نشریه علمی - پژوهشی علوم و فنون نقشه برداری، دوره ششم، شماره ۲، آبان ماه ۱۳۹۵

روشی برای اعتبارسنجی اطلاعات معابر در OSM بدون استفاده از اطلاعات مرجع و با استفاده از سایر اطلاعات OSM

سیدحسن حسینی[۱]، رحیم علی‌عباسپور[۲*]

[۱] کارشناس ارشد سیستم‌های اطلاعات مکانی - دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی - پردیس دانشکده‌های فنی - دانشگاه تهران
hasan.tlc@gmail.com

[۲] استادیار دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی - پردیس دانشکده‌های فنی - دانشگاه تهران
abaspour@ut.ac.ir

(تاریخ دریافت خرداد ۱۳۹۵، تاریخ تصویب مرداد ۱۳۹۵)

چکیده

کنترل و تضمین کیفیت اطلاعات مکانی داوطلبانه یک از بزرگترین چالشهای پیش روی این نوع از گردآوری اطلاعات مکانی است. روش‌های ارائه شده برای ارزیابی کیفیت اطلاعات مکانی داوطلبانه به صورت کلی شامل دو رویکرد مقایسه این اطلاعات با یک مرجع معتبر و یا روش‌های بدون نیاز به مقایسه با مرجع معتبر می‌شود. روش‌های گروه اول نیازمند دسترسی به اطلاعات مرجع هستند که دسترسی به این اطلاعات همواره امکان‌پذیر نمی‌باشد. روش‌های گروه دوم، یا مثل روش اعتبارسنجی اطلاعات سعی بر به دست آوردن اعتبار اطلاعات بر اساس معیارهای بدست آمده از خود اطلاعات دارند و یا با استفاده از قوانین حاکم بر داده‌ها مثل سازگاری منطقی داده-ها به ارزیابی کیفیت داده‌ها می‌پردازند. در این تحقیق تلاش بر ارائه روشی برای ارزیابی کیفیت اطلاعات مکانی داوطلبانه با استفاده از خصوصیات قابل محاسبه از داده مکانی و بدون نیاز به مقایسه با اطلاعات مرجع گردیده است. این تحقیق با استفاده از خصوصیات قابل محاسبه از داده‌های موجود در شبکه راه‌های OpenStreetMap، به صورت خاص به تحلیل میزان دقت دسته‌بندی خیابان‌ها، به عنوان بخشی از دقت معنایی مربوط به شبکه راه‌ها، می‌پردازد. در این تحقیق از درخت تصمیم و شبکه عصبی با ساختار چند لایه برای یادگیری این خصوصیات استفاده گردید. برای تعیین میزان دقت دسته‌بندی بدست آمده، معیارهای دقت و بازخوانی برای هر روش محاسبه گردید. درخت تصمیم با دقت وزن‌دار بدست آمده ۸۴/۱ ٪ نشان‌دهنده یک مدل مناسب برای این روش است. این روش بر اساس اطلاعات قابل استخراج از داده‌ها استوار است و می‌تواند برای داده‌های هر شهر دیگری تعمیم و استفاده گردد.

واژگان کلیدی: OpenStreetMap، کیفیت داده، داده مکانی داوطلبانه، آنالیز شبکه راه، ماشین یادگیری

* نویسنده رابط

۱- مقدمه

استفاده از داده‌های مکانی داوطلبانه، برای کاربری‌ها نیازمند تضمین کیفیت این اطلاعات است. بیشتر این اطلاعات توسط داوطلبان با مهارت کم یا بدون مهارت در ترسیم نقشه تولید شده و می‌شود و این موضوع باعث عدم گسترش استفاده از این داده‌ها در پردازشهای رسمی تا زمان تضمین کیفیت اطلاعات می‌گردد. درنتیجه، ارزیابی سطح کیفی این اطلاعات به یکی از موضوعات اساسی برای استفاده از این اطلاعات تبدیل شده است. ارزیابی و آگاهی از کیفیت داده‌ها، آگاهی کاربر از میزان ریسک محتمل، برای استفاده از داده‌ها را به دنبال دارد. روش‌های ارائه شده برای ارزیابی کیفیت این اطلاعات به صورت کلی شامل دو رویکرد مقایسه این اطلاعات با یک مرجع معتبر و یا روش‌های مستقل از داده مرجع معتبر است.

تحقیقات زیادی بر مبنای روش‌های مقایسه اطلاعات مکانی داوطلبانه با یک مجموعه اطلاعات مکانی مرجع یا اطلاعات گرداوری شده توسط متخصصین صورت گرفته است. تحقیقات صورت گرفته بیشتر شامل ارزیابی صحت موقعیتی یا میزان کامل بودن عوارض نقطه‌ای یا خطی است. Haklay به مقایسه دقت مکانی OSM با داده‌های OS Meridian 2 برای انگلستان پرداخت [۱]. در این تحقیق، برای بررسی صحت موقعیتی عوارض خطی، درصدی از عارضه مورد پرسش که در بافر با شعاع مشخص از همان عارضه در داده‌های مرجع قرار دارد، مورد استفاده قرار می‌گیرد [۱۲]. افزون براین، Haklay و همکاران [۲] روشی برای تطبیق خودکار پایگاه‌داده اطلاعات مکانی داوطبانه با پایگاه‌داده اطلاعات مکانی مرجع به صورت عارضه مبنا برای عارضه‌های خطی ارائه کردند. در این روش، با استفاده از قیدهای هندسی و ویژگیهای داده‌ها به بررسی میزان کامل بودن OSM در مناطق مختلف انگلیس پرداخته شده است. Touya و Girres [۱۷] به مقایسه داده‌های OSM و پایگاه داده BD TOPO برای کشور فرانسه پرداختند. در این تحقیق صحت موقعیتی عوارض نقطه‌ای، با استفاده از فاصله اقلیدسی محاسبه گردید که تغییرات فاصله برای نقاط در در بازه ۲/۵ تا ۱۰ متر و با میانگین ۶/۶ متر محاسبه گردید. نتایج بررسی صحت موقعیتی عوارض خطی و چندضلعی‌ها نشان‌دهنده تفاوت‌های قابل قبول بین داده‌ها در فرانسه دارند. نتایج مربوط به بررسی کمی، نشان‌دهنده قید شدن برچسب اصلی اکثر عوارض برای این داده‌هاست. Neis و همکاران [۱۸] به بررسی پیشرفت اطلاعات مکانی داوطلبانه در پروژه OSM مربوط به کشور آلمان از سال ۲۰۰۷ تا ۲۰۱۱ و مقایسه این اطلاعات با اطلاعات پایگاه داده TomTom پرداختند. در این مقایسه برای کل شبکه خیابان‌ها و راه‌های پیاده رو OSM نسبت به پایگاه داده TomTom دارای ۲۷٪ داده بیشتر است اما برای شبکه خیابان‌های ماشین رو این داده‌ها ۹٪ کمتر از داده‌های TomTom است. در تحقیق دیگری برای اطلاعات داوطلبانه آلمان، Zipf و همکاران [۱۶] صحت مفهومی و میزان کامل بودن داده‌های کاربری اراضی قسمتی از آلمان را با مقایسه با اطلاعات مرجع بررسی کردند. در حالی که نتایج نشان می‌داد که مناطق پرجمعیت دارای تطابق بالا با اطلاعات مرجع هستند، اما مناطق کم جمعیت لزوماً دارای میزان تطابق کم نیستند. اما مهمترین مشکل این روش‌ها اینست که اطلاعات مرجع در صورت در دسترس بودن گران‌قیمت هستند یا در موارد محدودی می‌توان از آنها استفاده کرد و عملاً همواره داده مرجعی برای مقایسه وجود ندارد. از اینرو، رسیدن به روش ارزیابی کیفیت با کارکرد کلی با استفاده از این روش امکان‌پذیر نمی‌باشد.

