فصلنامه مهندسی حمل و نقل، دوره ۱۶، شماره ۳، بهار ۱۴۰۴، ص. ۴۶۶۶-۴۶۴۵

مقاله علمی-پژوهشی

ارائه رویکردی به منظور ارزیابی الگوی مشارکت کاربران در اطلاعات مکانی داوطلبانه بزرگراه‌ها با استفاده از یادگیری ماشین

سیده سعیده ساداتی، کارشناس ارشد، دانشکده نقشه‌برداری و اطلاعات مکانی، دانشکدگان فنی، دانشگاه تهران، ایران

رحیم علی‌عباسپور (مسئول مکاتبات)، دانشیار، دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی، دانشکدگان فنی، دانشگاه تهران، ایران

Email:abaspour@ut.ac.ir

علیرضا چهرقان، دانشیار، دانشکده مهندسی معدن، دانشگاه صنعتی سهند، تبریز، ایران

عباس عابدینی، استادیار، دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی، دانشکدگان فنی، دانشگاه تهران، ایران

دریافت: ۱۴۰۲/۱۰/۲۵ پذیرش: ۱۴۰۳/۰۲/۰۱

چکیده

بزرگراه‌ها از پرترددترین و حیاتی‌ترین عوارض خطی در سیستم حمل‌ونقل کشور هستند که با گسترش زندگی شهری روزبه‌روز بر اهمیت آن‌ها افزوده شده است. ازاین‌رو دسترسی به اطلاعات به‌روز و کارآمد در ارتباط با این عوارض نقش به‌سزایی در مدیریت سیستم حمل‌ونقل کشورها دارد. ازجمله این اطلاعات می‌توان اطلاعات مکانی داوطلبانه (Volunteer Geographic Information (VGI)) را نام برد که توسط کاربرانی ایجاد می‌شود که دانش تجربی یا محلی خود را از یک مکان یا موقعیت در پایگاه‌داده مکانی وارد می‌کنند. این مطالعه با ارائه رویکردی، روندهای ترسیم عوارض (نظیر بزرگراه‌ها و جاده‌ها) ایجاد شده توسط مشارکت‌کنندگان در پایگاه داده OpenStreetMap (OSM) را مورد بررسی قرار می‌دهد. همچنین یک بررسی آماری از الگوی مشارکتی کاربران به‌هنگام ترسیم شریان‌های ارتباطی انجام شده است و ویژگی‌های یک راه بر تعیین طبقه آن مورد ارزیابی تحلیلی قرار گرفته است. در رویکرد پیشنهادی به منظور طبقه‌بندی انواع کلاس‌های بزرگراه‌ها از الگوریتم طبقه‌بندی جنگل تصادفی استفاده شده است. به منظور ارزیابی رویکرد پیشنهادی از داده‌های خطی شهر تهران استفاده شده است. نتایج طبقه‌بندی با استفاده از ویژگی‌های آزیموت خطوط بزرگراه‌ها، طول ژئودزیک خطوط، فاصله اولین نقاط ترسیمی تا نزدیک‌ترین خیابان و تراکم kernelاولین نقاط به F-Score برابر ۷۱ درصد رسیده است. بررسی پارامترهای معنایی نظیر نام بزرگراه و نام کاربر نیز در این پژوهش انجام شد که نشان از عدم تأثیر آن‌ها بر دقت طبقه‌بندی دارد. پس از محاسبه اهمیت هر کدام از ویژگی‌های مورد استفاده، پارامتر طول ژئودزیک به‌عنوان مهم‌ترین عامل تأثیرگذار در رفتار کاربران مشارکت‌کننده در ترسیم انواع بزرگراه‌ها شناسایی گردید.

واژه‌های کلیدی: اطلاعات مکانی داوطلبانه (VGI)، پایگاه‌داده OSM، بزرگراه، الگوی مشارکت، الگوریتم طبقه‌بندی جنگل تصادفی

۱. مقدمه

از چند دهه پیش، روش‌های تولید اطلاعات مکانی به شدت تغییر کرده است. با در دسترس بودن فناوری‌های تعیین موقعیت ارزان، GIS منبع باز و وب جهانی به عنوان یک پلت فرم همکاری داوطلبانه کاربران مختلف در سرتاسر جهان، به منظور ایجاد و تولید اطلاعات مکانی در حال افزایش می‌باشد. این رویکرد از چند سال پیش با عنوان اطلاعات مکانی داوطلبانه (VGI) مطرح گردید [Goodchild, ۲۰۰۷]. در اثر انتقال نسل قابل توجه در شبکه جهانی وب (WWW) که منجر به تغییر استفاده از وب نیز شد، کاربران دیگر فقط به عنوان مصرف‌کنندگان محض محتوای از پیش تعریف شده عمل نمی کنند. در عوض، آنها بخشی از یک فرآیند به اشتراک‌گذاری دانش و اطلاعات هستند [O'reilly, ۲۰۰۷]. در طول سالهای گذشته، VGI به عنوان یکی از مرتبط‌ترین منابع داده‌های مکانی در مقیاس جهانی جهت گردآوری و نمایش دانش محلی شناخته شده است. علاوه بر داده‌ها، با ارائه اطلاعات مربوط به فعالیت‌های جامعه، گزارش‌های ارزشمندی در مورد پیشرفت‌های نقشه‌برداری بدست می‌دهند که می‌توانند برای تخمین کیفیت داده‌ها، درک فعالیت جوامع VGI یا پیش‌بینی پیشرفت‌های آینده مفید باشند [Rehrl ۲۰۱۶ & Gröchenig,]. امروزه، با افزایش استفاده از سیستم‌های اطلاعات مکانی (GIS) و تمایل بیشتر مردم به ارائه اطلاعات داوطلبانه مکانی، بکارگیری اطلاعات نه تنها در همه حالت‌ها یکپارچه شده‌اند، بلکه توسط کاربران، در زمان واقعی و در هر مکانی توسط گوشی‌های هوشمند در دسترس هستند [Attard et al.,۲۰۱۶].

برای بیش از یک دهه، OpenStreetMap (OSM) که در سال ۲۰۰۴ تأسیس شده است، موفق‌ترین پروژه اطلاعات مکانی داوطلبانه در سطح جهانی بوده است. پوشش مکانی، دقت هندسی، کامل بودن معنایی و قابلیت اطمینان کلی داده‌ها نیز از زمان ایجاد به شدت افزایش یافته است [Jokar Arsanjani OSM .et al., ۲۰۱۵] تعداد بسیاری از داوطلبان را قادر می سازد تا بدون هیچ محدودیتی عارضه‌های مکانی را ایجاد، ویرایش و حذف کنند.

اطلاعات مرتبط با شبکه راه‌های درون شهری نیز به عنوان یکی از اصلی‌ترین مولفه‌های زیرساخت اطلاعات مکانی مطرحند که کاربردهای بسیار وسیعی در حوزه مدیریت حمل‌ونقل شهری دارند[Chehreghan&Abaspour, ۲۰۱۹]. در این خصوص می‌توان به نقش موثر این اطلاعات در حوزه مدیریت حمل‌ونقل شهری ;[Alenouri et ۲۰۱۴];[Mirbaha et al., ۲۰۱۶] ; [Saberian et al.,۲۰۱۴al., و طراحی خطوط تاکسیرانی و اتوبوسرانی [۲۰۱۰] et al., [Teymourian ۲۰۱۴] ; Afandizadeh et. al, اشاره داشت. مطالعه و بررسی روندها و مشارکت کاربران می‌تواند نقش موثری در مدیریت موارد ذکرشده داشته باشند. تجزیه و تحلیل مشارکت‌ها در VGI می‌تواند دانش ارزشمندی را در مورد حمل‌ونقل جاده‌ای ارائه دهد. برای مثال داده‌های VGI را می‌توان برای استخراج و به‌روزرسانی شبکه‌های جاده ای در تصاویر سنجش از دور استفاده کرد [Manandhar et al., ۲۰۱۹]. همچنین داده‌های VGI را می توان برای شناسایی مناطقی با تراکم ترافیک بالا، شرایط جاده‌ای نامناسب و سایر مسائل مربوط به حمل‌ونقل استفاده کرد.

تحلیل و ارزیابی رفتار کاربران هنگام ترسیم بزرگراه‌ها و جاده‌های سطح شهر، جهت‌گیری‌های ذهنی افراد را در سیستم‌های حمل‌ونقلی آشکار می‌سازد. همچنین میزان مشارکت افراد در هر یک از دسته‌ها، مشخص می‌کند که کدام دسته از بزرگراه‌ها و جاده‌ها پرتردد و آشناتر برای شهروندان هستند. همین امر می‌تواند به مدیران این عرصه برای احداث، چگونگی احداث، نگهداری و تجهیز بزرگراه‌ها و راه‌ها در مناطق مختلف شهر کمک کند. از سوی دیگر ارزیابی مشارکت داوطلبان می تواند یکی از شاخص‌های غیرمستقیم و ضروری برای تأیید کیفیت داده‌ها در تولید نقشه‌ی شریان‌های ارتباطی باشد. بنابراین،

درکی روشن از نحوه رفتار و دریافت بازخورد هر کاربر در VGI بسیار مهم است. با این حال شناسایی و بررسی رفتار کاربران مشارکت‌کننده OSM در ترسیم و ویرایش عوارض خطی مورد مطالعه قرار نگرفته است. به طور خلاصه، تجزیه و تحلیل مشارکت‌ها می‌تواند به بهبود دقت نقشه‌های راه، نمایش شبکه‌های مشارکتی کاربران، شناسایی مسائل مربوط به حمل‌ونقل و بهینه‌سازی جریان ترافیک کمک کند و منجر به حمل‌ونقل جاده‌ای کارآمدتر و ایمن‌تر شود. منظور از رفتار کاربر در ترسیم بزرگراه‌ها، سوگیری ذهن کاربر در ترسیم، اطلاعات و علم آن‌ها از انواع دسته‌بندی و اطلاعات تفضیلی بزرگراه‌های موجود است. ازاین‌رو در این تحقیق داده‌های بزرگراه‌های شهر تهران که از مشارکت کاربران در سایت OSM به‌دست‌آمده‌اند، مورد ارزیابی و تحلیل قرار می‌گیرند. مقصود از این پژوهش پاسخ به این سؤالات می‌باشد که مشارکت داوطلبان در رابطه با عوارض خطی الگوی تکرارشونده و مشابهی دارد یا خیر؟ و در صورت وجود الگو و روند در مشارکت‌ها، کاربران OSM چگونه در تولید و ویرایش اطلاعات مربوط به شریان‌های ارتباطی فعالیت داشته‌اند. این مطالعه الگوی مشارکتی کاربران OSM را با استفاده از پارامترهای استخراجی مرتبط با عوارض خطی مورد بررسی و مطالعه قرار می‌دهد. استفاده از پارامترهای مذکور و بررسی روند مشارکت برای عوارض خطی صرفاً در این پژوهش انجام گردیده است.

ابتدا پس از معرفی و بیان انگیزه و دلایل بررسی رفتار کاربران در ترسیم بزرگراه‌ها و جاده‌های شهری، مطالعات پیشین در حوزه‌های مرتبط به‌طور خلاصه مورد بررسی قرار می‌گیرند. سپس در بخش سوم رویکرد پیشنهادی این مطالعه شرح داده خواهد شد و رویکردهای موردنظر بر داده‌های انتخابی اعمال گشته و نتایج حاصل مورد بررسی قرار خواهند گرفت. در بخش آخر نیز گزارش مشروحی از روش‌های اعمال شده بر داده‌ها و نتایج و تحلیل‌ها، به همراه پیشنهادات برای تحقیقات آینده ارائه خواهند شد.

