
بسم الله الرحمن الرحیم

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر و فناوری اطلاعات

پایان نامه

برای دریافت درجه کارشناسی ارشد

در رشته مهندسی نرم افزار کامپیوتر

ارائه یک رویکرد جدید در تولید قوانین انجمنی بر اساس درخت الگوهای مکرر و روش خوشه بندی

وحید دهقان

استاد راهنما:

دکتر شهرام خدیوی

استاد مشاور:

دکتر احمد فراهی

اسفند/۹۰

دانشگاه پیام نور

بسمه تعالی

تصویب پایان نامه

پایان نامه تحت عنوان :ارائه رویکردی جدید در تولید قوانین انجمنی بر اساس درخت الگوهای مکرر و روش خوشه بندی

که توسط وحید دهقان در مرکز تهران تهیه و به هیئت داوران ارائه گردیده است مورد تأیید می‌باشد.

تاریخ دفاع: نمره: درجه ارزشیابی :

اعضای هیئت داوران:

نام و نام خانوادگی هیئت داوران	مرتبه علمی	امضاء
۱-	استاد راهنما	
۲-	استاد مشاور	
۳-	استاد داور	
۴-	نماینده تحصیلات تکمیلی	

تقدیم

این تحقیق را به رسم ادب به استاد بزرگ ادبیات آذربایجان، استاد شهریار، که منظومه حیدر بابایش همراه لحظه های شیرینی از زندگی‌ام بوده است تقدیم می‌کنم. روحش شاد و یادش گرامی.

سپاس گذاری

وظیفه خود می‌دانم سپاسگزار تمام آنهایی باشم که در این دوره ارزشمند بودنشان و امیدشان راه‌گشای من بود. پدر و مادر عزیزم و نیز اساتید عزیز و گران‌قدر دانشکده مهندسی کامپیوتر دانشگاه امیر کبیر، به خصوص جناب آقای دکتر شهرام خدیوی، همچنین جناب آقای دکتر فراهی، که با تلاش‌های بی‌شائبه خود نه‌تنها در انجام این پایان‌نامه مرا یاری نمودند و به هنگام نیاز برای حل مشکلات اینجانب از هیچ کمکی دریغ نورزیدند. برای ایشان آرزوی سلامتی، موفقیت و سر بلندی را دارم. کلیه همکاران و دوستان عزیزم در سازمان فن آوری اطلاعات و ارتباطات شهرداری تهران، واحد اطلاعات مکانی و نرم افزار، ونیز همکاران زحمت کش شرکت کنترل ترافیک شهرداری تهران به خصوص سرکار خانم مهندس نوشین سرور و خانم مهندس پریسا میر حسینی نیری و خانم مهندس مهدیه زنگی آبادی به جهت همکاری در اختیار داده های مورد نیاز تشکر و قدردانی می‌کنم.

چکیده

یافتن روابط معنی دار در اطلاعات خرید از موضوعات سابقه دار در داده کاوی است، که راه حل آن تحلیل و کاوش قوانین وابستگی[۱] است، در این تحقیق قصد داریم که رویکردی مناسب برای یافتن تکه مسیر هائی[۲] که ارتباط تنگاتنگی به لحاظ عبور و مرور بر هم دارند را در قالب قوانین وابستگی[۳] بررسی کنیم. تشخیص مسیرهائی که ترافیک آن‌ها بسیار به یکدیگر مرتبط و تاثیرگذاراست، منجر به ایجاد ترافیک روان و هدایت ترافیکی خواهد شد، این مسئله در گذشته معروف به دانش محدوده شهری بوده است. رویکرد ارائه شده توسعه ای است از روش CBAR[۴]، که بجای تقسیم پایگاه داده از روش خوشه بندی استفاده شده و الگوریتم الگوهای مکرر وظیفه تولید قوانین را بر عهده دارد. با توجه به نوع داده های موجود و نیاز به بررسی تأثیر مسیرها بر هم، استفاده از روش‌های معمول خوشه بندی میسر نیست، و لذا سازمانی به شکل یک گراف مورد نیاز است، پس از ساخت گراف ارتباطی تکه مسیرها، به دنبال روشهائی هستیم، که خوشه های واقعی گراف را در کمترین زمان و بهینه‌ترین تعداد مشخص کنند. با انجام تحقیقات در حوزه خوشه بندی گراف، الگوریتم های InfoMap[۵]، Blondel[۶] را برای خوشه بندی انتخاب می‌کنیم. گراف را با این دو روش خوشه بندی کرده و سپس قوانین وابستگی را از هرکدام از خوشه های بدست آمده استخراج می‌کنیم، برای تأثیر بر تشخیص بهینه خوشه‌ها علاوه بر معیاری که هر الگوریتم منتخب دارد، با توجه به تراکنش‌های واقعی، اقدام به وزن گذاری ارتباطات تکه مسیرها خواهیم نمود. در این روش مبتنی بر شبکه، علاوه بر اینکه نیاز به کاوش لیست‌های حجیم را کاهش داده و موجب تولید قوانین جذاب با سرعت بالا می‌شویم، بلکه مشکل روش‌های خوشه بندی که همانا تعیین پارامتر تعداد خوشه‌ها نیز هست را مرتفع می‌کنیم. با استفاده از شاخصه، پشتیبان[۷] و معیار قابلیت درک[۸]، روش مفید خوشه بندی که هزینه زمان و حافظه کمتری در مقایسه با روشهای دیگر دارد و قوانین مناسبی را استخراج می کند، معرفی می‌شود.