روش‌های ارزیابی بدون استفاده از اطلاعات مرجع را می‌توان به ۳ گروه کلی تقسیم کرد که شامل روش‌های ارزیابی کیفیت اطلاعات مکانی با استفاده از اطلاعات به دست آمده از خود داده‌ها، استفاده از قوانین و روابط حاکم بر داده‌های مکانی برای ارزیابی کیفیت این اطلاعات و روش بازنگری و اصلاح خطا توسط اعضایی که داده‌ها را به اشتراک گذاشته‌اند است.

روش بازنگری و اصلاح خطا توسط کاربران مبتنی بر این اصل است که اگر کاربران اجازه بازنگری و اصلاح خطاها را داشته باشند، این مشارکت سرانجام به سمت رسیدن به حقیقت همگرا می‌شود. اما این روش در حالت تعداد بالای بازنگریها و برای موضوعاتی که عمومی‌تر هستند، نسبت به موضوعات خاص و مبهم، کارکرد و نتایج بهتری دارد و برای پدیده‌های مکانی که توجه کمتری را جلب می کنند، مانند یک محل کم بازدید، یا پدیده‌هایی که در مناطق با جمعیت کم قرار دارند می‌تواند زمینه وارد کردن اطلاعات نادرست به صورت عمدی را فراهم کند.

در روش دیگر بدون نیاز به اطلاعات مرجع، از قوانین و روابط حاکم بر داده‌های مکانی برای ارزیابی کیفیت این اطلاعات استفاده می‌گردد. Hashemi [۱۴] با استفاده از قواعد سازگاری متناسب با اطلاعات مکانی داوطلبانه، بر مبنای روش‌های سازگاری در داده‌های تلفیقی و داده‌های دارای نمایش‌چندگانه، دو رابطه‌ی هم ارزی رابطه‌مبنا و عارضه مبنا را مورد بررسی قرار داد.

گروه دیگر روش‌ها شامل روش‌های ارزیابی کیفیت اطلاعات مکانی با استفاده از اطلاعات به دست آمده از خود داده‌ها است. به طور مثال Bishr و Janowicz [۳] اعتبار اطلاعات را به عنوان یک نماینده در ارزیابی کیفیت در اطلاعات مکانی داوطلبانه معرفی کردند. Kaβler و همکاران [۴] از اعتبار و منشاء اطلاعات، برای مطالعه الگوی مشارکت در OSM استفاده‌کردند. در ادامه این تحقیق Kaβler و De Groot [۵] چندین مقیاس برای ارزیابی اعتبار با استفاده از منشاء اطلاعات معرفی کردند. Antonio و همکاران [۶] با ارائه یک روش بر پایه تعیین اعتبار اطلاعات بر اساس تاریخچه خود اطلاعات، تولید، تغییر یا حذف و توپولوژی، ارزیابی اعتبار اطلاعات را فقط با استفاده از خود اطلاعات مورد ارزیابی قرار دادند.

در این تحقیق تلاش شده است تا یک روش ارزیابی اطلاعات مکانی داوطلبانه بر مبنای اطلاعات بدست آمده از خود داده‌های داوطلبانه و مستقل از داده‌های خارجی ارائه گردد. روش‌های ارزیابی کیفیت اطلاعات مکانی با استفاده از اطلاعات به دست آمده از خود داده‌ها، علاوه بر عدم نیاز به اطلاعات مرجع، از ابتدا با ساختار داده‌های مکانی داوطلبانه ایجاد شده و با ساختار این داده‌ها سازگارند.

OSM به عنوان یکی از موفق‌ترین منابع داده‌های مکانی داوطلبانه، مورد توجه بسیاری از محققین برای ارزیابی این اطلاعات است. اما بیشتر تحقیق‌های صورت گرفته برای ارزیابی اطلاعات مکانی داوطلبانه، بر ارزیابی دقت هندسی و کامل‌بودن داده‌ها تاکید دارند و اندکی از روش‌ها برای ارزیابی بخش‌های دیگر اطلاعات مثل صحت معنایی صورت گرفته است. OSM در بخش داده‌های معنایی شامل طیف وسیعی از گزینه‌های آماده است و این داده‌ها را با استفاده از برچسب‌ها ذخیره می‌کند. Ballatore and Bertolotto [۷] داده‌های OSM را از نظر داده مکانی غنی و از نظر داده معنایی ضعیف توصیف کرده است. Mooney and Corcoran [۸] متوجه شدند که برای عارضه‌های با ویرایش بالا (تعداد ویرایش بیش از ۱۵ بار) بین افزایش تعداد کاربران مرتبط با یک عارضه و تعداد برچسب‌های عارضه رابطه معناداری وجود ندارد، اما بین تعداد نسخه‌های ایجاد شده برای آن عارضه و تعداد برچسب‌ها رابطه آماری وجود دارد. این موضوع نشان می‌دهد که کاربران جدید برای یک عارضه، برچسب‌های موجود عارضه را بدون افزودن برچسب جدید می‌پذیرند. همچنین بیشتر ویرایش‌های انجام شده (۷۹٪ از ویرایش‌ها) شامل افزودن اطلاعات مکانی (نقاط) به عارضه‌ها است. از ویرایش برچسب‌ها، ۶۴٪ ویرایش‌ها برگشت به مقدار قبلی و ۳۲٪ بروزرسانی یا تغییر به مقدار جدید برچسب است. بررسی اینکه کدام نوع ویرایش برچسب‌ها (برگشت به مقدار قبلی یا بروزرسانی) دقیقاً درست است، مشکل است. به عنوان مثال، در یک مورد کشمکش برچسب‌ها[۱] بین کاربران OSM برای برچسب مربوط به یکی از خیابان‌ها در آلمان، تا اواسط بهمن ماه ۱۳۸۹، ۸۸ برچسب برای این خیابان ایجاد شد. درحالی که یکی از کاربران تاکید بر برچسب پررفت‌وامد برای این خیابان داشت، بقیه کاربران برچسب در حال‌ساخت را برای این خیابان انتخاب می‌کردند.

