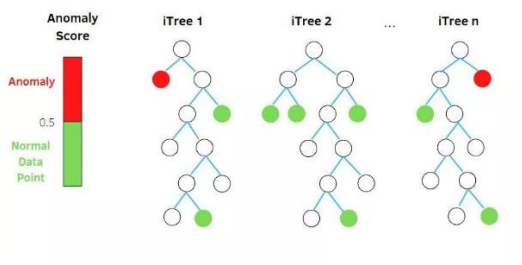
2.در مجموعه داده‌های با ابعاد بزرگ نیز قابل استفاده است

3.می‌توان الگوریتم را با استفاده از ناهنجاری‌های شناخته شده آموزش داد

4.بدون وجود آموزش مجدد می‌توان نتایج تشخیص با سطوح مختلف دسته‌بندی را ارائه داد.

یکي از مشکلات اصلي کاربرد iForest در تشخیص ناهنجاري، مربوط به خود مدل نیست، بلکه در روش محاسبه «نمره ناهنجاری» (Anomaly Score) است. این مشکل توسط «سهند حریری»، «ماتیاس کاراسکو» (Matias Carrasco Kind) و «رابرت برنر» (Robert J. Brunner) در مقاله‌ای که در سال 2018 منتشر کردند، مطرح گردید. آنان یک مدل بهبود یافته از iForest ارائه دادند که به نام «جنگل ایزوله توسعه یافته» (Extended Isolation Forest) یا به اختصار EIF شهرت دارد.





Isolation Forest ایده اصلی الگوریتم

+فرض اصلی این الگوریتم این است که داده‌های ناهنجار به‌راحتی قابل جداسازی هستند چون تعدادشان کم است و با سایر داده‌ها تفاوت دارند.

+الگوریتم با استفاده از درخت‌های تصمیم‌گیری تصادفی، داده‌ها را به‌صورت بازگشتی تقسیم می‌کند.

+هرچه یک داده زودتر در درخت ایزوله شود (یعنی مسیر کوتاه‌تری از ریشه تا برگ داشته باشد)، احتمال ناهنجار بودن آن بیشتر است.

مراحل اجرای الگوریتم:

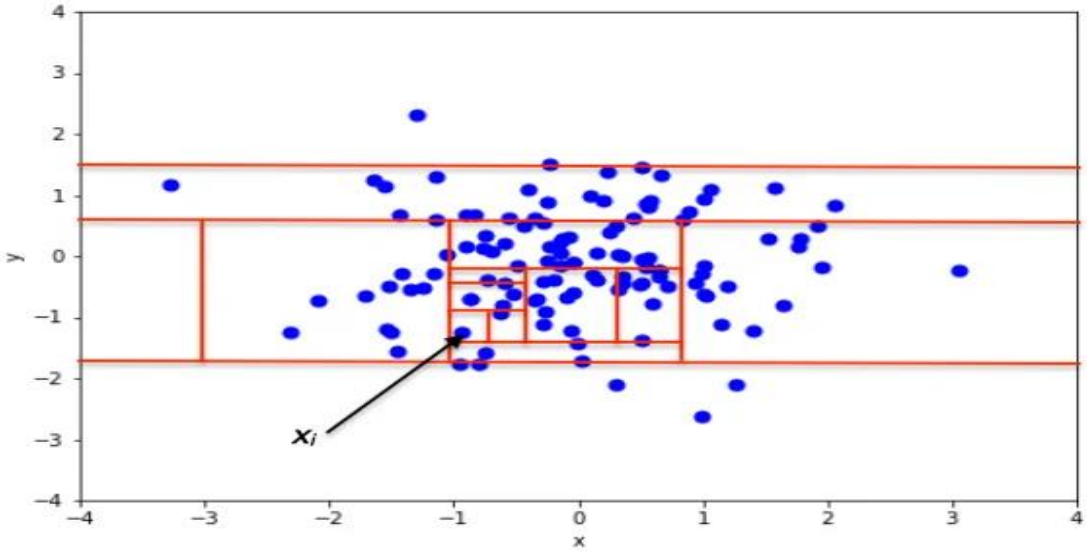
براساس الگوریتم جنگل ایزوله، تشخیص موارد غیر عادی و ناهنجار در مجموعه داده انجام شده که البته آسان‌تر از پیدا یا جداسازی داده‌ها یا نقاط نرمال یا هنجار است. به منظور جداسازی یک نقطه، الگوریتم به صورت بازگشتی با انتخاب تصادفی یک ویژگی، تقسیم‌بندی‌هایی را روی نمونه داده‌ها ایجاد می‌کند و سپس بطور تصادفی یک مقدار آستانه برای جداسازی یا تفکیک مقادیر به صورت هنجار یا ناهنجار، بین حداقل و حداکثر مقادیر مجاز برای آن صفت یا ویژگی، تعیین می‌کند

1.ساخت چندین درخت ایزوله با انتخاب تصادفی ویژگی‌ها و نقاط تقسیم .