کلمات کلیدی:

قوانین وابستگی، خوشه بندی گراف، داده کاوی، داده های، ترافیک شهری، تشخیص کمونها

---

[۱] Association rule mining
[۲] Road Segment
[۳] Association Rules
[۴] Cluster Based Association Rule
[۵] الگوریتم اینفومپ، که بر اساس معیار قدم زدن تصادفی اقدام به خوشه بندی گراف می کند
[۶] الگوریتم بلاندل، که بر اساس معیار حداکثر ماژولاریتی اقدام به خوشه بندی گراف می کند
[۷] Confidence
[۸] Comprehensibility

فهرست مطالب

فهرست جداول

فهرست شکل‌ها

فصل ۱: کلیات تحقیق

۱-۱- مقدمه

مشکل ترافیک شهری در حال حاضر یکی از دغدغه های مهم بسیاری از دولت‌ها در کلان‌شهرهایشان می‌باشد که این مهم حتی با توسعه زیرساخت‌های شهری ازجمله توسعه بزرگراه‌ها و تونل‌ها به قوت خود باقی است، به نظر می‌رسد آنچه باعث به وجود آمدن ترافیک‌های سنگین می‌شود عدم توجه به مقوله هدایت ترافیکی و نداشتن طرح تخلیه مسیرهای مسدود در مواقع بحرانی است و این به دلیل عدم سنجش رفتار ترافیکی مسیرهای ارتباطی و پیش بینی بحران در حوزه ترافیک شهری است، یافتن مسیرهای موثر در شبکه معابر یکی از مسائل مهم و مزیتی برای طراحان شهری، ایستگاه‌های پلیس، و بسیاری از سازمان‌های درگیر در ترافیک شهری محسوب می‌شود. تشخیص مسیرهائی که ترافیک آن‌ها بسیار به یکدیگر مرتبط و تاثیرگذاراست، منجر به ایجاد ترافیک روان و هدایت ترافیکی خواهد شد، این مسئله در گذشته معروف به دانش محدوده شهری بوده است[Xiaoli et al,2007]. یکی از روش‌های مفید در حل این مشکل استفاده از روش‌های داده کاوی خصوصاً مسئله کشف قوانین وابستگی است. مسئله کشف قوانین وابستگی در پایگاه داده های حجیم از جمله مطالعات چند دهه گذشته بوده و روش‌ها، و الگوریتم‌های متنوعی بدین منظور توسعه داده شده‌اند، هدف ما در این تحقیق بدست آوردن این قوانین در داده های ترافیکی شهر تهران، با در نظر گرفتن بهینگی و غلبه بر پیچیدگی مسئله است، با توجه به حجم بالا و نوع رکوردهای تراکنشی موجود در این پایگاه داده، الگوریتم‌های موجود به طور مستقیم قادر به کشف این قوانین نیستند، لذا با تمرکز بر روی این کاربرد در این تحقیق قصد داریم رویکردی را جهت استخراج این قوانین با چیرگی بر مشکل I/O و تولید قوانین جذاب ارائه کنیم. برای این منظور سعی می‌کنیم از روش خوشه بندی داده‌ها استفاده نمائیم، در اکثر روش‌های خوشه بندی به دنبال معیارهای شباهتی هستیم که بتوانیم با توجه به آن معیار اشیاء را به گروه‌های همگنی تقسیم کنیم، الگوریتم‌های مختلفی برای این منظور توسعه داده شده‌اند که همگی آن‌ها نیاز به پارامترهائی دارند که از طرف کاربر وارد

