نشریه علمی مهندسی فناوری اطلاعات مکانی

سال دهم، شماره سوم، پاییز ۱۴۰۱

Vol.10, No.3, Autumn 2022

۷۱ -۹۴

مقاله پژوهشی

DOR: 20.1001.1.20089635.1401.10.3.4.2

ارزیابی عملکرد روش‌های مختلف یادگیری ماشین در شناسایی نوع حمل و نقل با استفاده ازداده‌های خط سیر

مرتضی طیبی[۱]، پرهام پهلوانی[۲*]

۱- دانشجوی دکتری سیستم‌های اطلاعات مکانی - دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی - دانشکدگان فنی - دانشگاه تهران

۲- دانشیار دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی - دانشکدگان فنی - دانشگاه تهران

تاریخ دریافت مقاله: ۱۴۰۱/۰۲/۲۶ تاریخ پذیرش مقاله: ۱۴۰۱/۱۱/۰۱

چکیده

با ظهور گسترده گوشی‌های هوشمند که به سامانه موقعیت‌یاب جهانی (*GPS*) مجهز هستند، حجم انبوهی از داده‌های مکانی خط سیر کاربران ایجاد شده است. مطالعه بر روی این داده‌ها در راستای تسهیل مدیریت شهری و ارائه مناسب خدمات به کاربران به‌عنوان یک زمینه تحقیقاتی گسترده مطرح شده و در حال رشد است. در این تحقیق به شناسایی نوع حمل‌ونقل خطوط سیر کاربران بر مبنای داده‌های خام *GPS* آن‌ها پرداخته شده است. این داده‌ها غالباً دارای خطاهایی هستند که در این تحقیق با اعمال یک فرآیند پیش‌پردازش چندمرحله‌ای سعی شده است مقدار خطا به حداقل برسد. سپس به‌منظور شناسایی نوع حمل‌ونقل شامل پیاده‌روی، دوچرخه، قطار، اتوبوس و رانندگی ویژگی‌های متعددی استخراج می‌شود. در ادامه به‌منظور ساختن مدل پیش‌بینی‌کننده از چهار روش طبقه‌بندی درخت تصمیم، شبکه عصبی پرسپترون چندلایه، بیز ساده و ماشین بردار پشتیبان استفاده می‌شود. در جهت بهبود عملکرد روش‌های پیاده‌سازی، از درصد حضور نقاط هر خط سیر در فاصله یک انحراف معیار از میانگین کل سرعت نوع‌های حمل‌ونقل به‌عنوان یک ویژگی جدید استفاده شده است. پیاده‌سازی چهار مدل یادشده به ازای پارامترهای تنظیم‌کننده مختلف انجام‌شده و پس از یک جستجوی جامع شبکه‌ای پارامترهای مختلف موجود در این روش‌ها در بهینه‌ترین مقدار تنظیم می‌شوند. در ادامه از دو شاخص کاپا و دقت کلی برای ارزیابی روش‌های مختلف استفاده می‌شود. نتایج حاصل از این تحقیق نشان داد که شبکه عصبی پرسپترون چندلایه با دقت کلی ۰٫۸۸ توانست بهترین نتایج را نسبت به سایر مدل‌ها از خود نشان دهد.

کلید واژه‌ها : داده‌های خط سیر، تعیین نوع حمل‌ونقل، طبقه‌بندی، یادگیری ماشین.

* نویسنده مکاتبه کننده: خیابان کارگر شمالی، دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی، دانشکدگان فنی، دانشگاه تهران.

تلفن: ۰۲۱۶۱۱۱۴۵۲۴

Email: pahlavani@ut.ac.ir

۱- مقدمه

امروزه گسترش محیط‌های شهری و افزایش روزافزون جمعیت در کلان‌شهرها چالش‌های زیادی مانند ترافیک، آلودگی هوا، تأمین تسهیلات شهری و غیره با خود به همراه داشته است. از سوی دیگر منابعی نظیر گوشی‌های تلفن همراه، شبکه‌های اجتماعی، سامانه موقعیت‌یابی جهانی[۱] (*GPS*) خودروها و غیره موجب تولید داده‌های مکانی انبوه و ارزشمند شده است. دولتمردان در راستای حل چالش‌های یادشده و اتخاذ تصمیم‌های حوزه مدیریت شهری نیاز به شناخت محیط‌های شهری و درک رفتار حرکتی شهروندان دارند که می‌تواند از منابع مذکور حاصل شود. محاسبات شهری[۲] اصطلاحی است که به این حوزه از مطالعات مکانی شهرها اختصاص داده‌شده است [۱].

پیشرفت در تکنیک‌های محاسبه حرکت و دریافت موقعیت، داده‌های خط سیر مکانی حجیمی را تولید کرده است که حرکت اشیا متحرک گوناگونی مانند مردم، وسایل نقلیه و حیوانات را ارائه می‌کند. تکنیک‌های زیادی برای پردازش، مدیریت و کاویدن داده‌های خط سیر در دهه گذشته پیشنهاد شده است که توسعه‌دهنده دامنه وسیعی از کاربردها است. هر خط سیر مکانی یک دنباله تولیدشده توسط یک شی متحرک در فضای جغرافیایی است که معمولاً به‌وسیله یک سری نقاط مرتب بر اساس زمان ارائه می‌شود. مثلاً $p_1 \rightarrow p_2 \rightarrow \cdots p_n$ که هر نقطه شامل یک مجموعه مختصات مکانی و یک لحظه زمانی است مانند $p = (x, y, t)$ [۲].

در چند سال اخیر، تحقیقات متعددی در حوزه داده‌های خط سیر مکانی انجام‌شده است. این تحقیقات از یک دیدگاه کلی به سه‌شاخه اصلی تقسیم می‌شوند:

۱-مدل‌های داده: تعریف و توسعه نوع داده‌های خط سیر مانند نقاط و نواحی مشترک[۳ و۴].

۲-مدیریت داده: ذخیره‌سازی مؤثر داده‌های حرکت با نمایه‌گذاری و تکنیک‌های پرس و جوی ویژه [۵, ۶] که منجر به توسعه چندین سیستم مدیریت پایگاه داده مانند *SECONDO* [۷]، *HERMES* [۸] و *DOMINO* [۹] شده است.

۳-داده‌کاوی: طراحی و پیاده‌سازی الگوریتم‌های داده‌کاوی نظیر خوشه‌بندی، طبقه‌بندی، کشف خطا و الگو کاوی بر روی داده‌های خط سیر مانند ژئولایف[۳]، *GeoPKDD* و *MoveMine* [۲].

