رویکرد گراف محور برای تشخیص ناهنجاری در داده های متوالی و کار برد آن در بورس و هواشناسی

دنیا حمزئیان

دانشگاه صنعتی شریف

مقدمه

ناهنجاری ها یا داده های پرت داده هایی هستند که با توجه به سایر داده، به آن دنباله از داده ها تعلق ندارند. در این مقاله نوعی از داده های پرت بررسی می شوند که در ارتباط با دیگر داده ها پرت هستند. الگوریتم ارائ شده برای به دست آورن داده های پرت در سهام شرکت IMDB و در هواشناسی تحقیق شده است.

تعریف مسئله

ورودی: مجموعه ای از داده ها $i_1,i_2...i_n$

فرضیات:

- همه ی داده ها تعداد برابری ویژگی رفتاری[1] دارند.
- تمام داده ها تنها ۱ ویژگی متنی[2] دارند.
- تمام ویژگی های رفتاری متغیرهای پیوسته هستند.

خروجی:

برای مثال در سری زمانی ارزش سهام:

۱: قیمت ویژگی رفتاری است (ویژگی رفتاری می تواند بیش از ۱ عدد باشد). در اینجا از واژه ی ارزش استفاده می کنیم.

۲: زمان، متغیر ویژگی متنی است.

تعریف چند اصطلاح

- اسلاید: با استفاده از الگوریتم Sliding Window داده ها را بررسی می کنیم. از داده ی اول شروع می کنیم و بسته به اندازه ی اسلاید (CWS) که پارامتری است که خودمان تعیین می کنیم، بسته به نوع داده ها، داده هایی به تعداد اندازه ی اسلاید درون اسلاید قرار می گیرند و اسلاید هر بار به اندازه ی ۱ داده جلو می رود.
- داده ای که در بین داده های سمت چپ I در سری کمترین ارزش را داشته باشد و ویژگی رفتاری آن از I بیشتر باشد و با I در حداقل ۱ اسلاید آمده باشد؛ LSG آن نامیده می شود.
- تعریف LSG را اگر برای سمت راست L به کار بگیریم، آن داده را RSG آن می نامیم.
- داده ای که در بین داده های سمت چپ I در سری بیشترین ارزش را داشته باشد و ویژگی رفتاری آن از I کمتر باشد و با I در حداقل ۱ اسلاید آمده باشد؛ LLL آن نامیده می شود.
- تعریف LLL را اگر برای سمت راست I به کار بگیریم، آن داده را RLL آن می نامیم.
- داده ای را که در بین داده های سمت چپ ویژگی رفتاری اش با I برابرست و با آن در حداقل ۱ اسلاید آمده؛ LE می نامیم.
- اگر تعریف LE را برای سمت راست I به کار بریم، آن داده را LE آن می نامیم.

الگوریتم

ساختن گراف

داده ها برای بار دوم اسکن می شوند (با sliding window با طول اسلایدCWS).

۱. داده ها را براساس ویژگی متنی آن ها (که فرض کردیم یکی داریم) مرتب می کنیم.

۲. به هر داده یک راس نسبت می دهیم.

۳. اگر ۲ داده حداقل در ۱ اسلاید با هم آمده باشند، بینشان یال وجود دارد.

۴. وزن هر یال برابر با فاصله ی اقلیدسی میان ویژگی (های) رفتاری داده ها ی دو سر آن است.

خوشه بندی

از الگوریتم خوشه سازی کم وزن ترین درخت فراگیر استفاده می کنیم:

۱.کم وزن ترین درخت فراگیر را پیدا می کند.

۲. یال ها را بر اساس وزن مرتب می کند.

۳.سنگین ترین یال را حذف می کند.

۴. مولفه های همبندی را پیدا می کند.

۵. هر مولفه همبندی یک خوشه در نظر گرفته می شود و مقبولیت آن بررسی می شود.

۶. اگر خوشه مورد قبول باشد، خوشه را بر می گرداند اگر نه به مرحله ۳ بر می گردد.

تشخیص ناهنجاری

داده ها برای بار دوم اسکن می شوند (با sliding window با طول اسلایدVWS*).

۱. از داده ای که کمترین ویژگی متنی را دارد شروع می کنیم.

۲. به دست می آوریم که از هر خوشه چند داده در اسلاید جاری قرار گرفته.

۳. خوشه ای را که بیشترین تعداد داده را در اسلاید جاری دارد به عنوان خوشه ی محوری تعیین می کنیم.

۴. امتیاز ناهنجاری داده هایی که در خوشه ی محوری قرار ندارند را ۱ واحد اضافه می کنیم.

۵. اگر به آخرین اسلاید نرسیدیم اسلاید را جلو می بریم. اگر نه به مرحله ۶ می رویم.

۶.داده ها بابیشترین امتیاز ناهنجاری را به عنوان داده ی پرت معرفی می کنیم.

*پارامتر VWS با توجه به داده ها تعیین می شود.

ساختن کم وزن ترین درخت فراگیر

یال های متصل به داده های سمت چپ هر داده(i) با ارزش کمتر از LLL آن و ارزش بیشتر از LSG و یال های متصل به داده های سمت راست با ارزش بیشتر از RSG یا کمتر از RLL به کم وزن ترین درخت فراگیر تعلق ندارد. همین طور یال هایی که به سمت داده هایی با ارزش برابر با LSG, RSG, LE,RE,LLو RLL را هم می توان حذف کرد. *اثبات در مقاله

مقبولیت خوشه

معیار اول: اگر وزن یالی که در مرحله ی بعد قرار است حذف شود از EWT* بیشتر باشد، برنامه از حلقه خارج می شود.

معیار دوم: استفاده از پارامتر TWCV که در الگوریتم محاسبه می شود

*EWT با توجه به تفاوت وزن یال ها تعیین می شود.

توضیحات

- با دقت پایین VWS و CWS داده های پرت، محلی تعیین می شوند و در دقت بالای آن ها (افراطی) تمام داده ها داده ی پرت معرفی می شوند.
- زمان اجرا: O(N*VWS)

مراجع

- Graph-based approach for outlier detection in sequential data and its application on stock market and weather data, Feb 2014 pp. 1-9
- E.Hung, D.Cheung, Parallel algorithm for mining outliers in large database, in: Proceedings of IDC, Hong Kong, July 15-17 1999