می‌شود، تشخیص این مقادیر نیاز به دانش خاص در مورد حوزه کاربرد مورد نظر دارد و معمولاً تغییر این مقادیر نتایج متفاوتی را به وجود می‌آورد از طرفی تغییر این پارامترها در پایگاه داده های بزرگ، متضمن تحمل هزینه بالائی است. ازجمله معیارهای شباهت می‌توان به معیار مینکوفسکی، چبیشوف، منهتن و فاصله اقلیدسی اشاره کرد، با این حال در پایگاه داده مورد مطالعه این تحقیق، یعنی ثبت وقایع ترافیکی دو متغیر تکه مسیر و وضعیت ترافیکی وجود دارد که متغیر دوم سه حالت سنگین، بسیار سنگین و مسدود را شامل می‌شود، آنچه که در اینجا به عنوان معیار شباهت برای خوشه بندی مطرح می‌شود، ارتباط تکه مسیرهاست و نه مشابهت متغیرها، برای یافتن این ارتباط نیاز به تشکیل گراف تکه مسیرها وجود دارد، که با ساخت آن و تشخیص دسته گره هائی که چگالی ارتباط داخلی آن‌ها بیشتر است نسبت به سایر دسته‌ها اقدام به شناسائی این دسته گره‌ها خواهیم کرد، سپس با روشی که در ادامه توضیح خواهیم داد اقدام به استخراج رکورد از این دسته گره‌ها به فرم مورد نیاز برای کاوش قوانین وابستگی خواهیم کرد.

۱-۲- بیان مسئله تحقیق

تولید قوانین انجمنی یکی از پرکاربردترین و جذاب‌ترین روش‌های داده کاوی است که در سال‌های اخیر توجه زیادی را به خود جلب کرده است، و گستره ای از کاربردها را در برگرفته است. هدف از تولید قوانین وابستگی، کشف روابط وابستگی و معنی دار در یک مجموعه داده است.

کاوش قوانین وابستگی شامل دو زیر مسئله مهم است:

- یافتن مجموعه های مکرری که تکرار آن‌ها بیش از حد آستانه ای بنام پشتیبان در مجموعه داده ظاهر می‌شوند.
- تولید قوانین وابستگی با استفاده از این مجموعه داده های مکرر است.

اولین زیر مسئله نقش مهمی در کاوش قوانین بازی می‌کند. الگوریتم‌های کاوش قوانین وابستگی بعد از اینکه شخصی بنام آگراوال کراش برای اولین بار مسئله استخراج قوانین وابستگی از پایگاه داده

های تراکنشی را ارائه داد، توسعه یافتند[Agrawal et al,1997]. رویکرد این الگوریتم‌ها به دو بخش تقسیم می‌شوند:

- رویکرد تولید و بررسی کاندیدها [۹]
- رویکرد رشد الگوها[۱۰].

بخش اول شامل الگوریتمهائی همانند Apriori است، و بسیاری از مطالعاتی که بعداً روی این الگوریتم انجام شده است. به دلیل رجوع بیش از حد الگوریتم Apriori به پایگاه داده، در مواجهه با الگوهای مکرر[۱۱] و الگوهای بلند[۱۲] با مشکل سربار I/O مواجه است [Park et al,1995][Savaser et al,1995].

بخش دوم روش‌های رشد الگو هستند مانند FP_GROWTH.روش رشد الگو با استفاده از یک درخت، بجای تولید مجموعه کاندیدها، پایگاه داده را در یک درخت ذخیره می‌کند. و این درخت یعنی Fp-Tree را به صورت بازگشتی با استفاده از ساخت درخت‌های شرطی که به همان اندازه از اهمیت در تعداد همانند الگوهای مکرر هستند کاوش می‌کند. در مقایسه با رویکرد اول، رویکرد دوم بسیار کاراتر است اما نیاز به حافظه بیشتری برای نگهداری ساختار داده های میانی دارد. ساخت این حجم از درخت‌های شرطی در این رویکرد موجب می‌شود که این الگوریتم در پایگاه داده های بزرگ با حجم بیشتر از چند میلیون انعطاف پذیر نباشد.