داده‌کاوی خطوط سیر منجر به حصول نتایجی می‌شود که کاربردهای فراوانی نظیر شناسایی کاربران مشابه، تشخیص ناهنجاری‌های ترافیکی، شناسایی نقاط و مناطق جذاب و غیره دارند. تعیین نوع حمل‌ونقل[۴] یکی از کاربردهایی است که می‌توان با بهره‌گیری از تکنیک‌های داده‌کاوی به آن دست‌یافت. نوع حمل‌ونقل یکی از مهم‌ترین ویژگی‌ها در رفتارهای حرکتی کاربران است که در تحلیل تقاضای سفر، مدیریت ترافیک و برنامه‌ریزی حمل‌ونقل تأثیرگذار است [۱۰]. با استفاده از توزیع نوع سفر، آژانس‌های حمل‌ونقل می‌توانند استراتژی‌های مناسبی را برای کاهش زمان سفر کاربران، کاهش ترافیک و کاهش آلودگی هوا ایجاد کنند. به‌عنوان‌مثال یک نمونه بارز از کاربرد تحلیل‌های نوع سفر شناسایی مناطق وابسته و بهبود سامانه‌های حمل‌ونقل در راستای تشویق به استفاده از حمل‌ونقل عمومی است [۱۱]. اعمال محدودیت‌ها و قوانین ترافیکی مانند خطوط ویژه وسایل نقلیه عمومی[۵] در ساعات اوج شلوغی از دیگر کاربردهای مقدماتی تحلیل‌های نوع سفر است. پیش‌ازاین اطلاعات مربوط به نوع سفر از طریق مصاحبه‌های حضوری و تلفنی انجام

[۱] *Global Positioning System*

[2] *Urban computing*

[۳] *GeoLife*

[۴] *Transportation mode*

[5] *High Occupancy Vehicle (HOV) Lanes*

می‌پذیرفت که یک روش پرهزینه، کم‌دقت و ناقص بود، اما امروزه دسترسی ما به داده‌های خط سیر اطلاعات دقیق، کم‌هزینه، به‌روز و کاملی را در اختیار می‌گذارد [۱۰].

وجود منابع خطا داده‌های خام خطوط سیر را تحت تأثیر قرار می‌دهد و این خطا به ویژگی‌های خطوط سیر انتشار می‌یابد و دقت مدل‌ها را تحت تأثیر قرار می‌دهد[۱۲]. بنابراین پالایش داده‌ها و کاهش خطا جهت بهبود دقت شناسایی نوع‌های حمل‌ونقل ضروری است، در این راستا در این تحقیق یک فرآیند پیش‌پردازش چندمرحله‌ای جهت حذف داده‌های اشتباه و کاهش خطا پیاده‌سازی می‌شود و تأثیر آن بر دقت نتایج موردبررسی قرار می‌گیرد.

برای شناسایی نوع حمل‌ونقل با استفاده از داده‌های خط سیر، استخراج ویژگی‌ها از اهمیت بالایی برخوردار است. در این تحقیق علاوه بر بهره گرفتن از مهم‌ترین و تأثیرگذارترین ویژگی‌هایی که در تحقیقات پیشین معرفی شده‌اند [۱۰، ۱۲ و۱۳]، درصد حضور نقاط هر خط سیر در فاصله یک انحراف معیار از میانگین کل سرعت نوع‌های حمل‌ونقل به‌عنوان یک ویژگی جدید معرفی می‌شود. همچنین رفتار نوع‌های مختلف حمل‌ونقل در بازه‌های سرعت مختلف در جهت بهبود دقت شناسایی نوع‌های حمل‌ونقل موردبررسی قرار می‌گیرد.

در سال‌های اخیر تحقیقات گسترده‌ای در راستای تشخیص نوع حمل‌ونقل بر مبنای داده‌های خط سیر *GPS* انجام‌شده است. این تحقیق تلاش می‌کند عملکرد روش‌های مختلف یادگیری ماشین را به ازای پارامترهای مختلف مورد بحث قرار داده و نتایج آن‌ها را مقایسه کند.

در ادامه این مقاله، در بخش دوم مطالعات انجام‌شده در زمینه تعیین نوع حمل‌ونقل در داده‌های خط سیر مرور می‌گردد. بخش سوم به تشریح مبانی نظری تحقیق و بخش چهارم به بیان روش پیاده‌سازی تحقیق می‌پردازد. در بخش پنجم ضمن معرفی داده‌های موردمطالعه، جزئیات پیاده‌سازی روش تحقیق تشریح می‌شود. در بخش ششم نتایج تحقیق ارائه شده و مورد بحث قرار می‌گیرد. درنهایت، در بخش هفتم نتیجه‌گیری حاصل از تحقیق بیان شده و پیشنهادهایی برای تحقیقات آینده ارائه خواهد شد.

۲- مطالعات پیشین

در سال‌های اخیر داده‌کاوی بر روی داده‌های خام *GPS* که منجر به ایجاد خطوط سیر می‌شوند، به یک موضوع تحقیقاتی گسترده تبدیل شده است. این مطالعات شامل استخراج داده‌ها، پیش‌پردازش، مدیریت پایگاه داده و بازیابی اطلاعات، خوشه‌بندی و طبقه‌بندی، الگو کاوی خطوط سیر، عدم قطعیت[۱] و زمینه‌های دیگری است که در همکاری طی یک فرآیند داده‌کاوی به کاربردهای مختلفی منجر می‌شوند [۲ و۱۴]. در برخی تحقیقات مانند [۱۵ و۱۶] خطوط سیر به فرمت‌های دیگری مانند گراف، ماتریس و تنسور تبدیل می‌شوند تا تکنیک داده‌کاوی بیشتری بر روی آن‌ها قابل پیاده‌سازی باشد.

در تحقیقات داده‌کاوی خطوط سیر غالباً از داده‌های خام *GPS*، داده‌های شتاب سنج گوشی‌های همراه و داده‌های سامانه جهانی موبایل[۲] (*GSM*) استفاده می‌شود [۱۰]. هم‌چنین افزودن برخی از داده‌های زمینه‌ای مانند دمای بدن، ضربان قلب، رطوبت و شدت نور در کنار داده‌های مکانی نظیر شبکه راه‌ها، نقاط جذاب[۳] و کاربری‌ها در گروهی از مطالعات موردتوجه قرارگرفته‌اند [۱۳]. در این زمینه مطالعات قابل‌توجهی با به‌کارگیری تکنیک‌های داده‌کاوی ازجمله مدل‌های طبقه‌بندی، بر روی برآورد نوع حمل‌ونقل، برآورد نوع فعالیت و برآورد مقصد متمرکز شده‌اند. برخی از تحقیقات روش‌هایی برای برآورد نوع حمل‌ونقل ارائه

[۱] *Uncertainty*

[۲] *Global System for Mobile*

[۳] *Point Of Interest*

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]

کردند که از داده‌های ثابتی مانند نقشه راه‌ها استفاده می‌کند [۱۷ و۱۸]. با توجه به ماهیت پویای ساختار شهری دسترسی آسان و به‌روز به این‌گونه اطلاعات خارجی میسر نیست [۱۳].