۲. ادبیات پژوهش

در رابطه با ارتباط میان حمل‌ونقل و اطلاعات مکانی داوطلبانه تولید شده توسط کاربران، Nash در سال ۲۰۰۹ پتانسیل برنامه‌های کاربردی وب ۲,۰ را برای مشارکت عمومی در برنامه ریزی‌های حمل‌ونقل بررسی کرد [Nash, ۲۰۰۹]. مطالعات علمی مختلفی برای تعیین اینکه آیا داده‌های مکانی تولید شده از طریق VGI می‌توانند برای اهداف حرفه‌ای مانند سایر نقشه‌های تولید شده توسط نقشه‌نگاران مورد استفاده قرار گیرند یا خیر، انجام شده است. بیشتر مطالعات بر روی ارزیابی دقت و کامل‌بودن داده‌های معنایی و هندسی در VGI متمرکز شده‌اند. تنها چند مطالعه به طور مستقیم رفتار داوطلبان را بررسی کرده اند [Mooney et. al, ۲۰۱۰] ; [Hacar,۲۰۲۰] ; [۲۰۲۲ ,Hacar] .

Neis and Zipf در مقاله‌ای مشارکت‌کنندگان OSM را بر اساس تعداد مشارکت‌هایشان در پروژه به سه گروه senior ،junior و nonrecurring تقسیم کردند. همچنین مشارکت‌کنندگان از جنبه‌های دیگری چون محل عضویت، حوزه فعالیت و محدوده زمانی فعالیت مورد بررسی قرار گرفتند و نتایج مختلفی از تجزیه و تحلیل‌ها در مورد تعداد اعضای ثبت‌نام‌شده و واقعاً فعال یک جامعه آنلاین VGI در این مقاله ارائه شده است. این تحلیل‌های پیشنهادی با تمرکز بر فعالیت‌های کاربر و نتایج جمع‌آوری‌شده در این مقاله می‌توانند پایه‌ای ارزشمند برای پاسخ‌گویی به سؤالات مربوط به کیفیت داده‌های VGI و سوگیری‌های مشارکتی موجود باشند [Neis &Zipf,۲۰۱۲]. مطالعه‌ای دیگر، مشارکت‌کنندگان اصلی را با ترجیحات نوع عارضه و انتخاب منطقه نقشه‌برداری به‌منظور ارزیابی کامل‌بودن داده‌ها در OSM مشخص می‌کند. این مقاله، مطالعه فرآیندهای نقشه‌برداری مشارکت‌کنندگان را برای درک ویژگی‌ها و کیفیت داده‌های تولید شده پیشنهاد می‌کند. نویسندگان این مقاله استدلال کرده‌اند که رفتار مشارکت‌کنندگان هنگام نقشه‌برداری، منعکس‌کننده انگیزه مشارکت‌کنندگان، ترجیحات فردی در

انتخاب عوارض و تعیین مناطق نقشه‌برداری شده است. چنین دانشی از رفتار مشارکت‌کنندگان می‌تواند به استخراج اطلاعات در مورد کیفیت مجموعه داده‌های VGI منجر شود [Bégin et al., ۲۰۱۳] . Fogliaroni و همکاران در سال ۲۰۱۸ نیز ابتدا مشکل ارزیابی کیفیت داده‌های VGI را فرموله می‌کنند، سپس مدلی برای سنجش قابلیت اعتماد اطلاعات و شهرت مشارکت‌کنندگان با تحلیل جنبه‌های هندسی، کیفی و معنایی ویرایش‌ها در طول زمان ارائه می‌کنند. نویسندگان مدلی را برای به‌دست‌آوردن امتیازات قابلیت اعتماد و شهرت، برای عارضه‌های مکانی و مشارکت‌کنندگان یک سیستم VGI ارائه کرده‌اند [Fogliaroni et al., ۲۰۱۸].

در مطالعه‌ی دیگری که توسط Forati and Ghose انجام شده است، استدلال شده که عوارض با تعداد نسخه‌های کم و کاربران با سابقه عملکرد محدود نیز بایستی در کنار دیگر اطلاعات VGI در نظر گرفته شوند، زیرا این موارد حاوی دانش محلی ارزشمندی هستند [Forati & Ghose, ۲۰۲۰] .

Hacar و همکاران در سال ۲۰۱۸، تکامل شبکه‌های جاده‌ای OSM را در آنکارا بین سال‌های ۲۰۰۷ و ۲۰۱۷، با استفاده از معیارهای مرکزیت بررسی کرده‌اند. آنها پارامتر کامل بودن زمانی، سینوسی بودن جاده‌ها و تراکم فعالیت داوطلبان را در طول سال ها اندازه‌گیری کردند. مشاهده شد که با افزایش تجربه مشارکت کنندگان، آنها مشارکت‌های دقیق‌تری داشتند [Hacar et ۲۰۱۸ al. , .

Hacar در سال ۲۰۲۰ داده‌های planet.osm را بررسی کرده و سپس برچسب‌های متعلق به جاده‌ها را مقایسه نمود. نویسنده روند افزودن برچسب‌های داوطلبان را نیز مطالعه کرده است. در حالی که برچسب‌های سطحی، مبدا و یک طرفه در جاده‌های مسکونی با نرخی مشابه سایر جاده‌ها اضافه می‌شد، برچسب‌های نام به دفعات اضافه شده است. همچنین مشخص شد که در ۸۱ درصد از نقشه‌های جاده‌های مسکونی، منبع استفاده شده مشخص نشده است. وی خاطرنشان کرد: در حالی که OSM منبع داده خوبی از نظر تنوع برچسب است، اما از نظر کامل بودن داده‌ها دارای کمبودهایی است [Hacar, ۲۰۲۰].

بر اساس مطالعات انجام شده تا به اینجا، این سؤال که نحوه ترسیم عوارض مختلف توسط کاربران چگونه است و آیا روند قابل‌شناسایی دارد یا خیر تنها برای عوارض ساختمانی توسط Hacar در سال ۲۰۲۲ بررسی شده است. این مطالعه روندها را در اولین تصمیمات مشارکت‌کنندگان هنگام ترسیم ساختمان‌های OSM بررسی می‌کند. رویکرد پیشنهادی ارزیابی می‌کند که خواص یک نقطه چقدر در تعیین اولین نقطه نقشه‌های ساختمان اهمیت دارد. انواع مجاورت ساختمان‌ها را با استفاده از طبقه‌بندی جنگل تصادفی برای خواص طبقه‌بندی می‌کند و به استنباط روند ترسیم از تأثیر نسبی هر ملک کمک می‌کند. برای آزمایش رویکرد، از گروه‌های ساختمانی جدا و متصل در استانبول و ازمیر ترکیه استفاده شد که نتیجه ۸۳ درصد F-Score بوده است [Hacar, ۲۰۲۲]. بسیاری از مطالعات صورت گرفته اطلاعات مکانی داوطلبانه را از منظر ارزیابی کیفیت و دقت بررسی نموده‌اند و روندهای مشارکتی کاربران تنها برای دسته‌های محدودی داده ساختمانی مورد بررسی قرار گرفته‌اند؛ بنابراین در این مطالعه با تمرکز بر عوارض خطی به‌ویژه عوارض دارای تگ highway به بررسی روندهای ترسیمی در کلان‌شهر تهران پرداخته می‌شود. هدف این مطالعه یافتن جهت یا روند مشترک در میان مشارکت‌کنندگان OSM هنگام نگاشت بزرگراه‌ها است. یک عارضه خطی از اتصال نقاط به دست می‌آید؛ ازاین‌رو سعی شده است با استفاده از ویژگی‌های مختلف خطوط روندهای متمایز شناسایی شود و نقش هر ویژگی نیز در ترسیم خطوط مشخص گردد.

۳. روش پژوهش

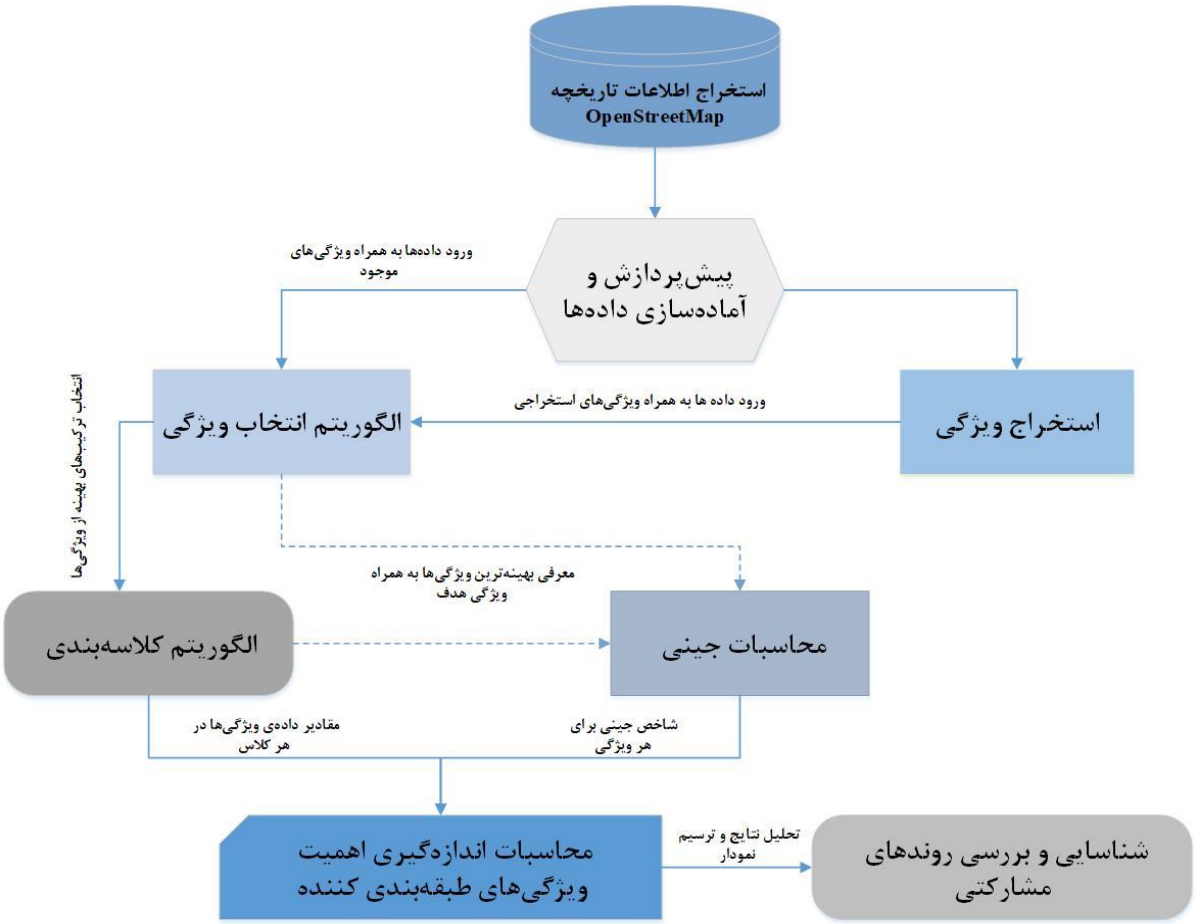
رویکرد پیشنهادی این مطالعه از ویژگی‌های هندسی نقاط تشکیل‌دهنده یک راه برای ارزیابی رفتار ترسیمی استفاده می‌کند. شکل (۱) روند اجرای روش پیشنهادی را به‌طور خلاصه نشان داده است. در این راهکار ابتدا داده‌های تاریخچه مشارکتی

کاربران از پایگاه داده استخراج می‌شود. در ادامه پس از اعمال پیش‌پردازش بر داده‌ها، ویژگی‌های مرتبط با شریان‌های ارتباطی از داده‌ها استخراج می‌گردند. سپس داده‌ها با استفاده از ویژگی‌های استخراجی از مرحله پیشین، مورد طبقه‌بندی قرار می‌گیرند تا روندهای مشارکتی کاربران در طبقات مختلف بزرگراه‌ها مورد بررسی و تحلیل قرار گیرند.

۳-۱ استخراج اطلاعات تاریخچه

برای اجرای رویکرد پیشنهادی از داده‌های موجود در پایگاه‌داده OSM استفاده شده است؛ زیرا از اولین و موفق‌ترین پروژه های VGIاست. OSM دانش محلی را در اولویت قرار می‌دهد، جامعه‌محور و مبتنی بر داده‌های باز است و همچنین برای فعالیت های VGI محبوب است [Forati&Ghose, ۲۰۲۰] . OSM در چند سال گذشته رشد سریعی را تجربه کرده است، به‌طوری‌که تعداد کل کاربران ثبت‌نام شده در ۴ دسامبر ۲۰۲۳ از ۱۱ میلیون نفر گذشته است. چندین مطالعه نشان داده‌اند که تعداد بالای کاربران منجر به به‌روز و باکیفیت‌تر شدن مجموعه داده OSM می‌شود [Neis& Zipf, ۲۰۱۲] ; [Girres ۲۰۱۰] ; & Touya, [Haklay et al.,۲۰۱۰].