در هردو رویکرد قبلی مشکلی بنام تولید قوانین بسیار و غیر جذاب را داریم، برای غلبه بر این مشکل در دو رویکرد ذکر شده و نیز مشکل مصرف حافظه و I/O قصد داریم از یک رویکرد جدیدی که تلفیقی است از روش‌های خوشه بندی و روش تولید قوانین وابستگی مبتنی بر درخت الگوهای مکرر استفاده کنیم. یکی از رویکرد های مواجهه با پایگاه داده هائی بزرگ خوشه بندی است. روش‌های خوشه بندی موجود نیاز به ورود پارامترهایی از طرف کاربر دارند که تعداد خوشه‌ها را مشخص می‌کند. با تغییر این پارامترها معمولاً خوشه های متفاوتی به وجود می‌آید که در غالب

[۹] Candidate generate and test
[۱۰] Pattern growth
[۱۱] Frequent pattern
[۱۲] Long pattern

اوقات نیز ساختار خوشه ذاتی را نمایان نمی‌سازند[Han et al,2006].

در اینجا سؤالاتی مطرح می‌شود که پایه این تحقیق است :

- رویکرد جدید در غلبه بر مشکل I/O چه رفتاری دارد؟
- رویکرد جدید در غلبه بر مشکل کمبود حافظه چه رفتاری دارد؟
- چطور می‌توانیم قوانین کمتر و قابل اتکائی تولید کنیم؟
- خوشه هائی که در روش خوشه بندی به وجود می‌آیند مسلماً شباهت زیادی بهم دارند ،این خوشه‌ها اگر مورد اعمال روش کاوش قوانین وابستگی قرار بگیرند چه خروجی هائی خواهند داشت؟
- آیا می‌توان با این تقسیم بندی سرعت عملیات را افزایش داد؟
- نویزها یا رکوردهای غیر مهم چگونه از بین می‌روند؟
- آیا ساخت درخت از روی داده های شبیه به هم یعنی خوشه هائی که تشخیص داده می‌شوند سریع‌تر است یا خیر؟
- آیا تلفیق روش خوشه بندی و تولید قوانین وابستگی برای تمام انواع داده‌ها عملی است؟
- تولید قوانین وابستگی با ویژگی‌های خصیصه های مجموعه داده های ترافیکی شهر تهران با روش جدید چگونه خواهد بود؟

۱-۳- معرفی موضوع تحقیق

ارائه رویکردی جدید در تولید قوانین وابستگی بر اساس درخت الگوهای مکرر و روش خوشه بندی، مورد مطالعه پایگاه داده ثبت وقایع ترافیکی شهر تهران است.

۱-۴- سابقه و ضرورت انجام تحقیق

اولین گام برای کشف قوانین وابستگی چهارچوب و بستر پشتیبانی-اطمینان است که توسط آگـراوال و همکارانش ارائه شده است[Agrawal et al,1997]. پرهزینه‌ترین قسمت به لحاظ زمان در الگوریتم کاوش قوانین وابستگی کشف مجموعه نمونه های بزرگ است. هاواسر و همکارانش الگوریتم تقسیم کننده ای را برای بهبود هرچه بهتر کارائی ارائه دادند که به طور کارآمدی تعداد رجوع به دیتـابیس را کاهش می‌داد، با این حال زمان زیادی که برای تولید مجموعه نمونه‌ها صـرف مـی‌کـرد قابـل توجـه بود[Savaser et al,1995]. پورک و همکارانش یک الگوریتم مفید بنام DHP[۱۳]، برای تولید مجموعه کاندید ابتدائی ارائه کردند، این روش به طور کارآمدی تعداد مجموعه های دو تائی کاندید را کنتـرل می‌کند و اندازه پایگاه داده را کـاهش مـی‌دهـد[Pork et al,1995]. چـی یونـگ و همکـارانش یـک الگوریتم توزیع شده بنام FDA [۱۴]را ارائه دادند که قوانین وابستگی را بـه صـورت کارآمـدی در یـک محیط توزیع شده کشف می‌کند[Cheung et al,1996]. یکی از الگوریتمهائی که در این زمینـه ارائـه شده روش CBAR است که قوانین وابستگی را مبتنی بر خوشـه هـائی کـه تعریـف مـی‌کنـد کـاوش می‌کند.