گروه تحقیقاتی شرکت مایکروسافت به سرپرستی ژنگ تحقیقات زیادی بر روی داده‌های خط سیر انجام داده‌اند. یکی از مهم‌ترین تحقیقات انجام‌شده توسط این گروه پروژه ژئولایف بود که هدف آن ارائه یک شبکه اجتماعی مکان‌مبنا بود که بر مبنای داده‌های خط سیر *GPS* که در شبکه به اشتراک گذاشته می‌شد عمل نماید [۱۹]. ژنگ و همکاران (۲۰۰۸) مدلی برای برآورد نوع حمل‌ونقل بر مبنای داده‌های پروژه ژئولایف که داده خام *GPS* محسوب می‌شوند، ارائه کردند. آن‌ها ابتدا با یک روش قطعه‌بندی نقطه مبنا تلاش کردند خطوط سیر با یک نوع حمل‌ونقل خاص ایجاد کنند. سپس به استخراج ویژگی‌های مهمی برای خطوط سیر پرداختند که پس از آن در تحقیقات دیگری نیز موردتوجه قرار گرفتند. ازجمله مهم‌ترین ویژگی‌های استفاده شده در این تحقیق سرعت، انحراف معیار سرعت، شتاب، نرخ توقف، تغییر جهت و غیره بود. سپس از با پیاده‌سازی مدل‌های طبقه‌بندی به مدلی برای پیش‌بینی نوع حمل‌ونقل دست یافتند [۲۰]. سپس در تحقیق دیگری در همان سال با افزودن یک مرحله پس پردازشی گراف مبنا به بهبود مدل پیشنهادی خود پرداختند. این مرحله پس پردازش بر مبنای احتمال وقوع هر نوع حمل‌ونقل در یک دنباله از نوع‌های حمل‌ونقل عمل می‌کند [۲۱]. آن‌ها در تحقیقات خود از مدل‌های مختلف طبقه‌بندی استفاده کردند و بهترین نتایج را از مدل‌های رندم فارست[۱] و درخت تصمیم[۲] (*DT*) به دست آوردند.

عمرانی و همکاران (۲۰۱۵) از شبکه عصبی مصنوعی[۳] (*ANN*) ازجمله شبکه عصبی پرسپترون چندلایه[۴] (*MLP*) و تابع پایه شعاعی[۵] (*RBF*) در راستای پیش‌بینی نوع حمل‌ونقل استفاده کردند. اندو و همکاران (۲۰۱۶) مدلی ارائه دادند که خطوط سیر را به یک تصویر دوبعدی تبدیل می‌کند و مقدار هر پیکسل برابر با مدت‌زمانی است که کاربر در آن مکان حضور داشته است. سپس با معرفی این تصویر به یک شبکه عصبی عمیق تعدادی ویژگی خودکار استخراج شده و پس از ادغام با ویژگی‌های سنتی مرسوم به مدل طبقه‌بندی معرفی می‌شود [۱۳]. آن‌ها روش‌های ماشین بردار پشتیبان[۶] (*SVM*)، درخت تصمیم و رگرسیون لجستیک[۷] پیاده‌سازی کردند و درخت تصمیم را به‌عنوان بهترین مدل معرفی کردند. این تحقیق علیرغم پیچیدگی زمانی بالایی که دارد موفق به کسب دقت قابل قبولی نشد. در تحقیق مشابه دیگری ونگ و همکاران (۲۰۱۷) با استخراج ویژگی‌هایی در سطح نقطه و معرفی این ویژگی به یک شبکه عصبی پیچشی[۸] ویژگی‌های بیشتری استخراج کرده و به مدل طبقه‌بندی معرفی کردند [۲۲].

در تحقیق دیگری خیائو و همکاران (۲۰۱۷) با تولید تعداد صد و یازده ویژگی آماری در سطح کل خطوط سیر و در سطح محلی برای قطعه‌های کوچک‌تر خطوط سیر، مدلی بر مبنای درخت تصمیم ارائه کردند [۲۳]. دبیری و همکاران در سال ۲۰۱۹، با به‌کارگیری شبکه عصبی پیچشی چهار ویژگی اصلی سرعت، شتاب، تغییرات شتاب و تغییر جهت را به لایه ورودی شبکه

[۱] *Random Forest*

[۲] *Decision Tree*

[۳] *Artificial Neural Network*

[۴] *Multi Layer Perceptron*

[۵] *Radial Basis Function*

[۶] *Support Vector Machine*

[۷] *Logistic Regression*

[۸] *Convolutional Neural Network*

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]

معرفی کردند و ویژگی‌های سطح بالا به شکل خودکار توسط شبکه استخراج شدند. با توجه به اینکه خطوط سیر با طول‌های متنوع قابل‌ارائه به ورودی شبکه عصبی پیچشی نبودند، آن‌ها هر خط سیر را به قطعات ۲۰۰ نقطه‌ای تقسیم کردند [۲۴].

نواز و همکاران (۲۰۲۰) نیز از شبکه عصبی پیچشی جهت استخراج ویژگی‌های سطح بالا استفاده کردند، سپس از مدل حافظه طولانی کوتاه‌مدت[۱] (*LSTM*) برای آموزش مدل شناسایی نوع حمل‌ونقل استفاده کردند. آن‌ها با افزودن داده‌های هواشناسی به لایه ویژگی‌ها موفق شدند دقت مدل پیشنهادی خود را هفت درصد نسبت به تحقیقات پیشین بهبود بخشند [۲۵]. جیمز و همکاران نیز در سال ۲۰۲۰، ضمن استفاده از شبکه عصبی عمیق از تبدیل موجک گسسته برای استخراج ویژگی‌های دامنه زمان بسامد برای کمک به شبکه عصبی استفاده کردند [۲۶].

در تحقیق دیگری هوانگ و همکاران (۲۰۲۰) با به‌کارگیری مدل جنگل ایزوله مشارکتی[۲] هر نوع حمل‌ونقل را به شکل جداگانه در یک مدل شناسایی کرده و نتایج را تلفیق کردند. این تحقیق نشان داد که مدل جنگل ایزوله مشارکتی در برخورد با داده‌های خط سیری که دارای خطای اتفاقی هستند انعطاف بیشتری نسبت به سایر مدل‌ها نشان می‌دهد [۲۷].

تحقیقات گذشته در راستای تشخیص نوع حمل‌ونقل داده‌های *GPS*، مدل‌های متنوعی را به کار گرفته‌اند و در این راستا ویژگی‌های مختلفی را از روش‌های متفاوت استخراج کرده‌اند. برخی از تحقیقات نیز از داده‌های کمکی دیگری در جهت بهبود عملکرد مدل پیشنهادی خود استفاده کرده‌اند. تحقیق حاضر بر مبنای داده‌های خام *GPS* و بدون کمک داده‌های مکانی خارجی، از چهار مدل طبقه‌بندی برای پیش‌بینی نوع حمل‌ونقل داده‌های خط سیر استفاده می‌کند. با توجه به اینکه مناطق شهری دارای نقاط کور برای دریافت سیگنال توسط *GPS* های کاربران هستند و وجود منابع خطای دیگر، انجام یک فرآیند پیش‌پردازش چندمرحله‌ای در این تحقیق موردتوجه قرارگرفته است که در تحقیقات گذشته غالباً کمتر به آن توجه شده است. همچنین در این تحقیق ویژگی‌های جدیدی برای خطوط سیر ارائه می‌شود.