یکی از روش‌های افزودن اطلاعات به OSM با استفاده از ورود اطلاعات از پایگاه‌های داده مرجع به پایگاه داده OSM است. Zielstra و همکاران [۹] تاثیر اطلاعات واردشده از پایگاه‌های داده مرجع، برکامل بودن اطلاعات OSM را ارزیابی کردند. این اطلاعات یک زیربنا برای فعالیت کاربران OSM ایجاد می‌کند، اما مشکلات مربوط به تبدیل برچسب‌ها این پایگاه‌های داده به برچسب‌های OSM، اعتمادپذیری و کاربردی بودن این اطلاعات را با مشکل مواجه می‌کند. پایگاه داده TIGER، که تحقیق برمبنای این پایگاه‌داده صورت گرفته، راه‌های روستایی، خیابان‌های شهری و راه‌های دسترسی‌محلی را جزء گروه خیابان مسکونی قرار می‌دهد، در حالی که OSM این خیابان‌ها را در گروه‌های متفاوت دسته‌بندی می‌کند. بنابراین، درنتیجه تفاوت در الگوی دسته‌بندی خیابان‌ها در پایگاه‌داده‌های دیگر و OSM، خطاهای ایجاد شده در بخش برچسب‌ها برای داده‌های وارد شده، همراه این داده‌ها باقی می‌ماند.

---

۱ Tag war

همه این موارد نشان‌دهنده وجود یک نیاز اساسی برای یک روش ارزیابی، برای داده‌های معنایی اطلاعات مکانی داوطلبانه است. بسیاری از کاربری‌ها مثل مسیریابی وابسته به اطلاعات شبکه راه‌ها هستند. اطلاعات معنایی به عنوان بخش مهمی از اطلاعات، که در داده‌های شبکه راه‌ها، شامل دسته‌بندی راه‌ها می‌گردد، برای آنالیز درست شبکه راه‌ها ضروری است [۱۰].

در پروژه‌های اشتراکی، مانند OSM، کاربران به راحتی قادر به اضافه کردن، ویرایش و حذف برچسب‌ها هستند. OSM برای دسته‌بندی راه‌ها دارای طیف وسیعی از گزینه‌های آماده است. همچنین در صورت نیافتن گزینه لازم، امکان اضافه کردن برچسب دلخواه به صورت دستی نیز وجود دارد. برای عارضه‌ها در OSM به روز رسانی این برچسب‌ها می‌تواند نشان‌دهنده تغییرات در زمان و مکان باشد و صحت این برچسب‌ها برای استفاده در کاربردهای مختلف ضروری است. در این مقاله، روشی خودکار برای ارزیابی کیفیت و میزان صحت داده‌های معنایی اطلاعات مکانی داوطلبانه و به طور خاص، دسته‌بندی‌های صورت گرفته برای شبکه راه‌های OSM ارائه شده که به اطلاعات مرجع نیازی ندارد.

در ادامه در بخش۲، به مرور میانی نظری برای این تحقیق پرداخته، در بخش ۳ روش پیاده‌سازی و الگوریتم-های مورد استفاده توضیح داده شده و در بخش ۴ نتایج به دست‌آمده بررسی می‌شود و در بخش نهایی نتیجه‌گیری و کارهای ممکن برای تحقیق‌های آتی ارائه می‌گردد.

۲- مبانی نظری

یادگیری درخت تصمیم‌گیری[۱] از پرکاربردترین روش‌های یادگیری استقرایی است که برای تخمین توابع هدف گسسته مقدار و داده‌های خطادار استفاده می‌شود. تابع تخمین‌زده شده با یک درخت تصمیم‌گیری مشخص می‌شود که درخت بدست‌آمده را می‌توان به صورت دسته‌ای از دستورهای if-then نیز نمایش داد. اکثر الگوریتم‌ها که برای یادگیری درختی ایجاد شده‌اند، نسخه‌های مختلف الگوریتم اساسی ID3 هستند که از جستجویی حریصانه و بالا به پایین[۲] برای جستجوی فضای درخت‌های ممکن استفاده می‌کند. فضای فرضیه‌ای جستجو شده توسط ID3 مجموعه تمام درخت-های تصمیم‌گیری است. ID3 جستجوی ساده به پیچیده و hill-climbing را در این فضای فرضیه‌ای انجام می‌دهد. این جستجو با یک درخت ساده تهی شروع شده و درنهایت به درختی می‌رسد که بتواند تمامی نمونه‌های آموزشی را دسته-بندی کند. بایاس استقرایی بر اساس ترجیح درخت کوتاه‌تر نسبت به درخت بلندتر و یا برگزیدن درختی که بهره اطلاعاتش در نزدیکی ریشه بیشتر است، عمل می‌کند [۱۳].

شبکه‌های عصبی مصنوعی با الهام از سازوکار فراگیری و آموزش مغز بنا شده‌اند. مقادیر یک شبکه عصبی برای انجام وظیفه‌ای مشخص مانند شناسایی الگوها و دسته‌بندی اطلاعات، در طول یک فرآیند یادگیری تنظیم می‌شوند. در حالت کلی یک شبکه عصبی مصنوعی سیستمی دارای ورودی‌ها و خروجی‌هاست که معمولاً از تعدادی عناصر پردازش غیر خطی تشکیل شده است. این عناصر پردازش با وزن‌های قابل تنظیم به هم وصل شده‌اند. تغییر این وزن‌ها رفتار ورودی-خروجی شبکه را تغییر می‌دهد. هدف از این سیستم انتخاب وزن‌های شبکه به نحوی است که رابطه مناسب ورودی-خروجی به دست آید. یکی از ساختارهای متداول شبکه‌های عصبی ساختار چندلایه است که با توجه به سادگی و توانایی آن کاربرد گسترده‌ای در علوم مختلف یافته است .در آرایش شبکه عصبی با ساختار چند لایه، نرون‌ها در چندین لایه به ترتیب قرار دارند به این صورت که هر لایه ورودی‌های خود را از لایه قبلی دریافت و خروجی-های خود را به لایه بعدی می‌فرستد. تعداد لایه‌های میانی و یا تعداد نرون‌های هر لایه، وابستگی مستقیمی با پیچیدگی اطلاعات اعمال شده به شبکه دارد. با توجه به کاربرد گسترده این ساختار از شبکه‌های عصبی در کاربردهای پیشبینی و تخمین و تشخیص الگو و دسته‌بندی، می‌توان گفت شبکه عصبی با ساختار چند لایه از پرکاربردترین ساختار شبکه‌های عصبی می‌باشد [۱۳].

برای استفاده از ماشین یادگیری، داده‌ها به دو بخش داده‌های یادگیری و آزمون تقسیم می‌گردد. داده‌های یادگیری شامل داده‌های ارزیابی نیز می‌گردد تا مدل ایجاد شده با داده‌های یادگیری، با استفاده از داده‌های ارزیابی تنظیم شده و بررسی گردد تا پارامترهای مدل به درستی شناسایی شود. برای آزمون درستی مدل، از داده‌های آزمون که در مرحله یادگیری استفاده نشده‌اند و برای مدل جدید هستند استفاده می‌شود.