روش CBAR خوشه هائی را به فرم جـدول بـا اسـکن پایگـاه داده و سـپس خوشـه بنـدی کـردن رکوردهای تراکنش به k تا جدول به وجود می‌آورد که طول هـر جـدول k تاسـت و مجموعـه داده های بزرگ بر اساس تطابق با این جداول تولید می‌شوند، این کار نه تنها تمهیدی است بـرای هـرس کردن و کاهش تعداد داده‌ها، زمان مورد نیاز برای انجام خواندن داده [۱۵]نیز کاهش می‌یابد، همچنین به صحت نتایج تولید شده هم اتکای زیادی می‌توان کرد[Yuh-Jiuan et al,2005] .

یافتن الگوهای دسترسی کاربران به وب سایت‌ها از طریق لاگهـا کـاربردی اسـت از داده کـاوی کـه

---

[۱۳] در هم سازي و هرس کردن مستقیم
[۱۴] Fast Distributed Algorithm
[۱۵] Data scan

تکنیک آن را وب کاوی[۱۶] می گوییم، از جمله روشهائی که بر روی این نوع از داده‌ها انجام می‌شود، خوشه بندی این داده‌ها برای کشف گروههائی است که علائق مشترکی در بازدید از وب سایت‌ها دارند[Santhisree,2010] . این خوشه بندی با روش DBSCAN و تعریف مجموعه های شباهت از جمله کارهائی است که در حوزه خوشه بندی انجام شده است. استفاده از خوشه بندی تراکنش‌ها به عنوان یک مکانیزم پیش پردازش داده برای تولید قوانین وابستگی نادر، در توسط [Yun Sing Koh et al,2008] انجام شد، از این رویکرد برای خوشه بندی تراکنش‌ها قبل از کاوش قوانین وابستگی استفاده شده است و نشان داده شده که پیش پردازش مجموعه داده به وسیله خوشه بندی باعث می‌شود که هر خوشه وابستگی‌های موجود در خود را بدون دخالت از سایر گروههائی که الگوهای متفاوتی در رابطه دارند را ارائه کند. نتایج نشان داده که قوانین نادر تولید شده در هر خوشه گویا تر از قوانینی است که در کاوش مستقیم بدست آمده است. بنیاد و پایه خوشه بندی تراکنش‌ها به کاوش قوانین وابستگی بر می‌گردد که در مرحله بعد روی پارتیشن هائی که ضرورتاً جدا از دیگر پارتیشن‌ها هستند انجام خواهد شد، بعضی از الگوریتم‌های موجود به نوعی به آگاهی از دامنه مسئله تاکید دارند. بنابراین دامنه کاربرد آن‌ها به کاربرد های خاصی محدود می‌شود.

یکی دیگر از روشهائی که برای کاوش قوانین وابستگی، خوشه بندی را مرحله پیش پردازش مجموعه داده قرار داده، یافتن ارتباط بین خصیصه های دودویی[۱۷] در یک مجموعه داده بزرگ است. این مجموعه داده مربوط به اطلاعات بیش از ۸۰۰۰۰ وسیله نقلیه است که هر وسیله نقلیه بیش از ۳۰۰۰ ویژگی یا خصیصه دارد. برای کاوش قوانین وابستگی در چنین مجموعه داده ای نیاز به تعریف آستانه ای بنام کمترین پشتیبانی است که با در نظر گرفتن عددی کوچک برای این متغیر تعداد قوانین تولید شده سریعاً و به مقدار زیادی رشد می‌کنند، و برعکس، برای غلبه بر این مشکل ابتدا با روش خوشه بندی، خصیصه‌ها به گروه‌های همگن تقسیم، و سپس بر روی هر کدام از این گروها عملیات کاوش قوانین انجام شده است، این کار باعث کاهش شدید تولید قوانین اضافی شده است. در این

[۱۶] Web mining
[۱۷] Binary attribute

مجموعه داده که همانند یک ماتریس بزرگ و خلوت است برای خوشه بندی از روش خوشه بندی متغیرها استفاده شده است، برای خوشه بندی نیز از معیار های شباهت استفاده شده است که عامل کاربرد مورد نظر و متغیرها از عوامل تأثیرگذار برای انتخاب روش خوشه بندی است. در کلیه این روش‌ها تشخیص تعداد واقعی خوشه‌ها کاری دشوار است و آزمایشات متنوعی برای تعیین این پارامتر انجام می‌شود.