در این تحقیق از چهار روش ماشین بردار پشتیبان، درخت تصمیم، شبکه عصبی پرسپترون چندلایه و بیز ساده جهت پیاده‌سازی مدل طبقه‌بندی استفاده می‌شود. روش ماشین بردار پشتیبان که در بسیاری از زمینه‌های تحقیقاتی عملکرد مناسبی از خود نشان داده است، در این تحقیق نیز مورداستفاده قرارگرفته است. در تحقیقات گذشته روش *SVM* برای طبقه‌بندی خطوط سیر بر اساس نوع حمل‌ونقل کمتر موردتوجه قرارگرفته و نتوانسته است دقت مناسبی از خود نشان دهد، درخت تصمیم نیز که پیش‌ازاین در تحقیقات مشابه با موفقیت به کار گرفته‌شده است، در این تحقیق با اعمال ضریب جینی، بهره اطلاعاتی و نسبت بهره به‌عنوان معیارهای طبقه‌بندی، مورداستفاده قرار می‌گیرد. شبکه‌های عصبی امروزه در تحقیقات مختلفی توانایی خود را در یادگیری رفتار داده‌های مختلف و پیش‌بینی آن‌ها نشان داده‌اند. در این تحقیق از شبکه عصبی پرسپترون چندلایه استفاده شده است. روش طبقه‌بندی کننده بیز ساده نیز که پیش‌ازاین به‌ندرت در تحقیقات مشابه موردتوجه قرارگرفته است، در این تحقیق مورد ارزیابی قرار می‌گیرد. این تحقیق تلاش می‌کند با یک جست‌وجوی جامع بهترین پارامترهای ورودی برای روش‌های مختلف طبقه‌بندی را کشف کند و نتایج حاصل با نتایج تحقیقات پیشین مقایسه شود.

۳- مبانی نظری تحقیق

در این بخش به تشریح مبانی نظری تحقیق پرداخته می‌شود. در این تحقیق از چهار روش طبقه‌بندی

[۱] *Long Short Term Memory*

[۲] *Collabrative Isolation Forest*

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]

مختلف برای ایجاد مدل پیش‌بینی کننده نوع حمل‌ونقل داده‌های خط سیر استفاده شده است.

۳-۱- ماشین بردار پشتیبان

روش ماشین بردار پشتیبان نخستین بار توسط وپنیک(۱۹۹۲) ارائه شد [۲۸]. هر ماشین بردار پشتیبان متشکل از یک یا چند ابرصفحه[۱] در فضای ابعاد بالا است که برای انجام وظیفه طبقه‌بندی یا رگرسیون کاربرد دارد. چنانچه کلاس‌ها به شکل خطی قابل جداسازی نباشند، اعمال یک تابع کرنل و انتقال داده‌های به فضای دیگر کارگشا خواهد بود. بدیهی است که هر چه فاصله نزدیک‌ترین داده از مجموعه داده‌های آموزشی از ابرصفحه بیشتر باشد، دقت روش در طبقه‌بندی افزایش می‌یابد. این فاصله حاشیه[۲] نامیده می‌شود. در اکثر مسائل طبقه‌بندی واقعی طبقات را نمی‌توان به شکل خطی از یکدیگر جدا کرد و پذیرش مقداری خطا غیرقابل‌اجتناب است [۲۹]. رابطه(۱) مسئله بهینه‌سازی نهایی را نشان می‌دهد.

رابطه(۱)

$$\min \frac{1}{2}W^2 + C\sum_{i=1}^{n}\xi_i \quad , y_i\left(w.x_i + b\right) - 1 + \xi_i \geq 0 \quad \forall_i$$

در رابطه(۱)، w و b پارامترهای ابرصفحه، ξ متغیر کمکی[۳] و C پارامتر تنظیم[۴] است. مقدار حاشیه $\frac{2}{\|w\|}$ خواهد بود.

۳-۲- درخت تصمیم

درخت تصمیم یک الگوریتم یادگیری ماشین نظارت شده است که از درخت‌ها برای نمایش استفاده می‌کند. درخت تصمیم از گره ریشه شروع می‌کند و بر اساس یک ویژگی[۵] داده‌ها را به دو یا چند زیر درخت تقسیم می‌کند به‌طوری‌که مشابه‌ترین داده‌ها کنار یکدیگر قرار گیرند. این روند تا زمانی که تمام داده‌ها ازنظر یک ویژگی هدف در یک طبقه قرار گیرند، ادامه پیدا می‌کند. انتخاب بهترین ویژگی برای شروع طبقه‌بندی و هم‌چنین یافتن یک حد آستانه مناسب برای اعمال طبقه‌بندی مهم‌ترین سؤال‌ها در این روش هستند. درخت تصمیم به‌طورکلی از مفهومی به نام آنتروپی[۶] برای بهینه‌سازی مسئله استفاده می‌کند [۳۰]. رابطه(۲) نحوه محاسبه آنتروپی را نمایش می‌دهد.

$$E(D) = -\sum_{i=1}^{m} P_i \log_2\left(P_i\right) \qquad \text{رابطه(۲)}$$

که در رابطه (۲)، P_i احتمال تعلق داده‌ها به کلاس i ام و m تعداد کلاس‌ها هستند.

بهره اطلاعاتی[۷]، نسبت بهره[۸] و ضریب جینی[۹] مهم‌ترین معیارهایی هستند که در مدل‌های درخت تصمیم استفاده می‌شود. بهره اطلاعاتی حاصل از طبقه‌بندی داده‌ها بر مبنای ویژگی A از رابطه (۳) محاسبه می‌شود و در هر مرحله از تقسیم ویژگی با بیشترین بهره اطلاعاتی جهت طبقه‌بندی داده‌ها انتخاب می‌شود [۳۱].

$$Gain(A) = E(D) - \sum_{j=1}^{n} \frac{|D_j|}{|D|} \times E\left(D_j\right) \qquad \text{رابطه(۳)}$$

در رابطه(۳)، $E(D)$ آنتروپی و $\frac{|D_j|}{|D|}$ وزن بخش j ام هستند.

معیار بهره اطلاعاتی به سمت داده‌ها با تنوع مقداری بالا دارای انحراف است [۳۱]. معیار نسبت بهره این مشکل را برطرف می‌کند و از رابطه(۴) محاسبه می‌شود.