---
۱ Decision Tree
۲ Top-down

۳- روش پیشنهادی

در پروژه‌های اشتراکی، مانند OSM، کاربران به راحتی قادر به اضافه کردن، ویرایش و حذف برچسب‌ها هستند. برای عارضه‌ها در OSM به روز رسانی این برچسب‌ها می‌تواند نشان‌دهنده تغییرات در زمان و مکان باشد و صحت این برچسب‌ها برای استفاده در کاربردهای مختلف ضروری است. در این مقاله، روشی خودکار برای ارزیابی کیفیت و میزان صحت داده‌های معنایی اطلاعات مکانی داوطلبانه و به طور خاص، دسته‌بندی‌های صورت گرفته برای شبکه راه‌های OSM ارائه شده که به اطلاعات مرجع نیازی ندارد.

طراحی خیابان‌های شهری بر اساس کارکرد هر گروه از خیابان‌ها صورت می‌گیرد. به عنوان مثال، بزرگراه‌ها و خیابان‌های اصلی با وظیفه جابه‌جایی ترافیک شهری، طراحی کاملا متفاوتی نسبت به خیابان‌های مسکونی یا سرویس‌دهنده دارند. در دسته اول طراحی‌ها برای سرعت بالای وسایل نقلیه و وسایل نقلیه سنگین مثل اتوبوس‌های دومحوره صورت می‌گیرد، اما در دسته دوم مسئله ایمنی و آسایش ساکنین اطراف خیابان‌ها بر سرعت جابه‌جایی اولویت دارد. هر دسته از این خیابان‌ها بر اساس کارکردی که در بافت شهری دارند، دارای خصوصیات خاصی هستند که این خیابان‌ها را از خیابان‌های دیگر دسته‌ها متمایز می‌کند. خصوصیاتی مثل شکل، اندازه و بقیه ویژگی‌های مکانی مثل موقعیت مطلق و نسبی خیابان نسبت به خیابان‌های دیگر قابل ذکر هستند. تلاش ما بر استخراج و استفاده مناسب از خصوصیات قابل اندازه‌گیری در خیابان‌ها برای ارزیابی دسته‌بندی خیابان‌هاست. برای ایجاد یک روش ارزیابی عمومی برای خیابان‌ها، خصوصیات مورد استفاده نباید بر اساس خصوصیت‌های خاص خیابان‌های یک شهر بوده و باید برپایه ویژگی‌های عمومی خیابان‌ها باشد، هرچند مقادیر خصوصیات مورد استفاده می‌تواند از شهری به شهر دیگر متفاوت باشد. خصوصیاتی که در این تحقیق استفاده شده است شامل چگونگی اتصال خیابان‌ها به یکدیگر، میزان مستقیم بودن، طول، تعداد تقاطع‌ها و تعداد بن‌بست‌ها است. در ادامه این ویژگیها به تفصیل مورد بررسی قرار گرفته‌اند.

نحوه اتصال: از خصوصیات بارز در خیابان‌های شهری، چگونگی اتصال انواع مختلف خیابان تشکیل‌دهنده شبکه معابر به یکدیگر است. در این پژوهش، برای شناسایی نحوه اتصال خیابان‌ها به خیابان مورد پرسش از مدل bag-of-word استفاده شد. در این مدل، ترتیب اتصال خیابان‌ها اهمیتی ندارد و صرفاً تعداد خیابان‌های هر دسته که به خیابان مورد پرسش متصل است، بررسی می‌گردد. در مرحله اول، براساس اندازه‌گیری تعداد اتصال خیابان‌های مختلف به یکدیگر، احتمال اتصال هر دسته از خیابان به دسته‌های دیگر محاسبه می‌شود. برای ارزیابی یک خیابان، با استفاده از نحوه اتصال آن به خیابان‌های دیگر، احتمال وقوع این نحوه اتصال برای خیابان مورد پرسش براساس احتمالات حاصل از مرحله اول بدست می‌آید.

مستقیم بودن[۱]: در طراحی شهری سعی می‌گردد خیابان‌های با وظیفه جابه‌جایی به صورت مستقیم‌تر طراحی گردد تا در حداقل زمان، امکان مسافرت بین دو نقطه را فراهم کنند. برای محاسبه مستقیم بودن از روش همواری منحنی‌میزان[۲] استفاده شد. براساس این روش نسبت مساحت شکل به مجذور محیط به عنوان میزان مستقیم بودن درنظر گرفته می‌شود [۱۵]. برای استفاده از این روش بر روی خیابان‌ها، سه نقطه متوالی در یک خیابان به عنوان یک مثلث فرض شده و مقدار مستقیم بودن برای این مثلث‌ها محاسبه می‌شود. با میانگین‌گیری برای تمام نقاط یک خیابان، این مقدار برای کل آن به دست می‌آید. با توجه به مفهوم مستقیم بودن خیابان‌ها، این روش برای شکل‌های مختلف خیابان، در مقایسه با روش‌های دیگر محاسبه مستقیم بودن خط، دارای نتیجه بهتری است.

طول: معابر بر اساس وظیفه‌ای که در بافت شهری نسبت به جابه‌جایی یا ایجاد دسترسی دارند، دارای طول‌های کاملا متفاوتی هستند. این خصوصیت از خصوصیات اصلی هر معبر می‌باشد. محاسبه طول خیابان‌ها با اندازه‌گیری فاصله دو نقطه متوالی و جمع‌کردن تمام فواصل مربوط به یک خیابان صورت می‌گیرد.

تعداد تقاطع‌ها: تقاطع‌ها در شبکه OSM ، با وجود نقطه مشترک در دو خیابان، شناسایی می‌گردد. در خیابان‌های با وظیفه جابه‌جایی، مثل بزرگراه‌ها، تعداد تقاطع‌ها محدودتر است و خیابان‌های با وظیفه دسترسی یا جمع‌کننده و پخش‌کننده، دارای تقاطع بیشتری هستند. در واقع تعداد تقاطع‌های یک خیابان می‌تواند نشان‌دهنده میزان ایجاد دسترسی و ارتباط با دیگر خیابان‌ها باشد.