۱-۵- اهداف تحقیق

اهداف اصلی: ارائه رویکردی برای تشخیص صحیح خوشه‌ها بدون نیاز به پارامتر ورودی از طرف کاربر، ونیز تولید قوانین وابستگی در پایگاه داده های بزرگ به کمک خوشه بندی و الگوریتم‌های مناسب.

اهداف فرعی: تشخیص گلوگاه‌های ایجاد ترافیک شهری، جهت کمک به هدایت ترافیکی شهر تهران. ونیز ارائه چهارچوبی جهت کمک به حل معضلات ترافیکی با مطالعه رفتار حمل و نقل درون شهری. در مطالعه این رفتار از داده کاوی استفاده می‌شود.

۱-۶- فرضیه های تحقیق

فرضیه اصلی تحقیق:

هنگامی که پایگاه داده‌ها بزرگ هستند، بعضی اوقات اجرای الگوریتم‌های استخراج قوانین وابستگی غیر واقعی به نظر می‌رسد، غایت این مسئله تقسیم کردن آن به مجموعه ای از Projected DataBase[۱۸] هاست و سپس اجرای الگوریتم‌ها در هر یک از آن‌ها، آنچه که در این تحقیق مد نظر است بجای استفاده از این تقسیمات ، استفاده از خوشه‌هاست. وقتی خوشه ای به طور ذاتی متمایز از خوشه

[۱۸] پایگاه داده های افراز شده

دیگری است پس مجموعه های مکرر آن خوشه در خوشه های دیگر روی نخواهد داد لذا عمل اسکن روی خوشه های دیگر به نوعی اتلاف زمان و حافظه است. یکی از مشکلات موجود در روش خوشه بندی تعیین تعداد خوشه‌هاست. استفاده از الگوریتم خوشه بندی به طور مثال خوشه بندی مبتنی بر چگالی optics[۱۹]ساختار ذاتی خوشه های موجود در یک پایگاه داده را به ما می‌دهد و ما می‌توانیم قوانین وابستگی را از روی خوشه هائی که خود انتخاب می‌کنیم کاوش کنیم، با این حال در مسئله مورد نظر ما ، خوشه بندی بر اساس معیار شباهت انجام پذیر نیست، چرا که نوع داده‌ها بگو نه ای است که می‌بایست ارتباط آن‌ها سنجیده شود و نه شباهت. بدین منظور می‌بایست ارتباط مسیرها به صورت یک گراف مشخص شود و سپس مسیرهائی که بیشترین ارتباط عبور و مروری را با هم دارند دریک خوشه واقع می‌شوند. رکوردهای موجود در یک پایگاه داده بزرگ(بیش از یک میلیون رکورد) دارای شباهت‌ها و تفاوت‌های زیادی در درون خود هستند، لذا کاوش قوانین همبسته و وابسته از رکوردهای شبیه به هم سرعت بالائی در تولید قوانین دارند.

از الگوریتم‌های خوشه بندی گراف InfoMap وBlondel استفاده خواهد شد.

از الگوریتم FP-Growth برای کاوش خوشه های بدست آمده برای تولید قوانین استفاده می‌کنیم.

با توجه به نتایج متفاوت الگوریتم‌های منتخب، با بررسی اطلاعات حاصله در قالب قوانین، روش مناسب برای این کاربرد معرفی خواهد شد.

در بررسی اطلاعات حاصله از معیارهای تشخیص قوانین بهینه استفاده خواهد شد.

فرضیه فرعی تحقیق:

چرخه معماری CRISP شامل، تعاملات، فهم داده، آماده سازی داده، مدل سازی، ارزیابی، و گسترش.

پیش فرض این پژوهش استفاده از پایگاه داده اطلاعات ترافیکی شهر تهران است.

تعداد متغیر های بانک داده اطلاعات ترافیکی شهر تهران ۳ متغیر است.

پیش فرض این مجموعه داده حدود ۱۰۰۰۰۰۰ رکورد دارد.که از سال‌های ۸۴ تا ۸۹ جمع آوری شده است.

[۱۹] Ordering points to indentify the clustering structure