---

[۱] *HyperPlane*

[۲] *Margin*

[۳] *Slack Variable*

[۴] *Regularization Parameter*

[۵] *Attribute*

[۶] *Entropy*

[۷] *Information Gain*

[۸] *Gain Ratio*

[۹] *Gini index*

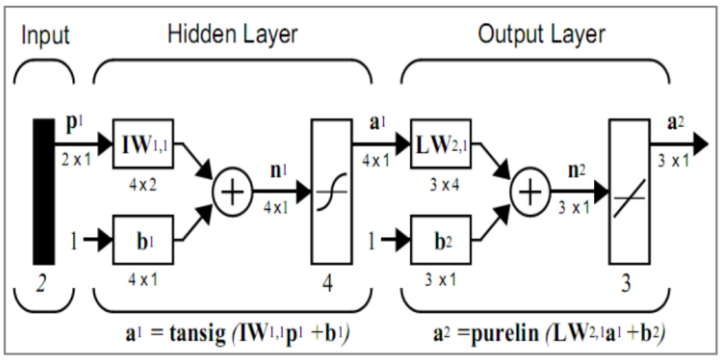
رابطه(۴)
$$GainRatio(A)=\frac{Gain(A)}{-\sum_{j=1}^{n}\frac{|D_j|}{|D|}\times\log_2\left(\frac{|D_j|}{|D|}\right)}$$

معیار ضریب جینی نیز یکی پرکاربردترین معیارها در مدل‌های درخت تصمیم است. هنگام استفاده از این ضریب مراحل طبقه‌بندی داده‌ها به نحوی ادامه پیدا می‌کند که ضریب جینی حداقل شود. ضریب جینی طبق رابطه (۵) محاسبه می‌شود که در رابطه (۵)، P_i احتمال تعلق داده‌ها به کلاس i ام و m تعداد کلاس‌ها هستند [۳۱].

رابطه(۵)
$$Gini(D)=1-\sum_{i=1}^{m}P_i^2$$

۳-۳- شبکه عصبی پرسپترون چندلایه

شبکه‌های عصبی مصنوعی الهام گرفته از مغز انسان و متشکل از واحدهای پردازشگری به نام نورون هستند. ساختار غیرخطی و تعمیم‌پذیری بالا ازجمله ویژگی‌های مهم این شبکه‌ها هستند. شبکه عصبی پرسپترون چندلایه (*Multi Layer Perceptron* (*MLP*)) در زمره پرکاربردترین شبکه‌های عصبی در مسائل مختلف داده‌کاوی است. این شبکه حداقل دارای سه لایه است. آموزش در این شبکه بر اساس قانون پس انتشار (*Back Propagation*) خطا صورت می‌گیرد و مقدار خطا پس از محاسبه در مسیر برگشت از طریق لایه‌های شبکه عصبی در کل شبکه توزیع می‌شود [۳۲]. شکل(۱) یک شبکه *MLP* سه لایه را نمایش می‌دهد که در آن بایاس در لایه‌ها با نماد b و وزن میان نورون‌ها با نماد IW نمایش داده‌شده است. *Purelin* و *tansig* نیز توابع عملکرد در لایه خروجی و میانی هستند.



شکل ۱: شبکه عصبی پرسپترون چندلایه

۳-۴- بیز ساده

روش بیز ساده[۱] (*NB*) با فرض استقلال متغیرهای تصادفی یک مدل احتمالاتی است که بر اساس قضیه بیز عمل می‌کند. این روش نظارت‌شده بر مبنای احتمال وقوع طبقه‌بندی‌ها را انجام می‌دهد [۳۱]. بر مبنای قضیه بیز احتمال وقوع کلاس C_i برای داده X طبق رابطه(۶) قابل‌محاسبه است.

رابطه(۶)
$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$

با مستقل در نظر گرفتن متغیرهای تصادفی $x_1, x_2, \dots, x_n$ کلاس هدف برای داده X مطابق رابطه(۷) قابل برآورد است که در آن K تعداد کلاس‌های هدف است.

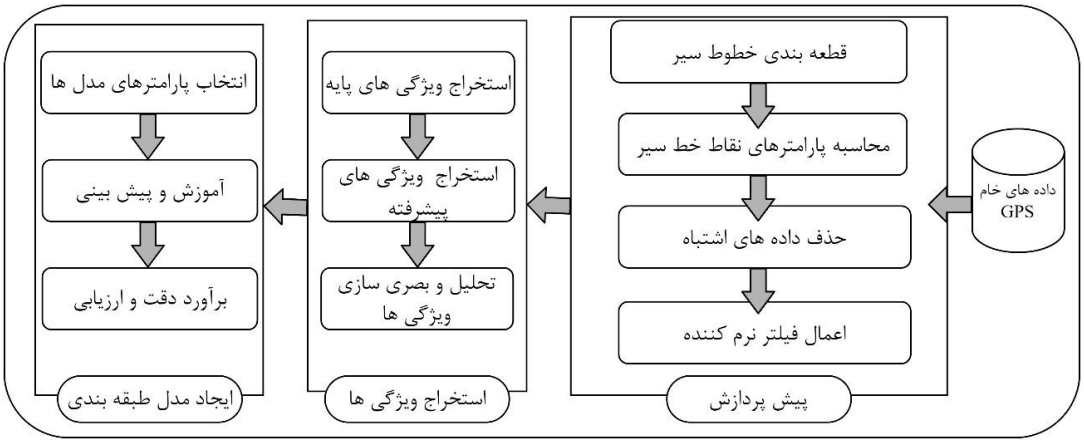
رابطه(۷)
$$\hat{y}=\operatorname{argmax}\left(P(C_k)\prod_{i=1}^{n}P(x_i|C_k)\right)\ ,\ k=1:K$$

۴- روش تحقیق

این تحقیق تلاش می‌کند با استخراج ویژگی از خطوط سیر داده‌های *GPS*، خطوط سیر را بر اساس نوع سفر

[۱] *Naive Bayes Classifier*

طبقه‌بندی کرده و عملکرد مدل‌ها و ویژگی‌ها را در پیش‌بینی نوع سفر موردبررسی قرار دهد. مراحل روش پیاده‌سازی این تحقیق در شکل(۲) نمایش داده شده است.



شکل ۲: مراحل روش پیاده‌سازی تحقیق

۴-۱- پیش‌پردازش و آماده‌سازی داده‌ها

داده‌های خام *GPS* شامل طول و عرض جغرافیایی و زمان برای هر نقطه هستند. در نخستین مرحله پنج پارامتر فاصله، سرعت، شتاب، تغییرات شتاب[۱] و تغییر جهت (زاویه) برای هر نقطه در داده‌های خام *GPS* محاسبه می‌گردد تا در ادامه تحقیق مورداستفاده قرار گیرند. به‌منظور محاسبه فاصله بین نقاط فرمول وینچنتی[۲][۳۳] مورد بهره‌برداری قرار گرفت، سپس پارامترهای سرعت، شتاب، تغییرات شتاب و تغییر جهت از روابط (۸)، (۹)، (۱۰) و (۱۱) به دست می‌آیند.

رابطه(۸)
$$S_{P_i} = \frac{\text{vincenty}\left(P_i, P_{i+1}\right)}{\Delta t = t_{i+1} - t_i}$$

رابطه(۹)
$$A_{P_i} = \frac{S_{i+1} - S_i}{\Delta t}$$

رابطه(۱۰)
$$J_{P_i} = \frac{A_{i+1} - A_i}{\Delta t}$$

رابطه(۱۱)
$$B_{P_i} = \left|H_{i+1} - H_i\right|$$

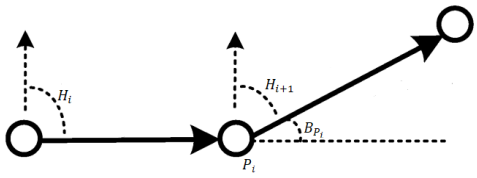
در روابط (۸)، (۹)، (۱۰) و (۱۱)، S_{P_i}، A_{P_i}، J_{P_i} و B_{P_i} به ترتیب سرعت، شتاب، تغییرات شتاب، تغییر جهت و زاویه با شمال برای نقطه P_i هستند. شکل(۳) پارامترهای رابطه(۱۱) را نمایش می‌دهد.