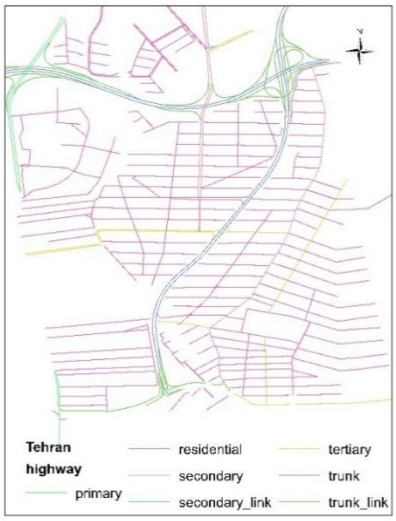
---

۱ Linearity
۲ Contour smoothness

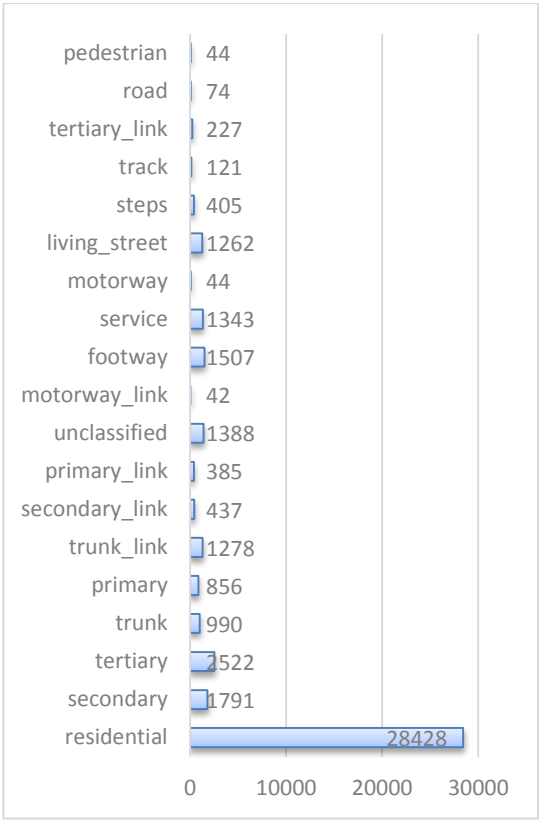
تعداد معابر بن‌بست: بعضی از انواع خیابان‌ها مثل خیابان مسکونی که وظیفه دسترسی دارند، نسبت به خیابان‌های دیگری مثل خیابان‌های اصلی که وظیفه جابه‌جایی را در جریان ترافیک به عهده دارند، به بن‌بست‌های بیشتری متصل هستند. در واقع یک بن‌بست فقط دسترسی به همان خیابان را فراهم می‌کند.

برای بررسی از داده‌های شهر تهران در OSM استفاده شد. شکل۱ نشان‌دهنده یک قسمت کوچک از شبکه خیابان‌های OSM برای شهر تهران است که چند دسته مختلف خیابان را شامل می‌شود. کاربران برای دسته‌بندی خیابان‌های تهران، شامل ۴۳۲۰۰ خیابان، ۲۴ برچسب به کار برده‌اند که ما تلاش کردیم تا همه این برچسب‌ها را مورد ارزیابی قرار دهیم. اما از ۵ برچسب به خاطر تعداد استفاده بسیار کم در شهر تهران، کمتر از ۴۰ مورد در پایگاه داده با ۴۳۲۰۰ مورد، صرفنظر شد. در نهایت تعداد ۱۹ برچسب برای ارزیابی خیابان‌های تهران مورد استفاده قرار گرفت. همانطور که در شکل۲ نشان داده شده است، تعداد خیابان‌ها در دسته‌های مختلف خیابان‌های تهران بسیار متفاوت است. به طور مثال خیابان مسکونی ۶۶٪ از کل خیابان‌های تهران را شامل می‌شود و برچسب خیابان پررفت‌وامد با تعداد ۲۵۲۲ خیابان، بعد از خیابان مسکونی بیشترین تعداد را شامل می‌شود.

میزان موفقیت روش‌های یادگیری وابستگی زیادی به در دسترس بودن داده‌های آموزشی خوب، مانند داده‌های با برچسب صحیح دارد. ارزیابی کیفیت داده‌های OSM مربوط به تهران، با مقایسه سازگاری با نقشه‌های شهرداری تهران، نشان‌دهنده دقت نسبتاً خوب این داده‌هاست [۱۱].



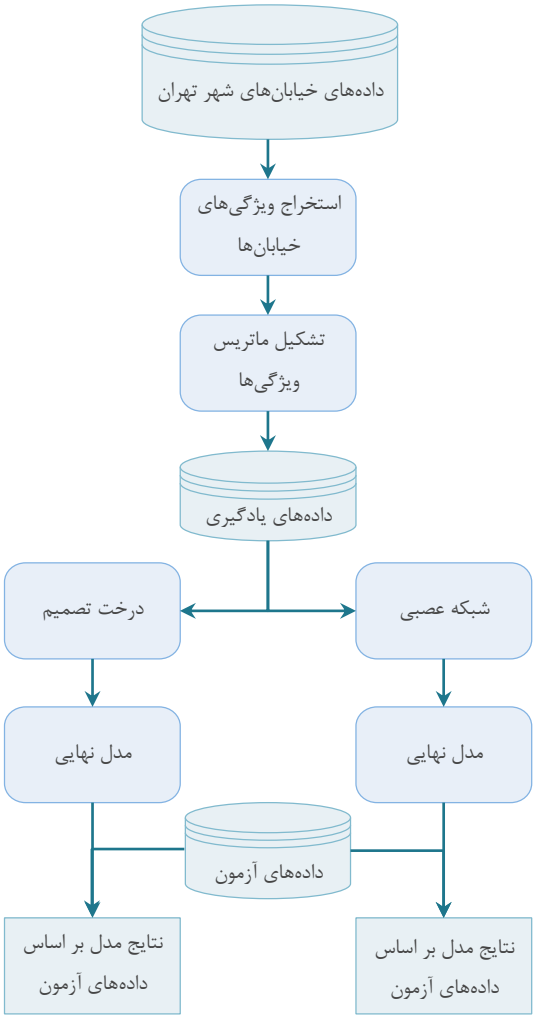
شکل۱- برچسب‌های استفاده شده برای خیابانها در قسمتی از تهران



شکل۲- فراوانی برچسب‌های استفاده شده در شهر تهران

در این روش ما به محاسبه ویژگی‌های ذکر شده برای خیابان‌ها بر اساس آخرین داده‌های بروزشده در OSM پرداختیم. در این روش تاریخچه داده‌ها مورد استفاده قرار نگرفته و آخرین نسخه داده‌ها مناسب‌ترین داده‌ها برای محاسبات است. هر یک از ویژگی‌های ذکر شده، برای تمام خیابان‌ها محاسبه گردید. این مقادیر برای هر خیابان به صورت یک بردار منظم با ۶ عضو ذخیره گردیده و آخرین مقدار در این بردار نشان‌دهنده دسته‌بندی انجام شده توسط کاربران است. با محاسبه مقادیر ویژگی‌ها برای تمام خیابان‌ها ماتریس با اندازه ۶×۴۳۱۹۹ برای داده‌ها ایجاد شد که شامل ۴۳۱۹۹ نمونه و ۵ ویژگی است. این داده‌ها به صورت اتفاقی و بدون هم‌پوشانی به نسبت ۷۰٪ و ۳۰٪ به ترتیب به داده‌های آموزشی و آزمون تقسیم گردید.

با استفاده از نرم‌افزار WEKA به تشکیل مدل با استفاده از درخت تصمیم و شبکه عصبی پرداختیم. مدل‌های بدست آمده با استفاده از هر الگوریتم، توسط داده‌های آزمون که در فرایند یادگیری از آن‌ها استفاده نشده است، بررسی گردید. روند پیاده‌سازی در شکل۳ نمایش داده شده است.