تاریخچه داده‌های OSM، نوعی از کلان‌داده[۱] هستند که حجم و تنوع بالایی دارند و وضعیت OSM را در یک نقطه زمانی خاص یا در یک دوره زمانی معین منعکس می‌کنند. این شامل هندسه و ویژگی‌های عوارضی مانند گره‌ها، راه‌ها و روابطی است که توسط کاربران وارد پایگاه‌داده شده‌اند. این داده‌ها همچنین تغییراتی که به‌مرورزمان توسط کاربران در OSM ایجاد شده است را نیز شامل می‌شوند که شامل ایجاد، اصلاح، و حذف عارضه‌ها و همچنین فراداده ویرایش‌ها، مانند نام کاربری، مهر زمانی، شناسه تغییرات و شماره نسخه است. شکل (۲) نمونه‌ای از ذخیره‌سازی تغییرات برای عوارض مکانی را در داده‌های تاریخچه OSM نشان می‌دهند. همچنین مشارکت کاربران در طول زمان نیز از طریق این داده‌ها قابل بررسی است که این شامل تعداد و نوع ویرایش‌ها، میزان و کیفیت داده‌ها و سطح فعالیت و مشارکت کاربران است. بنابراین، تاریخچه داده‌های OSM را می‌توان برای تجزیه‌وتحلیل تکامل نقشه، مشارکت کاربران و تغییرات در عوارض استفاده کرد. فایل تاریخچه داده‌ها که تقریباً تمام داده‌های OSM را از ابتدای پروژه در بر می‌گیرد، در فرمت‌های Pbf, osm و XML قابل‌دسترسی هستند. داده‌های تاریخچه OSM چالش‌هایی را برای پردازش و تحلیل ایجاد می‌کنند، مانند حجم داده، فرمت داده، کامل بودن داده‌ها و ناهمگونی داده‌ها؛ بنابراین، به برخی فناوری‌های کلان‌داده مانند محاسبات ابری[۲]، پایگاه‌داده توزیع‌شده[۳] و پردازش موازی[۴] نیاز دارد تا امکان دسترسی و تحلیل داده‌ها را کارآمد و مقیاس‌پذیر کند؛ بنابراین این داده‌ها برای پردازش بایستی با اتکا به فناوری‌های موجود کلان‌داده پیش رود.



شکل ۱. مراحل اصلی رویکرد پیشنهادی

```
<?xml version='1.0' encoding='UTF-8'?>
<osm version="0.6" generator="osmconvert 0.8.5" timestamp="2019-11-18T01:00:00Z">
    <bounds minlat="35.6883" minlon="51.3676" maxlat="35.768" maxlon="51.488"/>
    <node id="25920662" lat="35.7005049" lon="51.3779759" version="5" timestamp="2010-09-11T06:54:17Z" changeset="5747077" uid="238107" user="vmoghiti"/>
    <node id="25920703" lat="35.7059246" lon="51.3774842" version="2" timestamp="2010-05-13T09:42:26Z" changeset="4683967" uid="7161" user="Hooman Mesgary"/>
    <node id="25920703" lat="35.7060086" lon="51.3773876" version="3" timestamp="2010-08-17T16:33:58Z" changeset="5518498" uid="238107" user="vmoghiti"/>
    <node id="25920706" lat="35.7142648" lon="51.3681797" version="3" timestamp="2012-03-02T15:33:49Z" changeset="10849493" uid="13908" user="kesler"/>
    <node id="25920706" lat="35.7142683" lon="51.368185" version="4" timestamp="2014-11-13T18:45:18Z" changeset="26762267" uid="437598" user="dmgroom_ct"/>
    <node id="25920706" lat="35.7142992" lon="51.3682322" version="5" timestamp="2014-11-22T22:44:51Z" changeset="26962352" uid="13908" user="kesler"/>
    <node id="25920706" lat="35.714211" lon="51.3682975" version="6" timestamp="2016-02-27T19:53:54Z" changeset="37488096" uid="3659986" user="myyourpo"/>
    <node id="25920706" lat="35.7141895" lon="51.3683306" version="7" timestamp="2016-06-23T11:52:58Z" changeset="40229896" uid="4068122" user="Esmail Doolabi"/>
    <node id="25920706" lat="35.7141895" lon="51.3683306" version="8" timestamp="2016-10-13T18:14:54Z" changeset="42877270" uid="4449060" user="Khalil Laleh"/>
    <node id="25920706" lat="35.7141895" lon="51.3683306" version="9" timestamp="2016-10-13T19:41:45Z" changeset="42879891" uid="4449060" user="Khalil Laleh"/>
    <node id="25920706" lat="35.7142625" lon="51.3681703" version="10" timestamp="2017-03-04T16:27:15Z" changeset="46576214" uid="3360520" user="madjidf"/>
    <node id="25920707" lat="35.7120243" lon="51.3704135" version="4" timestamp="2012-07-20T12:43:48Z" changeset="12376373" uid="722137" user="OSMF Redaction Account
    <node id="25920707" lat="35.7120603" lon="51.3704543" version="5" timestamp="2014-11-22T22:44:52Z" changeset="26962352" uid="13908" user="kesler"/>
        <tag k="created_by" v="JOSM"/>
    </node>
```

شکل ۲. نمونه داده‌های تاریخچه OpenStreetMap

۲-۳ پیش پردازش داده‌ها

داده‌های استخراجی در فرمت PBF و XML استخراج می‌گردند و برای پردازش بایستی به فرمت‌های رایج نرم‌افزارهای داده‌های مکانی مانند *.shp دربیایند. منطقه مورد مطالعه این پژوهش نیز باید از سایر مناطق جهان جدا شود، بنابراین نیاز به جداسازی داده‌ها به‌صورت منطقه‌ای است. تمرکز اصلی پژوهش بر داده‌های خطی است و داده‌های دریافتی شامل تمامی اطلاعات مکانی وارد شده به پایگاه‌داده OSM مانند نقاط POI[5]، خطوط، روابط و چندضلعی‌ها است، بنابراین باید عوارض خطی از سایر عوارض موجود در داده‌ها جدا شوند و سپس مورد بررسی قرار گیرند. ازآنجایی‌که داده‌ها دارای حجم بسیار بالایی هستند، روش‌های بهینه‌ای باید برای این امور استفاده گردد. به‌این‌علت که داده‌های ورودی به پایگاه‌داده توسط کاربران ترسیم می‌شوند و برچسب‌های متفاوت متناسب با آنها نیز توسط کاربران وارد

می‌گردد، لازم است داده‌های برچسب‌ها نیز مورد بررسی و بازبینی قرار گیرند. تعداد محدودی از برچسب‌های مرتبط با بزرگراه‌ها دارای مشکلاتی از قبیل املای نادرست دسته‌های بزرگراهی یا فونت‌های ناشناس و دسته‌های تعریف نشده بوده‌اند که از مجموعه داده‌های ورودی به الگوریتم‌ها حذف می‌گردند تا میزان خطای پردازش‌ها و نتایج نامرتبط و خارج از محدوده کاهش یابد.

برچسب‌های موجود در عوارض خطی دریافتی از پایگاه‌داده در جدول (۱) معرفی شده‌اند. تعدادی از این اطلاعات توسط OSM وارد پایگاه‌داده می‌گردند (مانند شناسه هر عارضه، شناسه هر تغییر، شماره نسخه، تاریخ و ساعت و شناسه کاربر) و سایر اطلاعات به‌صورت برچسب‌های دارای مقدار توسط کاربران ضمیمه می‌گردد (مانند طول عارضه، نام عارضه، نام کاربری و غیره). از آنجائیکه این تحقیق با تمرکز بر بزرگراه‌ها انجام شده است، دسته‌های مختلفی برای بزرگراه‌ها توسط WikiOSM تعریف شده است تا به‌عنوان برچسب توسط کاربران مورداستفاده قرار گیرد. انواع رایج و پرتکرار از دسته‌های معرفی شده توسط OSM که در این پژوهش مورد بررسی قرار گرفته‌اند، در جدول (۲) شرح داده شده‌اند. برخی از این تعاریف در کشورهای مختلف با یکدیگر متفاوت هستند و اطلاعات بیشتر در این باره برای مشارکت‌کنندگان، در وب‌سایت WikiOSM موجود است.

جدول ۱. برچسب‌های موجود در جدول اطلاعات توصیفی داده‌های OSM

پارامترهای داده‌های تاریخچه	تعریف
OBJECTID	شناسه اختصاص داده شده به هر داده
Full_id	شناسه کامل اختصاص داده شده به هر عارضه یکتا (شامل عدد و حرف)
OSM_id	شناسه اختصاص داده شده به هر عارضه یکتا (تنها عدد)
OSM_Type	مشخص‌کننده نوع عارضه
OSM_Version	شماره نسخه عوارض
OSM_Time	تاریخ ورود هر نسخه به پایگاه‌داده
OSM_uid	شناسه اختصاصی هر کاربر در OSM
OSM_user	نام کاربری هر کاربر OSM
Name_en	نام عارضه وارد شده توسط کاربر
OSM_Change	شناسه هر ویرایش/حذف/ترسیم در داده توسط کاربر
Shape_Length	طول عارضه وارد شده توسط کاربر
highway	دسته‌بندی بزرگراه وارد شده توسط کاربر

جدول ۲. معرفی کلاس‌های بزرگراهی در OSM

نوع بزرگراه	تعریف
unclassified	جاده‌های عمومی که در حومه شهر فقط برای ترافیک محلی مورداستفاده قرار می‌گیرند.
construction	به بزرگراه‌هایی که در حال ساخت هستند اطلاق می‌گردد.
cycleway	راه جداگانه‌ای را برای استفاده دوچرخه‌سواران نشان می‌دهد.
footway	برای مسیرهای کوچکی که عمدتاً یا منحصراً توسط عابران پیاده استفاده می‌شود، به کار می‌رود.
living_street	جاده‌ای با محدودیت سرعت بسیار کم و سایر قوانین ترافیکی مناسب برای عابر پیاده است. این نوع جاده‌ها در مقایسه با خیابان‌هایی که دارای برچسب‌های مسکونی هستند، حداقل سرعت کمتر و قوانین ترافیکی و پارکینگی ویژه‌ای دارند.
motorway	برای شناسایی جاده‌هایی با بالاترین عملکرد استفاده می‌شود. به‌طورکلی محدودیت‌هایی برای انواع وسایل نقلیه یا ترافیکی که می‌تواند در این نوع بزرگراه در جریان باشد؛ مانند عدم وجود عابر پیاده، دوچرخه، دام، اسب و غیره اعمال می‌شود.
motorway_link	برای جاده‌های پیوندی منتهی به motorway.