داده‌های خام *GPS*، مجموعه‌ای متوالی از نقاط هستند که از مکان یک کاربر در زمان مشخصی ثبت شده‌اند. این داده‌ها به قسمت‌های کوچک‌تری به نام سفر[۳] تقسیم می‌شوند. هر سفر بخشی از داده‌ها است که فاصله زمانی بیشتر از یک حد آستانه مشخص در آن روی نداده باشد [۲۰]. سپس سفرها بر اساس نوع حمل‌ونقل قطعه‌بندی (*Segmentation*) می‌شوند که هر قطعه دارای یک نوع حمل‌ونقل مشخص خواهد بود. ازاین‌پس در این تحقیق هر یک از این قطعه‌ها، خط سیر نامیده می‌شوند. داده‌های خام *GPS* ممکن است به دلیل وجود خطاهای مختلف نیازمند پیش‌پردازش‌هایی جهت حذف ثبت‌های اشتباه و کاهش خطاهای اتفاقی باشند.

[۱] *Jerk*

[۲] *Vincenty's formulae*

[۳] *Trip*



شکل۳: نمایش تغییر جهت (زاویه) برای یک نقطه از خط سیر

در این تحقیق مراحل زیر برای پالایش داده‌ها انجام می‌گیرد:

- حذف قطعه‌های موردمطالعه با طول کمتر از ۱۰ نقطه
- حذف نقاطی که لحظه زمانی آن‌ها بزرگ‌تر یا مساوی نقطه بعدی است.
- حذف نقاطی که سرعت یا شتاب آن‌ها بر اساس قوانین و قابلیت‌ها غیر معتبر باشد. مثلاً سرعت بیش از ۲٫۵ متر بر ثانیه برای نوع پیاده‌روی انسان غیر معتبر است.
- اعمال یک کرنل نرم‌کننده برای حذف خطاهای اتفاقی

در این تحقیق از فیلتر ساویتزکی-گلی[۱] به‌منظور نرم کردن خطوط سیر استفاده شده است. به‌طورکلی نرم کردن فرآیندی است که طی آن هر نقطه از داده‌ها توسط همسایگان خود تحت یک تابع کرنل مشخص قرار می‌گیرد و به‌این‌ترتیب خطای اتفاقی کاهش پیدا می‌کند. از آنجایی هیچ اطلاعات قبلی در مورد شکل خطوط سیر وجود ندارد، باید از فیلتری استفاده شود که به شکل پیش‌فرض داده‌ها را به شکل خاصی متمایل نکند و روند خط سیر را دچار تغییر ننماید. فیلتر ساویتزکی-گلی ابتدا یک چندجمله‌ای درجه m به تعداد فرد n نقطه از داده‌ها که پنجره نامیده می‌شوند و نقطه موردنظر در مرکز این پنجره قرار دارد، برازش می‌دهد. سپس مقدار جدید نقطه مرکزی توسط تابع چندجمله‌ای برآورد می‌شود [۳۴]؛ بنابراین با این روش شکل کلی و الگوی خط سیر حفظ می‌شود.

[۱] *Savitzky-Golay Filter*

۴-۲- استخراج ویژگی‌ها

پس‌ازاینکه خطوط سیر پالایش‌شده از داده‌های خام به دست آمد، به‌منظور اعمال مدل‌های طبقه‌بندی ویژگی‌هایی از خطوط سیر که در میان نوع‌های مختلف حمل و نقل متفاوت هستند، استخراج می‌شود. به‌عنوان‌مثال ویژگی سرعت در نوع پیاده‌روی و قطار تفاوت چشم‌گیری خواهد داشت. در این بخش ویژگی‌های مختلفی که در این تحقیق مورد استفاده قرار گرفته‌اند، معرفی می‌شوند.

۴-۲-۱- ویژگی‌های پایه

طول خط سیر و مدت‌زمان آن مهم‌ترین ویژگی هر خط سیر هستند. از آنجایی کاربران عموماً با توجه به همین ویژگی‌ها نوع حمل‌ونقل خود را انتخاب می‌کنند، بنابراین این ویژگی‌ها می‌تواند در بهبود عملکرد مدل‌های پیش‌بینی کننده مؤثر باشند.

سرعت یکی از ویژگی‌هایی است که در نوع‌های مختلف حمل‌ونقل مقدار متفاوتی دارد و می‌تواند در تمیز این نوع‌ها توسط مدل‌های پیش‌بینی کننده نقش ایفا کند. ویژگی‌هایی مانند سرعت میانگین، امید ریاضی سرعت، حداکثر سرعت و انحراف معیار سرعت ازجمله ویژگی‌های مشتق شده از سرعت نقاط خط سیر هستند که در تحقیقات پیشین موردتوجه قرارگرفته‌اند و در این تحقیق نیز استفاده می‌شوند. از دیگر ویژگی‌های مهمی که می‌تواند در تعیین نوع حمل‌ونقل مفید باشد، شتاب است؛ زیرا شتاب وسایل مختلف حمل‌ونقل مانند اتوبوس، خودرو شخصی و قطار متفاوت است. میانگین شتاب، انحراف معیار شتاب و حداکثر شتاب ویژگی‌های مورداستفاده در این تحقیق هستند. کاربران زمانی که در حال پیاده‌روی هستند، معمولاً تغییر جهت‌های زیادی دارند، درحالی‌که در نوع حمل‌ونقلی قطار ممکن

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]

است به‌ندرت جهت تغییر کند. همچنین میزان تغییر جهت برای سایر حمل‌ونقل‌ها نیز متغیر است؛ بنابراین میانگین زاویه تغییر جهت نیز در این تحقیق مورداستفاده قرارگرفته است.

۴-۲-۲- ویژگی‌های پیشرفته

در نوع حمل‌ونقل پیاده‌روی یا دوچرخه ممکن است کاربران بارها جهت حرکت خود را به‌طور محسوسی تغییر دهند، درحالی‌که انتظار می‌رود در سایر نوع‌ها تغییرات محسوس کمتر باشد، مثلاً در نوع قطار به‌ندرت جهت حرکت به شکل محسوس تغییر می‌کند و این تعداد احتمالاً برای اتوبوس و خودرو بیشتر خواهد بود. در نظر گرفتن درصد تغییر جهت‌های محسوس به‌عنوان یک ویژگی در بهبود عملکرد مدل‌های پیش‌بینی‌کننده می‌تواند مفید باشد. در این تحقیق با در نظر گرفتن یک حد آستانه δ برای تشخیص تغیر جهت‌های محسوس این ویژگی برای خط سیری با n نقطه میانی طبق رابطه (۱۲) محاسبه می‌شود که در آن B_{P_i} تغییر جهت برای نقطه P_i است [۲۰].