شکل۳- روند انجام روش پیشنهادی

مدل به دست آمده با استفاده از درخت تصمیم، شامل یک درخت با ۳۰۹۹ برگ انتهایی است که ویژگی نحوه اتصال خیابان‌ها به یکدیگر، به عنوان متمایز کننده‌ترین ویژگی در دسته‌بندی برای گره ریشه استفاده شده است. درخت تصمیم بدست آمده با استفاده از الگوریتم J48 هرس‌شده، موفق به دسته‌بندی ۸۴/۵٪ از خیابان‌ها به صورت درست گردید. شبکه عصبی با ساختار چندلایه دارای دو لایه مخفی و هر لایه شامل ۴۳ نرون است. داده‌ها نرمال شده و با توجه به وجود برچسب‌های استفاده شده با تعداد کم در فضای داده‌ها از گام آموزشی ۰/۵ و بدون کاهش گام آموزشی در طول یادگیری برای جستجوی بهتر این فضا استفاده گردید. این شبکه قادر است تا ۷۲/۴٪ از خیابان‌ها را به صورت صحیح دسته‌بندی کند.

۴- نتایج و ارزیابی

برای بررسی سطح کیفی نتایج، دو معیار Precision و Recall به صورتی که در ادامه تعریف شده‌اند، استفاده گردید. درحالی‌که TP یا مثبت حقیقی[۱] نشاندهنده تعداد برچسب‌های دسته مورد بررسی که به درستی دسته‌بندی شده‌اند، FP یا مثبت کاذب[۲] برای تعداد برچسب‌های دسته‌های دیگر که به اشتباه در دسته مورد بررسی دسته‌بندی شده‌اند و FN یا منفی کاذب[۳] تعداد برچسب‌های دسته مورد بررسی که به اشتباه متعلق به دیگر دسته‌ها، دسته‌بندی شده‌اند[۱۹].

$$\text{precision} = \frac{TP}{TP + FP} * 100\% \qquad (۱)$$

$$\text{recall} = \frac{TP}{TP + FN} * 100\% \qquad (۲)$$

برای ارائه کیفیت دسته‌بندی برای همه دسته‌ها از Precision وزندار و Recall وزندار به صورتی که در فرمول نشان‌داده شده است، استفاده گردید. No برای هر دسته نشاندهنده تعداد خیابان‌های متعلق به دسته و n تعداد کل دسته‌ها است.

$$\text{w\_precision} = \frac{\sum_{k}^{n} precision(k) * No(k)}{\sum_{k}^{n} No(k)} \qquad (۳)$$

$$\text{w\_recall} = \frac{\sum_{k}^{n} \text{recall}(k) * No(k)}{\sum_{k}^{n} No(k)} \qquad (۴)$$

نتایج بدست آمده برای هر الگوریتم به صورت جداگانه در جدول‌های ۱ و ۲ نمایش داده شده‌اند. با توجه به پراکندگی نامتقارن خیابان‌ها در دسته‌های مختلف، میزان کیفیت دسته‌بندی برای هر برچسب، باید با توجه به تعداد خیابان‌های دارای همان برچسب بررسی گردد. تعداد بالای برچسب residential از یک‌سو و Precision و Recall بالای به دست آمده برای این برچسب از سوی دیگر، تأثیر زیادی بر Precision و Recall وزن‌دار کل دسته‌بندی دارد. لذا برای بررسی دقیق‌تر میزان کیفیت دسته‌بندی، یک Precision و Recall وزن‌دار برای بقیه برچسب‌ها

۱ True Positive
۲ False Positive
۳ False Negetive

بدون استفاده از برچسب محاسبه شد. Precision و Recall وزن‌دار به دست آمده به ترتیب برابر ۶۶/۸ و ۶۳/۴ برای روش درخت تصمیم و ۴۹/۶ و ۲۶/۲ برای روش شبکه عصبی با ساختار چند لایه هستند. نتایج نشان می‌دهد با وجود کاهش مقادیر نسبت به حالت استفاده از برچسب residential، مدل درخت تصمیم دارای کیفیت خوبی برای دسته‌بندی بقیه برچسب‌هاست.

جدول۱- دقت دسته‌بندی هر یک از برچسب‌های به کار رفته برای خیابان‌های تهران با استفاده از شبکه عصبی با ساختار چندلایه

برچسب	تعداد خیابان	شبکه عصبی با ساختار چند لایه	
		Precision	Recall
Residential	۲۸۴۲۸	۷۹	۹۶،۲
Secondary	۱۷۹۱	۲۴،۴	۴،۵
Tertiary	۲۵۲۲	۲۴،۲	۲۸،۹
Trunk	۹۹۰	۳۳،۳	۲۰،۷
Primary	۸۵۶	۳۰،۸	۱،۹
trunk_link	۱۲۷۸	۳۵،۸	۵۴،۵
secondary_link	۴۳۷	۱۰۰	۳۳
primary_link	۳۸۵	۸۲،۹	۲۹،۳
Unclassified	۱۳۸۸	۳۵،۹	۸،۱
motorway_link	۴۲	۰	۰
Footway	۱۵۰۷	۸۲،۵	۳۸،۳
Service	۱۳۴۳	۸۹،۸	۳۳،۷
Motorway	۴۴	۲۸،۶	۱۸،۲
living_street	۱۲۶۲	۹۷،۴	۴۸،۵
Steps	۴۰۵	۰	۰
Track	۱۲۱	۳۷،۵	۱۹،۴
tertiary_link	۲۲۷	۹۶	۳۶،۹
Road	۷۴	۰	۰
Pedestrian	۴۴	۰	۰
Weighted_average	۴۳۱۴۴	۶۹	۷۲،۴
Weighted_average*	۱۴۷۱۶	۴۹،۶	۲۶،۲

* بدون استفاده از نتایج برچسب residential

از عوامل موثر در مدل‌شدن یک دسته، تعداد خیابان‌های موجود در هر دسته، در پایگاه داده مورد بررسی است. در این تحقیق، بالاترین دقت دسته‌بندی برای هر دو الگوریتم درخت تصمیم و شبکه عصبی، برای دسته‌بندی خیابان‌های با برچسب مسکونی است که بیشترین تعداد خیابان را در داده‌های تحقیق داراست.

درحالی‌که در نتایج درخت تصمیم، معابر پیاده‌رو[1] و بزرگراه[2] که هر کدام شامل ۴۴ مورد در شهر تهران هستند، با دقت کمتری مدل شده‌اند و از اتصال‌دهنده بزرگراه که کمترین تعداد را در میان خیابان‌اتصال‌دهنده داراست نسبت به اتصال‌دهنده‌های دیگر دارای دقت پایین‌تری است.