نوع بزرگراه	تعریف
pedestrian	جاده‌های عابر پیاده در مراکز خرید و مناطق مسکونی
primary	برای مشخص نمودن بزرگراه اصلی که شهرهای بزرگ را به هم متصل می‌کند، اما الزامات عملکرد یک motorway را برآورده نمی‌کند و واجد شرایط دسته trunk نیز نیست، استفاده می‌شود.
primary_link	جاده‌ای که یک بزرگراه primary را به یک بزرگراه secondary ، tertiary یا دیگر بزرگراه‌ها متصل می‌کند.
residential	برچسب بزرگراه مسکونی در جاده‌هایی استفاده می‌شود که دسترسی به/یااز مناطق مسکونی را فراهم می‌کنند و عموماً برای تردد محلی در داخل شهرک‌ها استفاده می‌شود.
road	یک جاده / راه / خیابان / بزرگراه از نوع ناشناخته است که می‌توان برای هر چیزی استفاده شود. این برچسب فقط باید به طور موقت استفاده گردد تا زمانی که جاده مذکور به‌درستی بررسی شود.
secondary	برای برچسب‌گذاری بزرگراه‌هایی که بخشی از مسیرهای اصلی نیستند، اما بااین‌وجود پیوندی را در شبکه مسیرهای ملی تشکیل می‌دهند، استفاده می‌شود. در کشورهای توسعه‌یافته معمولاً یک جاده آسفالته با حداقل دو خط است که معمولاً توسط یک خط مرکزی در جاده از هم جدا می‌شوند.
secondary_link	یک جاده که یک بزرگراه secondary را به یک بزرگراه tertiary ، طبقه‌بندی‌نشده یا دیگر بزرگراه‌های فرعی متصل می کند.
service	برای جاده‌های دسترسی به داخل شهرک صنعتی، کمپ، پارک تجاری، پارکینگ، کوچه‌ها و غیره استفاده می‌گردد. راه‌های خدمات معمولاً بخشی از شبکه عمومی خیابان نیستند و گاهی اوقات ممکن است برای عموم مردم غیرقابل‌دسترسی باشند.
steps	پیاده‌راه‌های شامل پله
tertiary	برای جاده‌هایی که سکونتگاه‌های کوچک‌تر را به هم متصل می‌کنند و در شهرک‌های بزرگ برای جاده‌هایی که مراکز محلی را به هم متصل می‌کنند استفاده می‌شود. از نظر شبکه حمل‌ونقل، جاده‌های tertiary معمولاً خیابان‌های کوچک را به جاده‌های اصلی‌تر متصل می‌کنند.
tertiary_link	جاده‌های پیوندی که به/از یک جاده tertiary از/به بزرگراه کلاس پایین‌تر منتهی می‌شوند.
trunk	برای جاده‌های با کارایی بالا یا بااهمیت بالا استفاده می‌شود که الزامات دسته motorway را برآورده نمی‌کنند، به‌عنوان بزرگراه primary نیز طبقه‌بندی نمی‌شوند
trunk_link	جاده‌های پیوندی که به/از یک جاده اصلی از/به یک بزرگراه دسته trunk یا بزرگراه کلاس پایین‌تر منتهی می‌شوند.

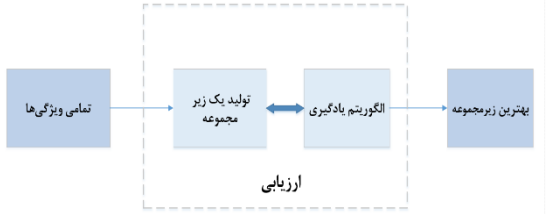
۳-۳ انتخاب ویژگی

ما اکنون در عصر کلان داده‌ها هستیم، جایی که حجم عظیمی از داده‌ها با ابعاد بالا در حوزه‌های مختلف در سراسر جهان استفاده می‌شوند. استفاده از تکنیک‌های داده کاوی و یادگیری ماشین برای کشف خودکار دانش از داده‌های مختلف مطلوب است. داده‌های با ابعاد بالا می‌توانند به طور قابل توجهی نیازهای ذخیره سازی حافظه و هزینه‌های محاسباتی را برای تجزیه و تحلیل داده ها افزایش دهند. این دسته از مسائل را می‌توان عمدتاً به دو جزء اصلی طبقه بندی کرد: استخراج ویژگی و انتخاب ویژگی. استخراج ویژگی، ویژگی‌های اولیه با ابعاد بالا را به یک فضای ویژگی جدید با ابعاد کم تبدیل می‌کند. فضای ویژگی جدید ساخته شده معمولاً ترکیبی خطی یا غیرخطی از ویژگی‌های اصلی است. از سوی دیگر، انتخاب ویژگی مستقیماً زیرمجموعه-ای از ویژگی های مرتبط را برای ساخت مدل انتخاب می کند [Guyon & Elisseeff, ۲۰۰۳] [Liu & ۲۰۰۷]. Motoda, هم استخراج ویژگی و هم انتخاب ویژگی دارای مزایای بهبود یادگیری در هر فرمت، افزایش کارایی محاسباتی، کاهش ذخیره سازی حافظه و ساخت مدل‌های تعمیم بهتر است.

از این رو، هر دو به عنوان تکنیک‌های موثر کاهش ابعاد در نظر گرفته می‌شوند.

داده‌های دنیای واقعی حاوی بسیاری از ویژگی‌های نامربوط، اضافی و خطادار هستند. حذف این ویژگی‌ها با انتخاب ویژگی، هزینه‌های ذخیره‌سازی و محاسباتی را کاهش می‌دهد در حالی که از از دست دادن قابل توجه اطلاعات یا کاهش عملکرد یادگیری جلوگیری می‌کند [Li et al.,۲۰۱۷]. این مرحله جهت شناسایی ویژگی‌های مفید و مرتبط انجام می‌شود. بنابراین از عمده‌ترین اهداف عملیات انتخاب ویژگی بر روی داده‌ها، می‌توان به کاهش حجم داده‌ها اشاره نمود. انتخاب ویژگی با کاهش مجموعه ویژگی‌ها، سبب صرفه‌جویی در وقت و هزینه در حین استفاده از داده‌ها و یا مرحله بعدی جمع‌آوری داده‌ها می‌گردد. همچنین این پردازش سبب بهبود عملکرد برای به دست آوردن دقت پیش‌بینی و درک داده‌ها، برای به‌دست‌آوردن دانش در مورد فرآیند تولید داده‌ها یا تجسم بهتر داده‌ها می‌گردد [Guyon & Elisseeff, ۲۰۰۳].

در این پژوهش از الگوریتم انتخاب ویژگی wrapper استفاده شده که یک رویکرد جستجوی حریصانه را با ارزیابی همه ترکیب‌های ممکن از ویژگی‌ها در برابر معیار ارزیابی دنبال می کند. به طور کلی، یک روش انتخاب ویژگی متعلق به یکی از دو خانواده اصلی filters و wrappers است [Liu & Yu, ۲۰۰۵]. یک رویکرد فیلتر سعی می‌کند مجموعه‌ای از ویژگی های برجسته را مستقل از الگوریتم‌های یادگیری پیدا کند، در حالی که یک رویکرد پوششی این کار را شامل یک الگوریتم یادگیری می‌کند و اگرچه رویکرد فیلتر از نظر محاسباتی ارزان است، عملکرد طبقه‌بندی آن در مقایسه با رویکرد پوششی ضعیف‌تر است [Dash & Liu,۱۹۹۷]. از طریق اعمال این مرحله ترکیب‌های بهینه‌ای از ویژگی‌های ورودی انتخاب گشته و معرفی می‌گردند. عملکرد الگوریتم پوششی در شکل (۳) نشان داده شده است. در این روش ویژگی‌های موردنظر به همراه داده‌ها به الگوریتم مذکور معرفی می‌گردند. الگوریتم در مرحله‌ی ارزیابی زیرمجموعه‌هایی از ویژگی‌های معرفی شده را تولید و وارد الگوریتم یادگیری ماشین می‌نماید تا دقت هر یک از زیرمجموعه‌های ساخته شده از ویژگی‌ها، محاسبه گردد. در نهایت ترکیبی از ویژگی‌ها که بالاترین دقت را بدست آورده‌اند، به عنوان ویژگی‌های انتخابی الگوریتم معرفی می‌گردند.



شکل ۳. روش انتخاب ویژگی wrapper

۳-۴ استخراج ویژگی

بر طبق تعاریف، داده‌ها با تعداد ثابتی از ویژگی‌ها نشان داده می شوند که می‌توانند باینری، موضوعی و یا پیوسته باشند [Guyon & Elisseeff,۲۰۰۳]. ویژگی به یک مشخصه یا خصوصیت قابل‌اندازه‌گیری از یک پدیده گفته می‌شود. از هر مجموعه داده ویژگی‌های مختلفی اندازه‌گیری می‌شود که به آن پروسه استخراج ویژگی گفته می‌شود. ویژگی‌های استخراج شده بردار ویژگی را برای یک مجموعه داده تشکیل خواهند داد. اگر از داده‌های خام دریافتی از پایگاه‌داده به‌طور مستقیم برای کلاسه‌بندی استفاده شود، علاوه بر کاهش عملکرد و دقت دچار افزایش پیچیدگی محاسباتی خواهیم شد؛ زیرا افزایش تعداد ویژگی‌ها باعث افزایش هزینه محاسباتی خواهد شد و همچنین مدل بایستی رابطه بین خروجی و ورودی با ابعاد بالا را بیابد. مطالعات نشان داده‌اند که در صورت مناسب بودن تمامی ویژگی‌های ورودی، می‌توان با افزایش تعداد داده‌های ورودی، با دقت بالاتری رابطه‌ی بین ورودی‌ها و خروجی را کشف نمود. ولی اکثر مواقع دسترسی به داده‌ها محدود است. به همین علت با استخراج ویژگی‌های مرتبط با این هدف که نماینده خوبی برای داده‌ی ورودی باشند، به افزایش دقت طبقه‌بندی کمک خواهد شد. برای بسیاری از برنامه‌هایی که داده‌های ورودی خام حاوی هیچ ویژگی قابل‌درک برای یک الگوریتم یادگیری خاص

نیستند، استخراج ویژگی ترجیح داده می‌شود. از سوی دیگر، ازآنجایی‌که استخراج ویژگی مجموعه‌ای از ویژگی‌های جدید را ایجاد می‌کند، تجزیه‌وتحلیل بیشتر مشکل‌ساز است؛ زیرا ما نمی توانیم معانی فیزیکی این ویژگی‌ها را حفظ کنیم. جهت بازدهی و دقت بالاتر الگوریتم، ویژگی‌های استخراجی بایستی رابطه مستقیمی با خروجی یا هدف داشته باشد، همچنین یک ویژگی باید مقادیر متغیری در بین کلاس‌ها داشته باشد. در حالت کلی ویژگی مطلوب بایستی واریانس درون کلاسی حداقل (بین نمونه های مشابه مقدار یکسان و یا نزدیک) و واریانس بین کلاسی حداکثر (بین نمونه‌های سایر کلاس‌ها مقدار متفاوت) داشته باشند. ساخت ویژگی یکی از مراحل کلیدی در فرآیند تجزیه‌وتحلیل داده‌ها است که تا حد زیادی موفقیت هر تحلیل آماری یا یادگیری ماشینی را بالا می‌برد [Guyon & Elisseeff, ۲۰۰۳].

بنابراین، برای افزایش دقت طبقه‌بندی داده‌ها در رویکرد پیشنهادی، تعدادی ویژگی با استفاده از داده‌های تاریخچه، استخراج و محاسبه می‌گردند. این ویژگی‌ها با استفاده از کدنویسی به زبان پایتون و نرم‌افزارهای Arcgis Pro و QGIS از اطلاعات هندسی موجود، استخراج می‌گردند و وارد مراحل بعدی از روش پیشنهادی خواهند شد. ویژگی‌های استخراجی در این پژوهش در جدول (۳) معرفی شده‌اند.