$$\frac{\left|B_{P_i} > \delta\right|}{n} \quad , \quad i = 1:n \qquad \text{رابطه(۱۲)}$$

در نوع‌های حمل‌ونقلی مختلف ممکن است به دلایل متفاوتی بارها سرعت حرکت به صفر نزدیک شود و کاربر متوقف باشد. مثلاً بازدید از یک فروشگاه در نوع پیاده‌روی ممکن است کاربر را متوقف کند. در مقایسه نوع‌های اتوبوس و رانندگی نیز این ویژگی می‌تواند تأثیرگذار باشد، زیرا اتوبوس‌ها برای سوار کردن مسافران تعداد دفعات بیشتری متوقف می‌شوند. در این تحقیق با در نظر گرفتن یک حد آستانه برای سرعت δ، درصد نقاط متوقف محاسبه‌شده و به‌عنوان ویژگی خطوط سیر در روش‌های طبقه‌بندی مورداستفاده قرار می‌گیرد [۲۰]. سه ویژگی مشابه دیگر نیز بر اساس درصد نقاط خطوط سیر در بازه‌های سرعتی معین محاسبه و استفاده خواهد شد (رابطه(۱۳)).

$$\frac{\left|V_i > \delta\right|}{n} \quad , \quad i = 1:n \qquad \text{رابطه(۱۳)}$$

معمولاً با افزایش تعداد داده‌ها توزیع آن‌ها به منحنی توزیع نرمال میل پیدا می‌کند. در توزیع نرمال، ٪۶۸٫۲ داده‌ها در فاصله کمتر از یک انحراف معیار نسبت به میانگین قرار دارند. در این تحقیق این فاصله برای سرعت تمام داده‌ها در نوع‌های حمل‌ونقل متفاوت محاسبه می‌شود، سپس درصد قرار گرفتن نقاط هر خط سیر در هریک از این فواصل به‌عنوان ویژگی به مدل معرفی می‌گردد.

۴-۳- مدل‌های طبقه‌بندی

پس از محاسبه ویژگی‌ها برای خطوط سیر، ۸۰ درصد داده‌ها به‌عنوان داده‌های آموزشی انتخاب‌شده و برای آموزش مدل‌ها مورداستفاده قرار می‌گیرند. ۲۰ درصد داده‌ها نیز به‌عنوان داده آزمایشی برای اندازه‌گیری دقت مدل‌ها در نظر گرفته خواهند شد. در این تحقیق از چهار روش طبقه‌بندی مختلف برای ایجاد مدل پیش‌بینی کننده نوع حمل‌ونقل داده‌های خط سیر استفاده شده است.

در این تحقیق از روش *SVM* با استفاده از رویکرد یک در مقابل همه[۱] و انتخاب کرنل‌های مختلف برای طبقه‌بندی چند کلاسه استفاده می‌شود. در رویکرد یک در مقابل همه، به ازای هر کلاس یک *SVM* استفاده می‌شود که وظیفه آن جدا کردن داده‌های یک کلاس از بقیه کلاس‌ها است. انتخاب مقدار مناسب برای پارامتر تنظیم نیز در این روش حائز اهمیت است [۳۱]. درخت تصمیم نیز در این تحقیق با اعمال ضریب جینی، بهره اطلاعاتی و نسبت بهره به‌عنوان معیارهای طبقه‌بندی، مورداستفاده قرار می‌گیرد. انتخاب عمق درخت و حداقل اعضای برگ‌ها پارامترهایی هستند که انتخاب صحیح آن‌ها می‌تواند در دقت مدل تأثیرگذار باشد [۳۱]. شبکه عصبی پرسپترون چندلایه روش دیگری است که در این تحقیق پیاده‌سازی می‌شود. انتخاب تعداد لایه‌ها و تعداد گره‌ها در هر لایه مهم‌ترین

[۱] *One vs all*

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]

پارامترهایی هستند که دقت این روش را تحت تأثیر قرار می‌دهند [۳۱]. در روش طبقه‌بندی کننده بیز ساده دقت با برآورد چگالی کرنل افزایش پیدا می‌کند. در این تحقیق از رویکرد برآورد حریصانه استفاده شده است که پارامترهای تعداد کرنل‌ها و حداقل عرض باند باید تنظیم شوند [۳۱].

۵- پیاده‌سازی

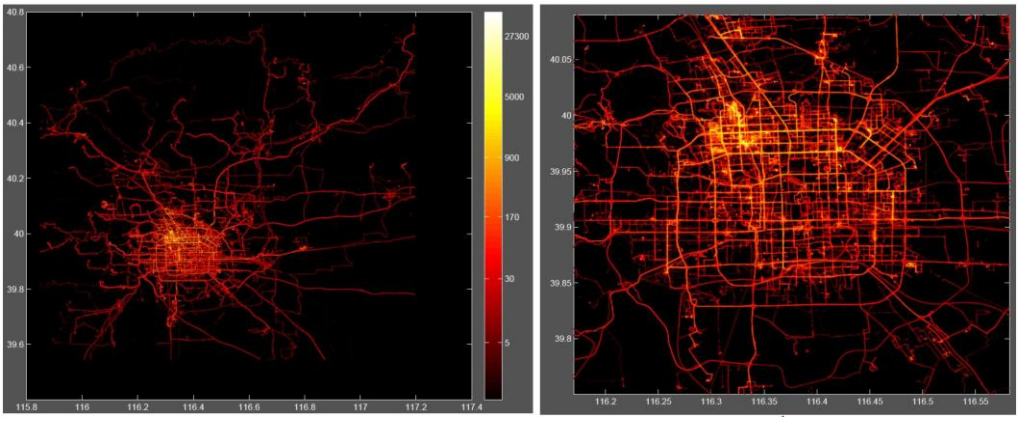
در این بخش به ارائه نتایج حاصل از پیاده‌سازی روش پیاده‌سازی تحقیق پرداخته خواهد شد. ابتدا داده‌های موردمطالعه معرفی شده و سپس توزیع آماری هر یک از ویژگی‌ها و دقت مدل‌های مختلف مورد بحث قرار می‌گیرند.

۵-۱- داده‌های موردمطالعه

در این تحقیق، از داده‌های پروژه ژئولایف استفاده شده است. این داده‌ها مربوط به شهر پکن است و توسط ۱۸۲ کاربر در مدت ۵ سال (آوریل ۲۰۰۷ تا اوت ۲۰۱۲) جمع‌آوری شده است. نرخ نمونه‌برداری این داده‌ها متفاوت است اما ۹۱٫۵ درصد از داده‌ها از چگالی مناسبی برخوردار هستند. بیشتر این داده با فاصله ۱ تا ۵ ثانیه و یا ۵ تا ۱۰ متر نمونه‌برداری شده‌اند. هر خط سیر در این مجموعه داده با استفاده از مجموعه‌ای از نقاط با فاصله زمانی مشخص، نمایش داده می‌شود و شامل طول و عرض جغرافیایی، ارتفاع و زمان نمونه‌برداری هستند [۳۵ و۳۶].