2.محاسبه طول مسیر برای هر داده تا زمانی که ایزوله شود .

3.میانگین طول مسیرها در تمام درخت‌ها به‌عنوان نمره ناهنجاری استفاده می‌شود .

4.داده هایی با مسیر کوتاه تربه عنوان ناهنجار شناسایی می شوند.



.4

«تشخیص ناهنجاری با الگوریتم جنگل ایزوله»

:فرآیندی است که از دو مرحله اصلی تشکیل شده است

در مرحله یا گام اول، از مجموعه داده‌های آموزشی training setبرای ساخت itreesاستفاده می شود.

این کار بوسیله عملیاتی که در قبل اشاره شد، صورت می‌گیرد

در مرحله یا گام دوم، هر نمونه از مجموعه آزمایشی از طریق ساخت itreesدر مرحله قبل منتقل می شود و یک نمره ناهنجاری مناسب با استفاده از الگوریتم به آن اختصاص می یابد.

هنگامی که به تمام موارد موجود در مجموعه آزمایشی، نمره ناهنجاری اختصاص داده شد، می‌توان هر نقطه‌ای را که نمره آن بیشتر از یک آستانه از پیش تعریف شده باشد را به عنوان «ناهنجاری» در نظر بگیریم.

<<نمره ناهنجاری>>

الگوریتم محاسبه نمره ناهنجاری یک نقطه، مبتنی بر دیگر مشاهدات است بطوری که ساختار iTrees آن معادل با ساختار درختی جستجو دودویی (BST) است. به این معنی که رسیدن به یک گره خارجی از iTree معادل با یک جستجوی ناموفق در BST است. در نتیجه، برآورد میانگین طول مسیر h(x)برای رسیدن به گره‌های خارجی به صورت زیر محاسبه

می‌شود.به طوری که n تعداد داده‌های آزمایشی وm حجم نمونه وH ، »عدد هارمونیک (Harmonic Number) «است که بوسیله رابطه H(i)=ln(i)+γ محاسبه می‌شود. توجه دارید کهγ همان«ثابت اویلر-ماسکرونی (Euler-Mascheroni constant)« است.

$$c(m) = \begin{cases} 2H(m-1) - \frac{2(m-1)}{n} & \text{for } m > 2 \\ 1 & \text{for } m = 2 \\ 0 & \text{otherwise} \end{cases} \qquad s(x,m) = 2^{\frac{-E(h(x))}{c(m)}}$$

مقدار c(m) در بالا نشانگر میانگین h(x) است که برحسب m مشاهده نوشته شده، بنابراین می‌توانیم از آن برای نرمال سازی h(x) استفاده کنیم و تخمین نمره ناهنجاری را برای نمونه معین بدست آوریم. در رابطه بالا، E(h(X))امید ریاضی یا مقدار مورد انتظار مقادیر مختلف h(x) روی مجموعه درختان iTrees است .

نکات قابل توجه:

+اگر s به ۱ نزدیک باشد، می‌توان نتیجه گرفت که x با احتمال زیاد یک نقطه ناهنجار تلقی می‌شود.

+نزدیکی s به 0.5 بیانگر هنجار یا متعارف بودن نقطه است.

+اگر برای همه مقادیر یک نمونه تصادفی، امتیاز یا مقدار s به 0.5نزدیک باشد، باید انتظار داشت که همه نقاط هنجار بوده و مجموعه داده شامل ناهنجاری نیست.

برخی از ویژگی‌های الگوریتم جنگل ایزوله

زیر نمونه‌گیری :(Sub-sampling) از آنجایی که iForest نیازی به جداسازی همه نقاط متعارف عادی ندارد، می‌تواند بیشتر نقاط نمونه آموزشی (Training Set) را نادیده بگیرد. در نتیجه، iForestهنگامی که اندازه نمونه، کوچک باشد، بسیار خوب کار می‌کند. این ویژگی در بین الگوریتم‌های دیگر کمتر دیده می‌شود.

غرق شدن swamping: وقتی نمونه‌های متعارف و هنجار خیلی نزدیک به ناهنجاری‌ها هستند، تعداد بخش‌بندی‌های مورد نیاز برای جدا کردن ناهنجاری‌ها افزایش می‌یابد. این باعث می‌شود که iForest تفاوت بین ناهنجاری‌ها و نقاط عادی را به کندی و با تکرارهای زیاد تشخیص داده یا به طور کامل دچار واگرایی در پاسخ‌ها شود. به این معنی که با هر بار تکرار الگوریتم، نقاط متفاوتی را به عنوان ناهنجاری شناسایی می‌شوند.

پنهان ماندنmasking: هنگامی که تعداد ناهنجاری‌ها زیاد باشد، ممکن است که برخی از این نقاط در یک خوشه متراکم و بزرگ قرار بگیرند و جدا کردن ناهنجاری‌های را توسط الگوریتم درخت ایزوله دشوارتر کند. درست مانند مشکل غرق شدن، مشکل پنهان ماندن (Masking) نیز در زمانی که تعداد نقاط در نمونه زیاد باشد، رخ می‌دهد.