جدول ۳. پارامترهای استخراج شده از داده‌های تاریخچه OSM

پارامترها	تعریف
HubDistance	فاصله‌ی نقطه‌ی اول ترسیمی هر داده تا نزدیک‌ترین خیابان مجاور
$KernelDens_{firstpoint}$	تراکم هسته نقاط ترسیمی ابتدایی
Length_GEO	طول ژئودزیک خطوط (فاصله‌ی بین دو نقطه‌ی روی یک دایره یا کره جهانی)
Bearing	آزیموت خطوط (زاویه بین خطوط با شمال مغناطیسی)
$X_{firstpoint}$, $Y_{firstpoint}$	مختصات نقاط ترسیمی اول
$X_{lastpoint}$, $Y_{lastpoint}$	مختصات نقاط ترسیمی آخر
$X_{centroid}$, $Y_{centroid}$	مختصات مرکز خطوط
Distance	فاصله نقاط ترسیمی ابتدایی از مرکز ثقل خطوط
NU_points	تعداد نقاط تشکیل‌دهنده‌ی خطوط
NEAR_DIST	فاصله‌ی بین نقاط ترسیمی ابتدایی و نزدیک‌ترین عارضه نقطه‌ای یا خطی

۳-۵ الگوریتم طبقه‌بندی

ازآنجایی‌که دلیل طبقه‌بندی بر طبق این فرض که «اگر یک طبقه‌بندی موفق (دسته‌های بزرگراه‌ها) با معیارهای محاسبه‌شده با استفاده از ویژگی‌های خطوط یا نقاط تشکیل‌دهنده‌ی خطوط امکان‌پذیر باشد، رفتار مشترکی در نقشه‌برداری از هر یک انواع مختلف بزرگراه‌ها وجود دارد.» به‌عبارت‌دیگر، هر کلاس به درک روندهای خاص ترسیم بزرگراه مربوطه کمک می‌کند. بر اساس پژوهش انجام شده توسط [Basiri et al., ۲۰۱۶] نمرات برخی از طبقه‌بندی‌کننده‌های یادگیری ماشین اندازه‌گیری شد و طبقه‌بندی‌کننده‌های K-nearest neighbor و طبقه‌بندی جنگل تصادفی به ترتیب برای طبقه‌بندی‌های نوع مکانی و هندسی به‌عنوان مناسب‌ترین روش‌های طبقه‌بندی معرفی شدند. در این مطالعه نیز روش طبقه‌بندی random forest به‌عنوان روش طبقه‌بندی برای ویژگی‌های هندسی داده‌های آموزشی و برای طبقه‌بندی داده‌های آزمون استفاده شده است. طبقه‌بندی کننده جنگل تصادفی یک روش یادگیری ماشینی نظارت شده است که از درخت‌های تصمیم‌گیری متعدد برای طبقه‌بندی داده‌ها استفاده می‌کند. در این روش، داده‌های آموزشی حاوی مقادیر ورودی و هدف هستند. الگوریتم الگویی را انتخاب می‌کند که مقادیر ورودی را به خروجی تعمیم می‌دهد و از این الگو برای پیش‌بینی مقادیر در آینده استفاده می‌کند. جنگل تصادفی از چندین درخت تصمیم تشکیل شده است. درخت تصمیم یک ساختار فلوچارت مانند است که هر گره داخلی آن، یک تصمیم را بر اساس ارزش یک ویژگی خاص نشان می‌دهد و هر گره برگ نشان‌دهنده نتیجه تصمیم است. این روش از

تکنیکی به نام bootstrap aggregating استفاده می‌کند. این شامل ایجاد چندین زیرمجموعه تصادفی از داده‌های آموزشی است و هر درخت تصمیم بر روی یکی از این زیر مجموعه‌ها آموزش داده می‌شود. در ادامه برای هر گره درخت تصمیم، یک زیرمجموعه تصادفی از ویژگی‌ها در نظر گرفته می‌شود. علت این امر این است که از همبستگی بیش از حد آنها جلوگیری شود و همچنین این باعث ایجاد تنوع در بین درختان می‌شود. در ادامه الگوریتم وارد مرحله‌ای به نام میانگین‌گیری یا رأی‌گیری می‌گردد که برای انجام طبقه‌بندی، پیش‌بینی نهایی با اکثریت آرا در میان درختان منفرد تعیین می‌گردد. به‌عبارت‌دیگر، هنگامی که تمام درخت‌های تصمیم ساخته شدند، مرحله بعدی پیش‌بینی مجموعه تست (داده‌های آزمون) است. این کار با جمع‌آوری پیش‌بینی‌ها از تمام درختان جنگل با استفاده از مکانیزم رأی‌گیری انجام می‌شود. هر درخت به یک کلاس رأی می‌دهد و کلاسی که بیشترین رأی را داشته باشد به پیش‌بینی نهایی تبدیل می‌شود. پیش‌بینی نهایی میانگین پیش‌بینی‌های انجام شده توسط درختان است. پیش‌بینی‌های الگوریتم با مقادیر واقعی داده‌های آزمون مقایسه شده و دقت نهایی اجرای الگوریتم محاسبه می‌گردد. Random Forest دارای فراپارامترهایی است که می‌توان آن ها را تنظیم کرد، مانند تعداد درختان در جنگل، حداکثر عمق هر درخت و اندازه زیرمجموعه‌های تصادفی مورداستفاده برای آموزش. در این پژوهش از پارامترهای n-estiamtor و random state برای تعیین تعداد درخت‌های تصمیم مورداستفاده در مجموعه Random Forest و تصادفی‌سازی الگوریتم استفاده شده است.

مقدار بیشتر پارامتر n-estimator می‌تواند عملکرد تعمیم و استحکام بهتری را ارائه دهند، زیرا مجموعه تعداد درختان بیشتری تولید می‌کند؛ ولی این امر سبب افزایش زمان و هزینه محاسبات الگوریتم می‌گردد. اگر پارامتر random state در الگوریتم تنظیم نشده باشد، مدل هر بار که آموزش داده می‌شود، نتایج متفاوتی تولید می‌کند، حتی اگر از پارامترها و داده‌های آموزشی یکسانی استفاده شود. این الگوریتم یک روش ترکیبی است، به این معنی که پیش‌بینی‌های مدل‌های مختلف را برای بهبود دقت و کاهش بیش‌برازش ترکیب می‌کند. مزایای کلیدی این روش عبارت‌اند از: ماهیت غیرپارامتری آنها، دقت طبقه‌بندی بالا و قابلیت تعیین اهمیت پارامترها [Pal , ۲۰۰۵]. به علت ورود داده توسط افراد غیرمتخصص، تعداد داده‌های نویز و غیرمرتبط در مجموعه داده تاریخچه زیاد است، ازاین‌رو استفاده از روش طبقه‌بندی جنگل تصادفی منطقی و مفید است.

۳-۶ محاسبات اندازه‌گیری اهمیت ویژگی‌ها

الگوریتم جنگل‌های تصادفی سریع و انعطاف‌پذیر است و رویکردی قوی برای تحلیل داده‌هایی با ابعاد بالا نشان می‌دهد. مزیت کلیدی این الگوریتم نسبت به الگوریتم‌های یادگیری ماشین جایگزین، محاسبه معیارهای اهمیت متغیر است که می تواند برای شناسایی ویژگی‌های مرتبط یا انجام انتخاب متغیر استفاده شود [Nembrini et al., ۲۰۱۸]. یکی از دلایل مهم محبوبیت آنها، دردسترس‌بودن معیارهای اندازه‌گیری اهمیت[۶] است. پرکاربردترین معیارهای اندازه‌گیری اهمیت، اهمیت ناخالصی[۷] و اهمیت جایگشت[۸] هستند [Breiman, ۲۰۰۱]. ازآنجایی‌که شاخص جینی[۹] معمولاً به‌عنوان معیار تقسیم در درختان طبقه‌بندی استفاده می‌شود، اهمیت ناخالصی مربوطه اغلب به نام اهمیت جینی[۱۰] نامیده می‌شود. شاخص اهمیت ناخالصی به نفع متغیرهایی با نقاط تقسیم احتمالی زیاد است، یعنی متغیرهای طبقه‌بندی با دسته‌بندی‌های زیاد و یا متغیرهای پیوسته [Breiman,۲۰۱۷] ; [Strobl et al.,۲۰۰۷] و همچنین به نفع متغیرهای با فرکانس‌های دسته‌بندی بالا [Nicodemus, ۲۰۱۱] برطبق مطالعات صورت‌گرفته، برای داده‌هایی با ابعاد بالا، روش اهمیت جایگشت از نظر محاسباتی بسیار فشرده است. [Calle and Urrea, ۲۰۱۱] نشان دادند که رتبه‌بندی بر اساس ناخالصی VIM در مقایسه با مواردی که با اهمیت جایگشت به‌دست می‌آیند، می‌تواند در برابر آشفتگی داده‌ها قوی‌تر باشد. هرگونه طبقه‌بندی، ناخالصی معمولاً با

ناخالصی جینی اندازه‌گیری می‌شود که از فرمول پایه ناخالصی جینی[11] در رابطه (۱) استفاده می‌گردد [Breiman, ۲۰۱۷]:

$$\mathcal{T}(t) = \sum_{j=1}^{J} \emptyset_j(t)(1 - \emptyset_j(t)) \qquad (۱)$$

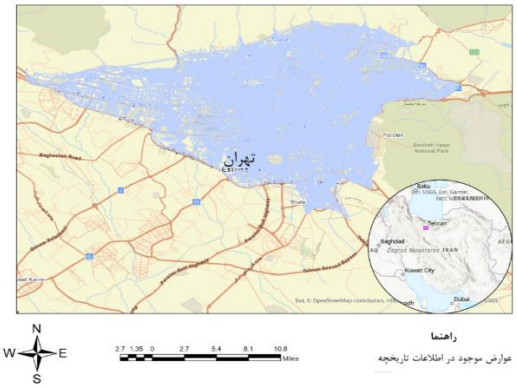
در رابطه (۱) $\emptyset_j(t)$ فرکانس کلاسی برای کلاس j در گره t است [Ishwaran, ۲۰۱۵] . پس از تعیین اهمیت هر معیار در فرآیند پیش‌بینی (PP) با استفاده از فرمول پایه ناخالصی جینی، اهمیت اندازه‌گیری در پیش‌بینی کلاس مربوطه (یعنی نوع بزرگراه) نیز با رابطه (۲) محاسبه می‌گردد.

$$MI_{\text{Classi}} = \text{Mean}_{\text{inclass}}(\text{X}_\text{i}) \times \text{MI}_{\text{i}'} \qquad (۲)$$

طبق رابطه (۲) MI_{Classi} اهمیت اندازه گیری i در پیش بینی نوع بزرگراه است، X_i مقدار مقیاس شده اندازه‌ی i ، و $MI_{i'}$ اهمیت اندازه‌گیری i در فرآیند پیش‌بینی (PP) است. این فرمول، اهمیت را با ضرب میانگین مقدار هر اندازه‌گیری در نوع بزرگراه پیش‌بینی شده، در اهمیت اندازه گیری مربوطه می‌یابد. در انتها روابط در کلاس‌ها شناسایی و تحلیل می‌شود.

۴. پیاده‌سازی و ارزیابی نتایج

برای ارزیابی رویکرد پیشنهادی این پژوهش داده‌های خطی مربوط به استان تهران، پایتخت ایران مورداستفاده قرار گرفت. نمایی از منطقه موردمطالعه در شکل (۴) ارائه شده است.



شکل ۴. منطقه مورد مطالعه به همراه داده‌های استخراج شده‌ی بزرگراه‌ها

برای استخراج داده‌های تاریخچه از وبسایت Planet.osm استفاده شد که تمامی داده‌های وارد شده به پایگاه داده OSM را به‌صورت فایل‌هایی با فرمت Pbf، osm و XML در اختیار افراد قرار می‌دهد. این داده‌ها به‌صورت هفتگی به‌روزرسانی می گردند. داده‌های این پژوهش در دی ماه سال ۱۴۰۱ با حجم ۱۲۰ گیگابایت در فرمت Pbf دانلود شد. برای پردازش، داده‌ها بایستی به فرمت‌های رایج نرم‌افزارهای داده‌های مکانی مانند *.shp تبدیل شوند. برای این منظور و همچنین استخراج داده‌های شهر تهران از سایر داده‌ها، از برنامه OSMconvert استفاده شد. بنابراین منطقه مورد مطالعه از سایر داده‌ها جدا و سپس داده‌های خطی منطقه نیز برای سایر مراحل استخراج شد. تعداد داده‌های مورد استفاده ۴۸۸۵۸۵ داده خطی با تمامی اطلاعات تاریخچه و تگ‌های مربوط است. داده‌های خام دریافت شده دارای اطلاعات نامرتبط نیز بود (حروف و اشکال بی‌معنی و تعریف نشده مانند # ، i ، À´Ã و...) که قبل از ورود به الگوریتم، پاک‌سازی شده و فقط داده‌های مرتبط استخراج گردید. تعداد کلاس و دسته بزرگراه‌های شرکت‌کننده در پژوهش ۲۰ کلاس است که در جدول (۴) تعداد عارضه‌های هر دسته از کلاس‌ها ذکر شده است.