جدول۲- دقت دسته‌بندی هر یک از برچسب‌های به کار رفته برای خیابان‌های تهران با استفاده از درخت تصمیم

برچسب	تعداد خیابان	درخت تصمیم	
		Precision	Recall
Residential	۲۸۴۲۸	۹۳	۹۵،۴
Secondary	۱۷۹۱	۴۴،۳	۴۲،۳
Tertiary	۲۵۲۲	۴۹،۹	۴۷،۷
Trunk	۹۹۰	۷۲،۱	۶۹،۶
Primary	۸۵۶	۴۰،۵	۴۱
trunk_link	۱۲۷۸	۷۳	۷۳،۸
secondary_link	۴۳۷	۶۴،۸	۷۲،۲
primary_link	۳۸۵	۷۴،۴	۶۷،۷
Unclassified	۱۳۸۸	۷۰	۵۹،۲
motorway_link	۴۲	۴۵،۵	۵۵،۶
Footway	۱۵۰۷	۸۹،۵	۸۵،۹
Service	۱۳۴۳	۷۸،۵	۷۳،۹
Motorway	۴۴	۵۵،۶	۴۵،۵
living_street	۱۲۶۲	۸۶،۳	۷۸،۹
Steps	۴۰۵	۹۳،۶	۹۵،۴
Track	۱۲۱	۶۳	۵۴،۸
tertiary_link	۲۲۷	۸۱،۷	۷۵،۴
Road	۷۴	۶۹،۲	۴۷،۴
Pedestrian	۴۴	۲۵	۱۰
Weighted_average	۴۳۱۴۴	۸۴،۱	۸۴،۵
Weighted_average*	۱۴۷۱۶	۶۶،۸	۶۳،۴

* بدون استفاده از نتایج برچسب residential

تعداد خیابان‌ها برای هر برچسب، در نتایج مربوط به شبکه عصبی تاثیرگذارتر بوده است. با توجه به نتایج ارائه شده در جدول ۱، شبکه عصبی برچسب‌های با تعداد کم خیابان را شناسایی نکرده و تعداد زیادی از خیابان‌های دیگر دسته‌ها را به صورت اشتباه با برچسب residential، دسته‌بندی می‌کند که این مورد در برچسب‌های با تعداد کم خیابان شدیدتر است. یکی از دلایل اصلی برای این قضیه در

[1] Pedestrain
[2] Motorway

روش جستجوی فضای فرضیه‌ای توسط دو الگوریتم است. به طوری‌که درخت تصمیم فضای فرضیه‌ای کاملی را جستجو می‌کند اما جستجوی فضای فرضیه‌ای در شبکه عصبی متفاوت بوده و به عوامل زیادی وابسته است که در نهایت می‌تواند به عدم جستجوی تمام فرضیه‌ها برسد.

با توجه به نتایج درخت تصمیم، خیابان‌های با خصوصیات خاص مثل خیابان‌های با برچسب اتصال‌دهنده و یا پله‌ها با دقت نسبتاً خوب مدل شده‌اند. همانطور که ذکر شد شبکه عصبی وابستگی بیشتری به تعداد خیابان‌های هر برچسب داشته و در این موارد با وجود ویژگی‌های خاص برای این برچسب‌ها در موارد با تعداد کم نمونه دقت پایینی دارد. این نوع خیابان‌ها به خاطر اینکه صرفاً در مناطق و شرایط خاصی قابل استفاده هستند، با خیابان‌های دیگر هم‌پوشانی کمتری داشته و با دقت بیشتری دسته‌بندی شده‌اند. درحالی‌که برچسب‌های خیابان اصلی[۱]، شریانی درجه اول[۲] و شریانی درجه دوم[۳] با وجود تعداد خیابان‌ها با این برچسب‌ها در نتایج الگوریتم، دقت نسبتاً پایینی دارند، زیرا خیابان‌ها از نظر خصوصیات به هم شبیه بوده و کاربران در تشخیص این خیابان‌ها از یکدیگر موفق نبوده‌اند.

به‌غیر از تاثیر عوامل ذکر شده در مدل شدن بهتر، می‌توان به تعداد بالای دسته‌ها در دسته‌بندی مربوط به خیابان‌ها، در OSM اشاره کرد، به صورتی که افراد غیر حرفه‌ای با بعضی از انواع خیابان‌ها به ندرت آشنایی دارند و بعضی از برچسب‌ها متعلق به دسته‌بندی همه کشورها نیست. به طور مثال خیابان با برچسب طبقه‌بندی نشده، در واقع یک برچسب برای یک نوع خیابان در انگلستان به شمار می‌آید. همانطور که در بخش‌های قبل اشاره شد، یکی از مشکلات شامل تفاوت در الگوی دسته‌بندی خیابان‌ها در پایگاه داده‌های متفاوت است که این تفاوت، شامل تفاوت در الگوی دسته‌بندی در کشورهای مختلف نیز می‌شود. ما با استفاده از تغییر در تعداد دسته‌های استفاده شده برای دسته‌بندی، شاهد تغییر در میزان دقت دسته‌بندی در بعضی گروه‌های خیابان‌های تهران بودیم. به صورتی که با کاهش در تعداد دسته‌ها، دقت دسته‌بندی افزایش یافت.

۵- نتیجه‌گیری

بررسی و تضمین کیفیت اطلاعات مکانی یک نیاز اساسی برای استفاده از داوطلبانه این اطلاعات است. روش‌های ارزیابی کیفیت اطلاعات مکانی داوطلبانه، به طور کلی به دو دسته مقایسه با اطلاعات مرجع و یا روش‌های بدون نیاز به اطلاعات مرجع تقسیم می‌گردد. اطلاعات مرجع همیشه وجود ندارد و یا در صورت وجود، در دسترس نیستند. علاوه بر این ماهیت اطلاعات مکانی داوطلبانه، بروزرسانی نقشه‌ها با سرعت بیشتر از اطلاعات مرجع است مثل فضای سبزی که به تازگی ساخته شده است، و همچنین اطلاعات مکانی داوطلبانه می‌تواند شامل اطلاعاتی مثل نام محلی خیابان‌ها باشد، که اطلاعات مرجع این اطلاعات را شامل نمی‌شود. لذا این روش‌های ارزیابی در بسیاری موارد موثر نیستند. درنتیجه باید تلاش گردد تا اطلاعات مکانی داوطلبانه با استفاده از خود این اطلاعات و یا قوانین حاکم بر این اطلاعات، ارزیابی گردد. در این تحقیق، روشی برای ارزیابی کیفیت اطلاعات مکانی داوطلبانه با استفاده از اطلاعات به دست آمده از خود این اطلاعات ارائه شده است. با توجه به اینکه بیشتر تحقیق‌های صورت گرفته به دقت هندسی و میزان کامل بودن اطلاعات پرداخته‌اند و تحقیقات کمی بر ارزیابی صحت معنایی اطلاعات مکانی داوطلبانه انجام گردیده است، ما به ارزیابی کیفیت داده‌های معنایی اطلاعات مکانی داوطلبانه و به صورت خاص دسته‌بندی خیابان‌ها در پایگاه داده OSM پرداختیم.

در طراحی خیابان‌های شهری، سهم هر خیابان در تامین دسترسی، جابه‌جایی و نقش اجتماعی در نظر گرفته شده و شاخص‌هایی از قبیل معیارهای هندسی، ترافیکی و کنترلی اعمال می‌شود. لذا با استفاده از این ویژگی‌ها می‌توان به کارکرد خیابان و دسته‌ای که خیابان متعلق به آن است، رسید.