داده‌های چند بعدی با ابعاد بالا: یکی از اصلی‌ترین محدودیت‌های روش‌های مبتنی بر روش‌های استاندارد که برمبنای کار می‌کنند، ناکارآمدی آنها در برخورد با مجموعه‌های داده با ابعاد زیاد (Distance Function) محاسبه تابع فاصله است. دلیل اصلی این امر آن است که در یک فضای ابعاد بالا، نقاط تقریبا دارای فاصله یکسانی نسبت به یکدیگر هستند، بنابراین استفاده از یک اندازه‌گیری مبتنی بر فاصله بسیار ناکارآمد عمل می‌کند. متأسفانه، داده‌های با ابعاد بالا بر عملکرد نیز تأثیر می‌گذارند، اما می‌توان با افزودن یک آزمون به منظور انتخاب ویژگی‌های خاص و موثر، iForest تشخیص عملکرد را به شکلی تغییر داد تا ابعاد فضای نمونه کاهش یابد.

مشاهدات هنجار و متعارف:

الگوریتم جنگل ایزوله

(iForest) حتی اگر مجموعه آموزشی دارای فقط نقاط متعارف باشند نیز به خوبی عمل می‌کند. دلیل این امر آن است توزیع داده‌ها را به گونه‌ای توصیف می‌کند که مقادیر بزرگ برای طول مسیر h(xi) ملاک عمل است.

در نتیجه وجود ناهنجاری‌ها نسبت به عملکرد تشخیص iForest بی‌ارتباط است.

مزایا

بدون نیاز به برچسب :(Unsupervised) نیازی به داده‌های برچسب‌خورده نیست.

مقیاس‌پذیر و سریع: برای داده‌های حجیم و با ابعاد زیاد بسیار مناسب است.

مقاوم در برابر داده‌های پرت: حتی اگر داده‌ها کمی آلوده باشن، عملکرد خوبی داره.

محدودیت ها

حساسیت به انتخاب پارامترها مانند max_samples و contamination

تقسیم‌های تصادفی ممکن است در داده‌های بسیار پیچیده ناکارآمد باشند.

در برخی موارد تفسیرپذیری نتایج دشوار است.

تحلیل پیچیدگی و کارایی

یکی از بزرگ‌ترین مزایای Isolation Forest پیچیدگی زمانی آن است. ساخت هر درخت ایزوله به‌طور میانگین $O(n \log n)$ زمان می‌برد. اگر تعداد درخت‌ها را t در نظر بگیریم، کل جنگل در زمان $O(t \cdot n \log n)$ ساخته می‌شود. پیش‌بینی برای یک نمونه نیز تنها $O(t \cdot \log n)$ زمان نیاز دارد. این در حالی است که روش‌هایی مانند One-Class SVM پیچیدگی مربعی دارند و در داده‌های بزرگ عملاً غیرقابل استفاده می‌شوند.

توسعه‌های پیشرفته

از زمان معرفی اولیه، پژوهشگران نسخه‌های متعددی از iForest را توسعه داده‌اند :

Extended Isolation Forest (EIF): به‌جای تقسیم‌های محوری (محورهای مختصات) ، از تقسیم‌های مورب استفاده می‌کند تا داده‌های پیچیده‌تر را بهتر ایزوله کند.

Deep Isolation Forest (DIF):

ترکیب iForest با شبکه‌های عصبی برای ایجاد نمایش‌های غیرخطی از داده .

Time Series iForest:

نسخه‌ای برای داده‌های سری زمانی که وابستگی‌های زمانی را نیز در نظر می‌گیرد .

:OptiForest

بهینه‌سازی انتخاب آستانه‌ها برای افزایش دقت ایزوله‌سازی.

کاربرد ها

امنیت سایبری: شناسایی حملات و نفوذ در شبکه‌های کامپیوتری

مالی: کشف تقلب در تراکنش‌های بانکی

پزشکی: تشخیص داده‌های غیرعادی در آزمایش‌های بالینی یا سیگنال‌های زیستی

صنعت: پایش کیفیت محصولات و شناسایی خطاهای تولید

پیاده‌سازی عملی

```
python
from sklearn.ensemble import IsolationForest
import numpy as np
داده‌ی مصنوعی
X = np.random.randn(1000, 5)
X[990:] += 5  # افزودن ناهنجاری
مدل
clf = IsolationForest(nestimators=200, contamination=0.01, randomstate=42)
clf.fit(X)
پیش‌بینی
scores = clf.decision_function(X)
labels = clf.predict(X)  # -1 ناهنجار، 1 نرمال
```

منابع

- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. IEEE ICDM.

- Hariri, S., Kind, M., & Brunner, J. (2021). Extended Isolation Forest. ACM TKDD.

Scikit-learn Documentation: [IsolationForest](https://scikit-learn.org/stable -