جدول ۴. اطلاعات آماری داده‌های highway

شناسه کلاس کلاس	تعداد عوارض در هر کلاس	نام کلاس
۰	۹۹۳۸	unclassified
۱	۱۱۷۶	construction
۲	۱۱۶۸	cycleway
۳	۲۴۲۴۷	footway
۴	۴۰۹۱	living_street
۵	۱۱۴۱	motorway
۶	۱۲۶۰	motorway_link
۷	۱۱۰۴	pedestrian
۸	۱۸۶۹۶	primary
۹	۴۷۲۲	primary_link

شناسه کلاس کلاس	تعداد عوارض در هر کلاس	نام کلاس
۱۰	۲۲۰۳۷۲	residential
۱۱	۶۴۴۳	road
۱۲	۳۹۰۳۸	secondary
۱۳	۵۶۲۸	secondary_link
۱۴	۳۴۵۸۰	service
۱۵	۳۳۰۷	steps
۱۶	۴۳۸۶۲	tertiary
۱۷	۲۳۸۶	tertiary_link
۱۸	۱۷۱۷۷	trunk
۱۹	۱۴۴۸۰	trunk_link

اطلاعات توصیفی موجود در داده‌ها شامل اطلاعات کاربران، تاریخ، شناسه‌های اختصاصی کاربران و عوارض، شماره نسخه عوارض، اطلاعات ورودی کاربران و غیره است که به صورت برچسب توسط کاربران به عوارض ترسیمی اضافه می‌گردد. پس از پیش‌پردازش اولیه بر روی داده‌ها، در مرحله‌ی دوم پژوهش، پارامترها و ویژگی‌ها و برچسب‌های وارد شده توسط کاربران مورد بررسی قرار می‌گیرند تا در فاز بعدی وارد الگوریتم طبقه‌بندی گردند. برای بررسی تأثیر پارامترهای هندسی بر عملکرد طبقه‌بندی، مشاهدات گوناگونی صورت‌گرفته شد و نتایج از طریق دقت حاصل شده با یکدیگر مقایسه شدند. پارامترهای بررسی شده در مرحله اول شامل پارامترهایی بودند که در داده‌های دریافتی از پایگاه‌داده‌ی OSM موجود بود. در ابتدا با استفاده از داده‌های در دسترس برای کشف روندهای احتمالی موجود، الگوریتم اجرا گردید. ویژگی‌های انتخابی موجود در جدول شماره (۵) معرفی گردیده‌اند. ترکیبی از معیارهای معرفی شده، پس از انتخاب تعداد ویژگی موردنظر در الگوریتم feature selection، به‌دست آمده و سپس به الگوریتم طبقه‌بندی Random Forest داده شده که نتایج حاصل در جدول (۵) آورده شده است.

برای اجرای الگوریتم انتخاب ویژگی ابتدا بایستی داده‌های ورودی و ویژگی‌ها وارد الگوریتم شوند و سپس تعداد ویژگی‌های مدنظر برای اجرای طبقه‌بندی و ویژگی هدف معرفی گردند. این الگوریتم بهترین ترکیب از ویژگی‌ها را که بالاترین دقت را برای طبقه‌بندی ایجاد می‌کنند معرفی می‌نماید. ترکیب‌های گوناگون از ویژگی‌ها و تعداد ویژگی‌های انتخابی بررسی شده است. در ادامه ویژگی‌های انتخابی و ویژگی هدف وارد الگوریتم طبقه‌بندی جنگل تصادفی شده و پس از مشخص کردن داده‌های آموزشی و داده‌های آزمون، فرآیند طبقه‌بندی انجام شده و دقت آموزش محاسبه می‌گردد.

در ادامه برای افزایش دقت پژوهش، وارد مرحله استخراج ویژگی شده و پارامترهای مرتبط‌تری انتخاب و استخراج شدند تا مورد بررسی قرار گیرند. به‌ازای هر داده خطی، مقدار تراکم هسته نقطه‌ی ترسیمی اول (kernel density value)، فاصله‌ی هر داده‌ی خطی تا نزدیک‌ترین خیابان (hubdistance)، طول ژئودزیک خطوط، آزیموت، تعداد نقاط تشکیل‌دهنده خطوط و فاصله نقطه ترسیمی ابتدایی از مرکز ثقل خطوط علاوه بر مختصات نقاط ترسیمی ابتدایی و انتهایی و مرکز خط محاسبه گردید. علاوه بر پارامترهای معرفی شده در جدول (۳) دو پارامتر معنایی موجود در داده‌ها نیز برای بررسی تأثیر پارامترهای معنایی وارد عملیات طبقه‌بندی شدند که شامل Username (نام کاربری کاربران) و Streetname (نام بزرگراه‌ها) می‌باشد.

در رویکرد پیشنهادی پس از استخراج ویژگی، مرحله انتخاب ویژگی اعمال می‌گردد. در این مرحله نیز حالت‌های مختلفی از انتخاب ویژگی بررسی می‌شود تا تأثیر پارامترها مورد ارزیابی قرار گیرند. در ابتدا ویژگی‌های موجود در پایگاه‌داده OpenStreetMap (جدول (۱)) به الگوریتم معرفی گردید و با تغییر تعداد ویژگی‌های درخواستی الگوریتم انتخاب ویژگی اجرا گردید. برای بررسی بهتر پارامترها، تعداد ویژگی‌ها و حتی ویژگی‌های ورودی به الگوریتم انتخاب ویژگی در حالت‌های

مختلفی مورد آزمایش قرار گرفته شد. برای مثال مختصات نقاط استخراجی روندهای قابل‌توجهی را مشخص نمی‌کنند؛ بنابراین سعی شده است کمتر در طبقه‌بندی به‌عنوان ویژگی ورودی مشارکت داده شوند. حالت‌های گوناگون از ویژگی‌های انتخاب شده توسط الگوریتم انتخاب ویژگی، وارد الگوریتم طبقه‌بندی شده است. نتایج در جداول (۵) و (۶) گزارش شده‌اند. در ابتدا این عملیات با استفاده از یک ویژگی انتخابی و تنها بر روی ۱۰۰ داده‌ی نمونه (۸۰ داده آموزشی و ۲۰ داده آزمون) اجرا گردید. در ادامه این عمل بر روی تمامی داده‌ها اعمال گردید که شاهد کاهش دقت حاصل از طبقه‌بندی تمامی داده‌ها در برابر طبقه‌بندی داده‌های نمونه بوده‌ایم. بنابراین، ترکیبی از پارامترهای انتخابی به همراه تمامی داده‌های موجود وارد الگوریتم طبقه‌بندی شدند تا دقت حاصل اندازه‌گیری شود. نتایج این مرحله در جدول (۶) نشان داده شده است. برای اجرای الگوریتم طبقه‌بندی از نسبت رایج ۸۰ درصد داده آموزشی و ۲۰ درصد داده آزمون استفاده شد و همچنین random state و n_estimator به ترتیب ۴۲ و ۱۰۰ انتخاب شدند. بدین ترتیب الگوریتم از ۸۰٪ داده‌ها برای آموزش خود استفاده کرده و ۲۰٪ از داده‌ها برای ارزیابی دقت نهایی آموزش درنظر گرفته شده‌اند.

جدول ۵. دقت حاصل از طبقه‌بندی با پارامترهای اولیه

تعداد داده آموزشی	تعداد داده آزمون	ویژگی‌ها	ویژگی	دقت
۸۰	۲۰	Shape_length	highway	۰/۶۵۰
۳۶۳۸۳۳	۹۰۹۵۸	Shape_length	highway	۰/۵۰۴
۳۶۳۸۳۳	۹۰۹۵۸	OSM_Type	highway	۰/۵۰۱
۳۶۳۸۳۳	۹۰۹۵۸	OSM_Type, OSM_Version, OSM_uid	highway	۰/۵۰۳
۳۶۳۸۳۳	۹۰۹۵۸	Shape_length, OSM_Type, OSM_Version, OSM_uid , OSM_change	highway	۰/۴۹۳

جدول ۶. دقت حاصل از طبقه‌بندی با پارامترهای استخراج شده

داده آموزشی	داده آزمون	ویژگی‌ها	ویژگی	دقت
۳۶۳۸۳۳	۹۰۹۵۸	Bearing , Length_GEO , dist , $X_{firstpoint}$, $Y_{firstpoint}$, KernelDens , HubDistance	Highway	۰/۷۱۰
۳۶۳۸۳۳	۹۰۹۵۸	Bearing , Length_GEO , dist , $X_{firstpoint}$, $Y_{firstpoint}$, $X_{centroid}$, $Y_{centroid}$	Highway	۰/۷۰۵
۳۶۳۸۳۳	۹۰۹۵۸	Bearing , Length_GEO , dist , $X_{firstpoint}$, $Y_{firstpoint}$,	Highway	۰/۷۰۰
۳۶۳۸۳۳	۹۰۹۵۸	Bearing , Length_GEO , dist , $X_{firstpoint}$, $Y_{firstpoint}$, $X_{centroid}$, $Y_{centroid}$, Shape_length	Highway	۰/۷۰۱
۳۶۳۸۳۳	۹۰۹۵۸	Bearing , Length_GEO , dist , KernelDens, HubDistance, $X_{firstpoint}$, $Y_{firstpoint}$, NU_points, NEAR_DIST	Highway	۰/۶۸۷
۳۶۳۸۳۳	۹۰۹۵۸	Bearing , Length_GEO , HubDistance , $X_{firstpoint}$, $Y_{firstpoint}$, $X_{lastpoint}$, $Y_{lastpoint}$ KernelDens, Username	Highway	۰/۶۷۹
۳۶۳۸۳۳	۹۰۹۵۸	Bearing , Length_GEO , HubDistance , $X_{firstpoint}$, $Y_{firstpoint}$, $X_{lastpoint}$, $Y_{lastpoint}$ KernelDens, Streetname	Highway	۰/۷۰
۳۶۳۸۳۳	۹۰۹۵۸	Bearing , Length_GEO , HubDistance , $X_{firstpoint}$, $Y_{firstpoint}$, $X_{lastpoint}$, $Y_{lastpoint}$, KernelDens	Highway	۰/۷۱۱

در این مطالعه از انواع دسته‌بندی بزرگراه‌ها به‌عنوان متغیر وابسته برای اندازه‌گیری تأثیر هر یک از پارامترهای HubDistance ، $KernelDens_{firstpoint}$ ، LengthGEO ، Bearing استفاده شده است که بالاترین دقت طبقه‌بندی را در میان سایر پارامترها داشته‌اند و برای بررسی روندهای مشارکتی روابط معناداری را ایجاد می‌نمایند. بر اساس نتایج به‌دست‌آمده در جدول (۶)، پارامترهای معنایی دقت طبقه‌بندی را کاهش می‌دهند بنابراین حضورشان تاثیر چندانی در افزایش دقت نداشته است. بهترین ترکیب حاصل از پارامترهای استخراج شده که بالاترین دقت را ایجاد می‌کند، ترکیب Length_GEO، Bearing، $Y_{lastpoint}$ ،$X_{lastpoint}$ ،$Y_{firstpoint}$ ،$X_{firstpoint}$ ،HubDistance و KernelDens است. نمرات پیش‌بینی داده‌های آزمون با مقایسه کلاس‌های پیش‌بینی شده و انواع دسته بزرگراه‌های واقعی به‌دست آمد. درحالی‌که دقت precision را می‌توان ارزش پیش‌بینی‌کننده مثبت یا توانایی طبقه‌بندی‌کننده تعریف کرد، recall حساسیت و توانایی طبقه‌بندی‌کننده برای یافتن تمام نمونه‌های مثبت است [Pedregosa et al., ۲۰۱۱]. F-scoreیک میانگین هارمونیک وزنی از precision و recallاست، که امتیاز به بهترین مقدار در ۱ و بدترین مقدار آن در ۰ می‌رسد. نتیجه طبقه‌بندی در جدول (۷) آمده است.

ازآنجایی‌که امتیاز جینی کمتر به معنی تأثیر بهتر و بیشتر پارامتر در فرآیند پیش‌بینی طبقات است، طبق جدول (۷) و نتایج حاصل از محاسبات جینی، پارامترهای آزیموت و طول ژئودزیک خطوط بیشترین تأثیر را در پیش‌بینی‌های انجام شده داشته‌اند. در نهایت پس از انجام محاسبات نهایی، نتایج به‌دست آمده در نمودارهای شکل (۵) نمایش داده شده‌اند. محور افقی نمودارها ۲۰ کلاس بزرگراه‌ها را نشان می‌دهند که در جدول (۱) معرفی شده‌اند. محور عمودی اعداد محاسبه شده برای اهمیت هر یک از پارامترهای طبقه‌بندی‌کننده هستند.