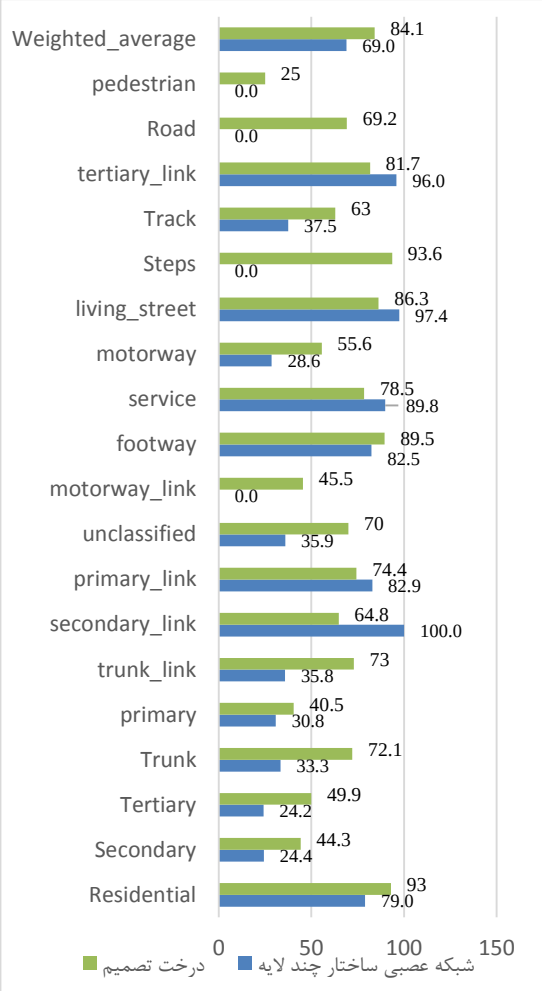
این روش با استخراج ویژگی‌های موثر در دسته‌بندی خیابانها و استفاده از ماشین‌های یادگیری شبکه عصبی و درخت تصمیم، به ایجاد مدلی برای ارزیابی صحت دسته‌بندی این خیابان‌ها می‌پردازد. ویژگی‌های استفاده شده شامل چگونگی اتصال خیابان‌ها به یکدیگر، میزان مستقیم بودن، طول، تعداد تقاطع‌ها و تعداد بن‌بست‌ها است.

---

۱ Primary
۲ Secondary
۳ Tertiary

برای ارزیابی کیفیت روشهای دسته‌بندی معابر، پارامترهای دقت و بازخوانی برای هر دو ماشین یادگیری به کار برده شد. مقایسه دقت دو روش برای دسته‌بندی در شکل۴ نشان‌دهنده عملکرد بهتر درخت تصمیم در برابر شبکه عصبی برای این داده‌هاست. دقت وزن‌دار حاصل برای دسته‌بندی با درخت تصمیم، ۸۴/۱٪، بیان‌گر مدل موفق برای ارزیابی کیفیت دسته‌بندی خیابان‌هاست.

شبکه خیابان‌های OSM دارای تعداد زیادی برچسب برای دسته‌بندی خیابان‌هاست که با دسته‌بندی استاندارد بسیاری از کشورها مطابقت ندارد. لذا یکی از راه‌حل‌های بالابردن کیفیت دسته‌بندی، یافتن یک دسته‌بندی بهینه، از دسته‌بندی‌های OSM برای کشور ایران است. همچنین روش‌های اندازه‌گیری‌های خصوصیات ذکر شده برای خیابان‌ها، مثل میزان مستقیم بودن، شامل طیف وسیعی از روش‌هاست که می‌تواند برای رسیدن به روش‌هایی با کارکرد بالاتر برای خیابان‌ها بررسی گردد. از اصلی‌ترین راه‌ها برای ارتقاء این روش، یافتن معیارهای جدید متمایز کننده برای دسته‌های متفاوت است.



شکل۴- مقایسه نتایج دقت یادگیری درختی و شبکه عصبی

مراجع

[1] 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, 682-703

[2] Koukoletsos, T., Haklay, M., and Ellul, C. (2012). Assessing data completeness of VGI through an automated matching procedure for linear data. Transactions in GIS 16, 477-498

[3] Bishr, M., and Janowicz, K. (2010). Can we trust information?-the case of volunteered geographic information. In Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium, volume 640.

[4] Keßler, C., Trame, J., and Kauppinen, T. (2011). Tracking editing processes in volunteered geographic information: The case of OpenStreetMap. In Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory.

[5] Keßler, C., and de Groot, R.T.A. (2013). Trust as a proxy measure for the quality of volunteered geographic information in the case of OpenStreetMap. In Geographic information science at the heart of Europe. Springer), pp 21-37

[6] F. Antonio, P. Fogliaroni and T. Kauppinen. (2014). "VGI Edit History Reveals Data Trustworthiness and User Reputation." AGILE 2014, Castellon, June 3-6.

[7] Ballatore, A., and Bertolotto, M. (2011). Semantically enriching VGI in support of implicit feedback analysis. In Web and Wireless Geographical Information Systems. (Springer), pp 78-93.

[8] Mooney, P., and Corcoran, P. (2012). Characteristics of heavily edited objects in OpenStreetMap. Future Internet 4, 285-305.

[9] Zielstra, D., Hochmair, H.H., and Neis, P. (2013). Assessing the effect of data imports on the completeness of OpenStreetMap–a United States case study. Transactions in GIS 17, 315-334.

[10] Jilani, M., Corcoran, P., and Bertolotto, M. (2014). Automated highway tag assessment of openstreetmap road networks .In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. (ACM), pp 449-452

[11] Forghani, M., and Delavar, M.R. (2014). A quality study of the OpenStreetMap dataset for Tehran. ISPRS International Journal of Geo-Information 3, 750-763.

[12] Hunter, G.J., and Goodchild, M.F. (1996). A new model for handling vector data uncertainty in geographic information systems. URISA-WASHINGTON DC- 8, 51-57.

[13] Mitchell, T.M. (1997). Machine learning. In. (McGraw-Hill, Inc., New York

[14] Hashemi, P., and Ali Abbaspour, R. (2015). Assessment of Logical Consistency in OpenStreetMap Based on the Spatial Similarity Concept. In OpenStreetMap in GIScience. (Springer), pp 19-36.

[15] Stojmenović, M., Nayak, A., and Zunic, J. (2008). Measuring linearity of planar point sets. Pattern Recognition 41, 2503-2511.

[16] Dorn, H., Törnros, T., and Zipf, A. (2015). Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany. ISPRS International Journal of Geo-Information 4, 1657-1671.

[17] Girres, J.F., and Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. Transactions in GIS 14, 435-459.

[18] Neis, P., Zielstra, D., and Zipf, A. (2011). The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007-2011. Future Internet 4, 1-21.

[19] Witten, I.H., and Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. (Morgan Kaufmann).