جدول ۷. نتایج پیش‌بینی طبقه‌بندی

type	Precision (%)	Recall (%)	F-score (%)
highway	۷۱	۷۱	۷۱

در مرحله بعد از پژوهش، اهمیت هر یک از ویژگی‌های انتخابی در طبقه‌بندی از طریق رابطه (۲) محاسبه می‌گردد. برای این منظور ابتدا از طریق محاسبات جینی (طبق رابطه (۱)) میزان اهمیت هر معیار در فرآیند پیش‌بینی (PP) محاسبه می‌گردد. سپس اهمیت هر اندازه‌گیری در پیش‌بینی کلاس‌های هر دسته بزرگراه نیز محاسبه می‌گردد. نتایج بررسی اهمیت هر معیار ورودی در فرآیند طبقه‌بندی در جدول (۸) گزارش شده است.

جدول ۸. محاسبه امتیاز جینی پارامترهای طبقه‌بندی

نام ویژگی ورودی	امتیاز جینی
BEARING	۰/۱۸۹
length_GEO	۰/۱۲۴
HubDist	۰/۲۸۱
$X_{firstpoint}$	۰/۲۹۰
$Y_{firstpoint}$	۰/۲۹۳
$X_{lastpoint}$	۰/۳۰۹
$Y_{lastpoint}$	۰/۳۱۳
KernelDens	۰/۵۱۵

کلاس ۱۲ که secondary_ link نام دارد، نسبت به سایر پارامترها مقدار bearing یا آزیموت بیشتری پس از کلاس‌های motorway_link و living_street دارد که نشان می‌دهد این دسته از بزرگراه‌ها زوایای بزرگ‌تری داشته‌اند که هنگام ترسیم مورد توجه مشارکت‌کنندگان قرار گرفته است.

کلاس شماره ۶ کلاس motorway_link است که جاده‌های پیوند دهنده منتهی به یک بزرگراه هستند. در این کلاس مقدار آزیموت بیشتر از تمامی کلاس‌ها است که نشان از مقدار زاویه بزرگتر در ترسیم این بزرگراه‌ها و همچنین اهمیت این پارامتر در کلاسه‌بندی آن بوده است. این کلاس کمترین مقدار را نیز در پارامتر hubdistance دارد که نشان می‌دهد مشارکت‌کنندگان ترسیم این دسته از بزرگراه‌ها را از فاصله کمتری نسبت به خیابان‌های مجاورشان شروع کرده‌اند. مقدار چگالی نقاط

ترسیمی ابتدایی نیز در این کلاس پس از کلاس living_street دارای بیشترین مقدار است که نشانگر شروع روند ترسیم در مکان‌هایی با چگالی نقاط بالا است.

کلاس شماره ۵ نیز motorway نام دارد که بزرگراه‌هایی با ظرفیت بالا هستند. مقدار عددی محاسبه شده این کلاس برای پارامتر طول ژئودزیک بیشترین مقدار را دارد و مشارکت‌کنندگان ترسیم طولانی‌تری در این دسته از بزرگراه‌ها داشته‌اند. مقدار kernel density مربوط به این کلاس نیز کمترین مقدار را داشته است که نشان می‌دهد افراد تمایل دارند ترسیم این دسته از بزرگراه‌ها را از مناطقی که چگالی نقاط کمتر است شروع کنند. این کلاس دارای hubdistance بالایی نیز است که به مانند پارامتر چگالی، نشان‌دهنده تمایل ترسیم در مناطقی دورتر از سایر خیابان و بزرگراه‌های مجاور است.

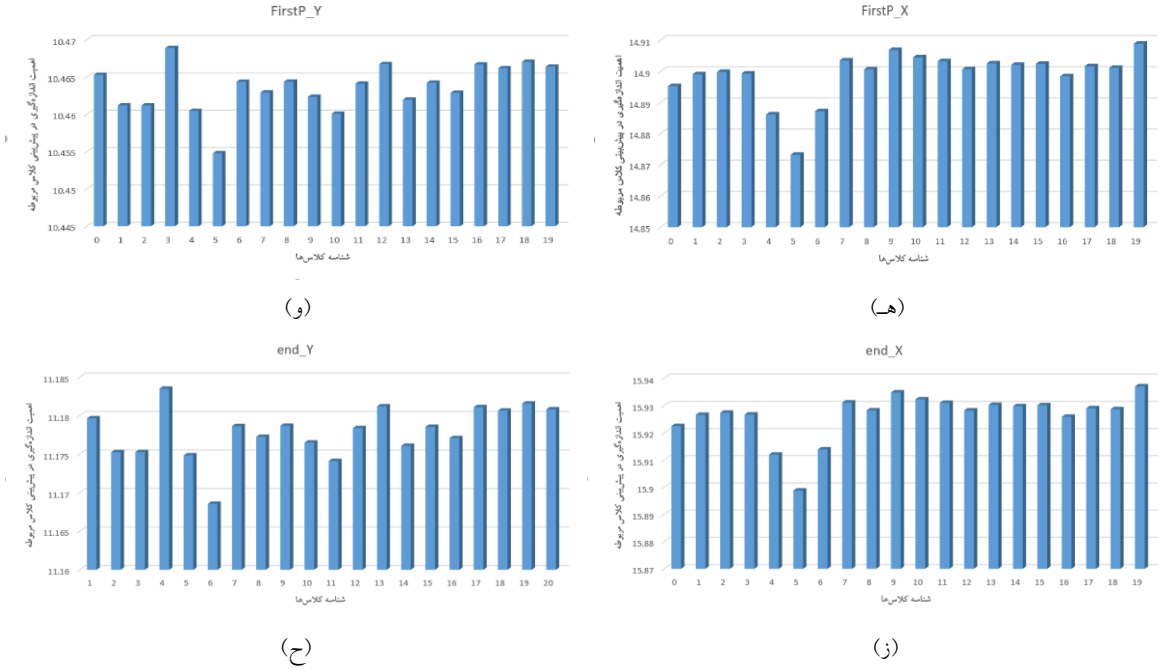
کلاس شماره ۴ نیز living_street است. پارامتر قابل اهمیت در این دسته از بزرگراه‌ها، پارامتر kernel density است که دارای بیشترین مقدار نسبت به سایر کلاس‌هاست. بنابراین شروع روند ترسیم این بزرگراه‌ها در مناطقی با چگالی نقاط بالا بوده است. از نظر آزیموت نیز این دسته دارای مقدار آزیموت بالایی پس از کلاس ۶ است که ترسیم در زوایای بیشتر را نشان می‌دهد.

کلاس بعدی که دارای تفاوت‌هایی نسبت به سایر کلاس‌ها است، کلاس شماره ۱۶ یا tertiary است. این دسته بیشترین مقدار hubdistance را داشته‌اند که نشان از بیشترین فاصله در ترسیم‌شان از بزرگراه‌های مجاور دارد. از نظر پارامتر طول نیز بزرگراه‌های tertiary (درجه سوم) پس از بزرگراه‌های motorway بیشترین مقدار را دارد که نشان‌دهنده اهمیت این پارامتر در این کلاس و ترسیم طولانی مشارکت‌کنندگان است.

کلاس بعدی، کلاس شماره ۱ است که construction نام دارد. این کلاس کمترین مقدار آزیموت را دارد که نشانگر روند ترسیم در زوایای کمتر نسبت به امتداد مبنا است. با توجه به اینکه امتداد مبنا جهت شمال در نظر گرفته می‌شود، مشارکت‌کنندگان بیشتر در حال ترسیم جاده‌های ارتباطی شمالی و شرقی در تهران هستند. همچنین این کلاس کمترین مقدار طول ژئودزیک را در بین کلاس‌ها به خود اختصاص داده است که نشان می‌دهد مشارکت‌کنندگان طول کمتری از این دسته بزرگراه‌ها را ترسیم کرده‌اند که ممکن است به علت عدم شناخت از بزرگراه‌های در حال ساخت یا عدم تکمیل ساخت آن‌ها باشد.

(الف) (ب)

(ج) (د)



(و) (هـ)

(ح) (ز)

شکل ۵. اهمیت هر یک از معیارها در پیش‌بینی ۲۰ کلاس ایجاد شده

۵. نتیجه‌گیری و پیشنهادها

ارزیابی روند ترسیم در بین راه‌های OSM چالش‌برانگیز است؛ زیرا راه‌ها در مقایسه با سایر عارضه‌های مکانی از تعداد زیادی نقطه تشکیل شده‌اند. همین امر پردازش داده‌ها را به علت حجم بسیار زیاد داده با معضل روبرو می‌کند. این مطالعه با استفاده از داده‌های خطی استخراج شده از پایگاه‌داده سایت OpenStreetMap انجام گردید و با تمرکز بر تگ highway طبقه‌بندی‌ها انجام گرفت. داده‌های خطی که توسط کاربران تولید می‌شوند با استفاده از تگ الصاقی highway در دسته‌بندی‌های مختلفی قرار می‌گیرند. هر یک از دسته‌های تعریف شده در تگ highway دارای تعاریف مختص خود هستند که با مراجعه به سایت OSM.Wiki می‌توان تعاریف و تصاویر هر کدام را مشاهده نمود. پس از طبقه‌بندی که با استفاده از ویژگی‌های استخراجی انجام می‌شود، نمرات پیش‌بینی با مقایسه کلاس‌های پیش‌بینی‌شده با کلاس‌های واقعی (دسته‌های highway) داده‌های آزمون، اندازه‌گیری می‌شوند. در این مطالعه از معیارهای تشابه رایج مانند فاصله نقطه به نقطه و نقطه به خط [Haklay et al., ۲۰۱۰] ، چگالی [Breiman,;۱۹۹۶ ۱۹۹۶] [Pal, ۲۰۰۵] و زاویه استفاده و درنهایت معیارهای منتخب با بالاترین دقت طبقه‌بندی انتخاب شده‌اند.

این مطالعه از انواع دسته‌بندی بزرگراه‌ها به عنوان متغیر وابسته برای اندازه‌گیری تاثیر هر یک از پارامترهای انتخابی استفاده می‌کند. به عبارتی انواع بزرگراه با استفاده از مقادیر (متغیرهای مستقل) پیش‌بینی می‌شوند، سپس با مقایسه انواع بزرگراه‌های واقعی با طبقات پیش بینی شده، نمرات پیش بینی به دست می آید. رویکرد پیشنهادی به امتیازات پیش‌بینی نتایج طبقه‌بندی وابسته است. در این مطالعه، پیش‌بینی بیشتر داده‌های موردی برای بررسی رفتار ترسیمی مشارکت‌کنندگان و استنتاج قابلیت اعتماد در نظر گرفته شده است. بنابراین، نمرات پیش‌بینی بیشتر از ۵۰ درصد نشان‌دهنده اطمینان بیشتر در رویکرد است [Hacar,۲۰۲۲]. در این پژوهش میزان دقت طبقه‌بندی ۷۱ درصد بوده است. سپس اهمیت معیارهای اشاره شده در پیش‌بینی انواع بزرگراه مشخص می‌شود. نمرات پیش‌بینی چگونگی اثرگذاری معیارها بر نتایج را مشخص نمی‌کند. از این رو برای مشخص شدن تاثیر هر معیار بر نتایج از فرمول پایه

ناخالصی جینی استفاده می‌گردد [Breiman, ۲۰۱۷] . پس از تعیین اهمیت هر معیار در فرآیند پیش‌بینی (PP)، اهمیت اندازه‌گیری در پیش‌بینی کلاس مربوطه نیز محاسبه می‌گردد. در این مطالعه از یک طبقه‌بندی کننده ML برای تفسیر مشارکت‌های هندسی بزرگراه‌ها در OSM استفاده شده است. چهار معیار هندسی (فاصله اولین نقطه ترسیمی تا نزدیک‌ترین خیابان، چگالی نقاط ترسیمی ابتدایی، طول ژئودزیک و آزیموت)، برای ارزیابی رفتار ترسیمی داوطلبان استفاده شد و روندها در بین بزرگراه‌های نقشه‌های OSM مشخص شد. براساس پیگیری روندهای مشارکت، مشاهده شد که در برخی کلاس‌های بزرگراه‌ها رفتار ترسیمی خاصی در بین مشارکت‌کنندگان وجود دارد. در بررسی روندها و آمار مشخص شد که مشارکت‌کنندگان بیشتر تمایل به مشارکت در ترسیم جاده‌های محلی اطراف محل سکونت خود را دارند، مانند بزرگراه‌های residential با ۲۲۰۳۷۲ مشارکت و tertiary با ۴۳۸۶۲ مشارکت،که این می‌تواند به علت شناخت بیشتر آن‌ها نسبت به جاده‌های نزدیک به محل زندگی‌شان باشد. همچنین امکان تعیین ترتیب اهمیت در بین اقدامات وجود داشت. پارامتر طول ژئودزیک با مقدار ناخالصی جینی ۰/۱۲۴ دارای بیشترین اهمیت در تمایز کلاس‌ها بوده است. سپس معیارهای آزیموت و فاصله تا نزدیک‌ترین خیابان با ناخالصی جینی به ترتیب ۰/۱۸۹ و ۰/۲۸۱ اثرگذارترین پارامترها بوده‌اند. این مطالعه نشان می دهد که یک طبقه‌بندی‌کننده ML با استفاده از ویژگی‌های هندسی مرتبط می‌تواند برای تعیین روند ترسیم بزرگراه‌ها توسط مشارکت‌کنندگان OSM استفاده شود. تازگی این مطالعه این است که روندهای ترسیم رایج در اقدامات نقشه‌برداری بزرگراه‌ها را نشان می‌دهد و تأثیر پارامترهای گوناگون را بر شناسایی روندها بررسی می‌نماید. در مراحل انجام پژوهش، پارامترهای معنایی مانند نام خیابان‌ها و کاربران نیز وارد روند طبقه‌بندی شدند اما به نظر می‌رسد که افزودن آن‌ها در مرحله دوم تأثیر کمی در پیش‌بینی طبقات مجاورت داشته و باید پارامترهای معنادارتری که مرتبط با روندهای کلاس‌ها باشند، بررسی شوند. روش مورد مطالعه تنها بر روی یک کلان‌شهر مورد بررسی قرار گرفت و احتمال می‌رود در سایر شهرها یا حتی روستاها شناسایی و استخراج روندها نتایج جالب‌توجهی داشته باشند که بایستی موردمطالعه و بررسی در آینده قرار گیرد. محدودیت اصلی مطالعه، معیارهای مورداستفاده به‌عنوان متغیرهای مستقل در PP است که به علت حجم بسیار بالای داده‌ها امکان استخراج و وارد کردن آن‌ها به مطالعه وجود نداشت. استفاده از معیارهای گوناگون و بررسی آن‌ها در تحقیقات آینده می‌تواند قابل‌توجه باشد.

۶. پی‌نوشت‌ها

1. Big Data
2. Cloud computing
3. Distributed database
4. Parallel processing
5. Point of interest
6. Variable importance measures (VIMs)
7. Impurity importance
8. Permutation importance
9. Gini index
10. Gini importance.
11. Gini impurity

۷. مراجع

– Afandizadeh , S., Javanshir , H., & Elyasi , R. (2010). Development of a model for designing urban bus transit network based on tabu search. Transportation Engineering, 5, 13-26. (in Persian)

– Alenouri, H., Meshkani, S. M., Saffarzadeh, M., & Sherafaty Pour, S. (2014). Locating Cameras at the Entrances of Plate Number Rationing Zone to Maximize Violations Detection. Quarterly Journal of Transportation Engineering, 6(2), 181-196. (in Persian)

– Attard, M., Haklay, M., & Capineri, C. (2016). The potential of volunteered geographic

for measuring transportation performance and information distribution of urban bus using volunteer geographic information (VGI). Transportation Engineering 6, 225-236. (in Persian)

– Fogliaroni, P., D'Antonio, F., & Clementini, E. (2018). Data trustworthiness and user reputation as indicators of VGI quality. Geo-Spatial Information Science, 21(3), 213-233.

– Forati, A. M., & Ghose, R. (2020). Volunteered Geographic Information Users Contributions Pattern and its Impact on Information Quality.

– Girres, J. F., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. Transactions in GIS, 14(4), 435-459.

– Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. GeoJournal, 69, 211-221.

– Guo, H., Nguyen, H., Vu, D.-A., & Bui, X.-N. (2021). Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach. Resources Policy, 74, 101474.

– Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

– Hacar, M. (2020). Analyzing the Contribution Trends of Volunteers by Comparing Tag Metadata of OpenStreetMap Residential Roads [Original title in Turkish: OpenStreetMap Yerleşim-içi Yollarına Ait Etiket Bilgilerinin Karşılaştırılmasıyla Gönüllülerin Katkı Sağlama Eğilimlerinin İncelenmesi]. Harita. Dergisi, 164, 77-87.

information (VGI) in future transport systems. Urban Planning, 1(4), 6-19.

– Bégin, D., Devillers, R., & Roche, S. (2013). Assessing Volunteered Geographic Information (vgi) Quality Based on CONTRIBUTORS'Mapping Behaviours. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 40, 149-154.

– Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.

– Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

– Breiman, L. (2017). Classification and regression trees. Routledge.

– Chehreghan, A., & Abaspour, R. (2019). "Evaluation of the geometric similarity of voluntary spatial data in the inner city road network. Transportation Engineering, 10, 357-370.. (in Persian)

– Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(1-4), 131-156.

– Davidovic, N., Mooney, P., & Stoimenov, L. (2016). An analysis of tagging practices and patterns in urban areas in OpenStreetMap. Proceedings of the AGILE 2016 Conference, Helsinki, Finland.

– Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, 40, 139-157.

– Farideh Teymourian , A. A. A., Abass Alimohammadi and Abolghasem SadeghiNiaraki (2014). Developing a system

– Hacar, M. (2022). Analyzing the behaviors of OpenStreetMap volunteers in mapping building polygons using a machine learning approach. ISPRS International Journal of Geo-Information, 11(1), 70.

– Hacar, M., Kılıç, B., & Şahbaz, K. (2018). Analyzing openstreetmap road data and characterizing the behavior of contributors in Ankara, Turkey. ISPRS International Journal of Geo-Information, 7(10), 400.

– Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and planning B: Planning and design, 37(4), 682-703.

– Haklay, M., Basiouka, S., Antoniou, V., & Ather, A. (2010). How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. The cartographic journal, 47(4), 315-322.

– Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2). Springer.

– Ishwaran, H. (2015). The effect of splitting on random forests. Machine learning, 99, 75-118.

– Jokar Arsanjani, J., Helbich, M., Bakillah, M., & Loos, L. (2015). The emergence and evolution of OpenStreetMap: a cellular automata approach. International Journal of Digital Earth, 8(1), 76-90.

– Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1-45.

– Liu, H., & Motoda, H. (2007). Computational methods of feature selection. CRC press.

– Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on knowledge and data engineering, 17(4), 491-502.

– Manandhar, P., Marpu, P. R., Aung, Z., & Melgani, F. (2019). Towards automatic extraction and updating of VGI-based road networks using deep learning. Remote Sensing, 11(9), 1012.

– Mirbaha, B., SHERAFATIPOUR, S., & MAHPOUR, A. (2016). Congestion pricing model for urban congested roads (case study: sadr elevated highway). Transportation Engineering, 7, 353-365.

– Mooney, P., Corcoran, P., & Winstanley, A. C. (2010). Towards quality metrics for OpenStreetMap. Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems.

– Nash, A. (2009). Web 2.0 applications for improving public participation in transport planning. Transportation Research Board 89th Annual Meeting,

– Neis, P., & Zipf, A. (2012). Analyzing the contributor activity of a volunteered geographic information project—The case of OpenStreetMap. ISPRS International Journal of Geo-Information, 1(2), 146-165.

– Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? Bioinformatics, 34(21), 3711-3718.

– Nicodemus, K. K. (2011). On the stability and ranking of predictors from random forest variable importance measures. Briefings in bioinformatics, 12(4), 369-373.

– O'reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. Communications & strategies(1), 17.

– OpenStreetMap. OSMstats. Retrieved 04/12/2023 from https://wiki.openstreetmap.org/wiki/Stats

– Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222.

– Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

– Rehrl, K., & Gröchenig, S. (2016). A framework for data-centric analysis of mapping activity in the context of volunteered geographic information. ISPRS International Journal of Geo-Information, 5(3), 37.

– Saberian, J., Malek, M., & Hamrah, M. (2014). Using Dual Graph and Wavelet Transform for Evaluation and Planning Transportation systems. Transportation Engineering, 5, 317-328. (in Persian)

– Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1), 1-21.

– Zhao, P., Jia, T., Qin, K., Shan, J., & Jiao, C. (2015). Statistical analysis on the evolution of OpenStreetMap road networks in Beijing. Physica A: Statistical Mechanics and its Applications, 420, 59-72.

سیده سعیده ساداتی، درجه کارشناسی در رشته مهندسی نقشه‌برداری را در سال ۱۳۹۹ از دانشگاه صنعتی نوشیروانی بابل أخذ نمود و درزمان نگارش این مقاله دانشجوی کارشناسی ارشد مهندسی نقشه برداری – سیستم‌های اطلاعات مکانی دانشگاه تهران می‌باشد. زمینه های پژوهشی مورد علاقه ایشان، اطلاعات مکانی داوطلبانه، داده کاوی مکانی و یادگیری ماشین است.

رحیم علی‌عباسپور، درجه کارشناسی در رشته مهندسی نقشه‌برداری را در سال ۱۳۷۹ از دانشگاه تهران و درجه کارشناسی ارشد در رشته مهندسی نقشه برداری – سیستم‌های اطلاعات مکانی را در سال ۱۳۸۱ از دانشگاه تهران اخذ نمود. در سال ۱۳۸۹ موفق به کسب درجه دکتری سیستم‌های اطلاعات مکانی از دانشگاه تهران گردید. زمینه های پژوهشی مورد علاقه ایشان داده کاوی مکانی–زمانی، بهینه‌سازی مکانی، اطلاعات مکانی داوطلبانه و خدمات مکان‌مبنا (LBS) بوده و در حال حاضر عضو هیات علمی با مرتبه دانشیار در دانشکده مهندسی نقشه برداری و اطلاعات مکانی دانشکدگان فنی دانشگاه تهران است.

علیرضا چهرقان در سال ۱۳۸۸رشته کارشناسی مهندسی نقشه‌برداری را از دانشگاه تهران اخذ نمود. همچنین کارشناسی ارشد و دکتری خود را به ترتیب در سال های ۱۳۹۰ و ۱۳۹۶ از دانشگاه تهران دریافت کرد. زمینه‌های پژوهشی مورد علاقه ایشان شامل تناطریابی عوارض در پایگاه های داده مکانی، اطلاعات مکانی داوطلبانه، داده کاوی مکانی–زمانی، تصمیم گیری های مکان مبنا، بهینه سازی مکانی، خدمات مکان مبنا و محاسبات هندسی می باشد. در حال حاضر ایشان عضو هیات علمی با مرتبه دانشیار در دانشگاه صنعتی سهند است.

عباس عابدینی، درجه کارشناسی در رشته مهندسی عمران _نقشه‌برداری را در سال ۱۳۷۳ از دانشگاه تهران و درجه کارشناسی ارشد در رشته مهندسی ژئوانفورماتیک را در سال ۱۳۸۶ از دانشگاه کاربردی اشتوتگارت آلمان و همچنین درجه معادل کارشناسی ارشد در رشته مهندسی ژئوماتیک را در سال ۱۳۹۳ از دانشگاه اشتوتگارت آلمان اخذ نمود. در سال ۱۴۰۰ موفق به کسب درجه دکتری مهندسی عمران _ ژئودزی از دانشگاه اصفهان گردید. زمینه های پژوهشی مورد علاقه ایشان مهندسی ژئوماتیک، تحلیل آماری داده‌ها و بصری‌سازی مکانی، هیدرولوژی مهندسی و مهندسی هیدروگرافی است. در حال حاضر عضو هیات علمی با مرتبه استادیار در دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی دانشکدگان فنی دانشگاه تهران است.