این داده‌ها شامل ۱۷۶۲۱ خط سیر هستند که درمجموع در طی ۵۰۱۷۶ ساعت و به طول ۱۲۹۲۹۵۱ کیلومتر جمع‌آوری شده است. داده‌ها شامل انواع گسترده‌ای از فعالیت‌های حرکتی کاربران هستند و تنها به سفرهای معمول بین خانه و محل کار محدود نمی‌شوند و فعالیت‌های سرگرمی و ورزش مانند خرید، غذا، هاکی، گردش و غیره را شامل می‌شوند. الگو کاوی حرکتی، شبکه‌های اجتماعی مکان‌مبنا، تشخیص فعالیت کاربر، حریم خصوصی و توصیه مکان ازجمله کاربردهایی است که می‌توان از این داده‌ها در آن بهره گرفت [۲۱ و۳۷].

اگرچه این داده‌ها از چندین شهر کشور چین و حتی شهرهای امریکا و اروپا جمع‌آوری شده است ولی بخش عمده این داده‌ها مربوط به شهر پکن هستند، شکل (۴) نقشه حرارتی از پراکندگی این داده‌ها را در شهر پکن نشان می‌دهد.



الف) شهر پکن ب) شهر پکن و حومه

شکل ۴: نقشه حرارتی داده‌های موردمطالعه [۳۵].

۵-۲- پیش‌پردازش و آماده‌سازی داده‌ها

از بین ۱۸۲ کاربر در پروژه ژئولایف، ۶۹ کاربر نوع حمل‌ونقل خود را نیز مشخص کرده‌اند. ازجمله نوع‌های موجود پیاده، خودرو، تاکسی، اتوبوس، مترو، قطار، دوچرخه و غیره (دویدن، هواپیما، موتور و قایق) هستند و این بخش از داده‌ها موردتوجه این تحقیق هستند.

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]

جدول(۱) خلاصه‌ای از آمار کلی نوع‌های حمل‌ونقل را ارائه می‌کند.

جدول ۱: آمار نوع حمل‌ونقل در داده‌های ژئولایف

نوع سفر	مسافت (km)	زمان (h)
پیاده	۱۰۱۲۳	۵۴۶۰
دوچرخه	۶۴۹۵	۲۴۱۰
اتوبوس	۲۰۲۸۱	۱۵۰۷
رانندگی (ماشین و تاکسی)	۳۲۸۶۶	۲۳۸۴
قطار	۳۶۲۵۳	۷۴۵
هواپیما	۲۴۷۸۹	۴۰
سایر (قایق، موتور و دویدن)	۹۴۹۳	۴۰۴
کل	**۱۴۰۳۰۴**	**۱۲۹۵۳**

در این تحقیق نوع‌های سفر کم‌اهمیت نادیده گرفته‌شده و پنج نوع سفر رانندگی (خودرو شخصی و تاکسی)، پیاده‌روی، دوچرخه، اتوبوس و قطار (قطار و مترو) مورد بررسی قرار گرفتند.

در اولین قدم هرگاه فاصله بیشتر از ۲۰ دقیقه در طی یک خط سیر وجود داشت، خط سیر به واحدهای کوچک‌تری به نام سفر تقسیم شد [۲۰]. به‌منظور آماده‌سازی داده‌ها برای انجام تحلیل‌های مکانی بعدی داده‌های فاقد نوع حمل‌ونقل حذف شدند زیرا برخی از کاربران نوع حمل‌ونقل را فقط برای قسمتی از خطوط سیر تعیین کرده‌اند. همچنین هر قسمت پیوسته از سفرها با نوع حمل‌ونقل خاص به‌عنوان واحد کاری انتخاب شدند و ازاین‌پس خط سیر نامیده می‌شوند.

سپس بر مبنای قوانین ترافیکی و محدودیت‌های فیزیکی، حداکثر سرعت و حداکثر شتاب برای هر نوع حمل‌ونقل تعیین شد [۱۰]. با اعمال کردن این شرایط بر روی نقاط خطوط سیر، نقطه‌های اشتباه از خط سیر حذف شده و پارامترهای نقاط همسایه به‌روزرسانی می‌شوند. شرایط ذکر شده در جدول(۲) شرح داده شده‌اند.

جدول ۲: حداکثر سرعت و شتاب مجاز برای نوع‌های حمل‌ونقل

نوع حمل‌ونقل	حداکثر سرعت (m/s)	حداکثر شتاب (m/s^2)
پیاده	۲/۵	۳
دوچرخه	۱۱	۳
خودرو و تاکسی	۳۳	۱۰
اتوبوس	۲۷	۲
قطار	۹۷	۳

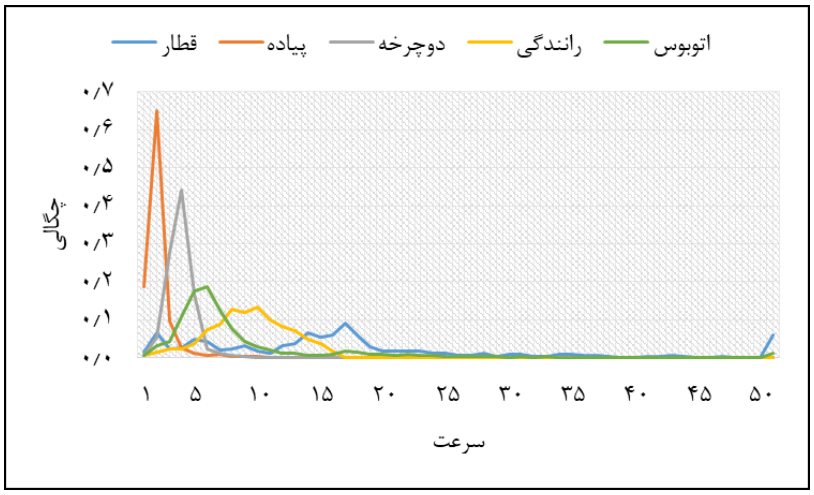
در آخرین مرحله پیش‌پردازش فیلتر ساویتزکی-گلی با تنظیم کردن پنجره با ابعاد ۷ نقطه و با تابع چندجمله‌ای درجه ۳ بر روی داده‌ها اعمال شد و پارامترهای اصلی نقاط به‌روزرسانی گردید. جدول (۳) تعداد خطوط سیر و میانگین سرعت مربوط به آن‌ها را پیش و پس از مرحله پالایش شرح می‌دهد.

جدول ۳: آمار خطوط سیر پس از پالایش

نوع حمل‌ونقل	پیش از پالایش		پس از پالایش	
	تعداد خطوط سیر	میانگین سرعت (*m/s*)	تعداد خطوط سیر	میانگین سرعت (*m/s*)
پیاده	۴۸۲۲	۲	۴۱۷۶	۱٫۷
دوچرخه	۱۶۰۱	۳٫۶	۱۵۷۰	۳٫۶
رانندگی	۱۴۰۳	۱۱٫۲	۱۳۹۳	۱۱٫۳
اتوبوس	۱۹۹۶	۶٫۷	۱۸۹۲	۶٫۶
قطار	۷۹۵	۲۱٫۸	۷۹۲	۲۱٫۹

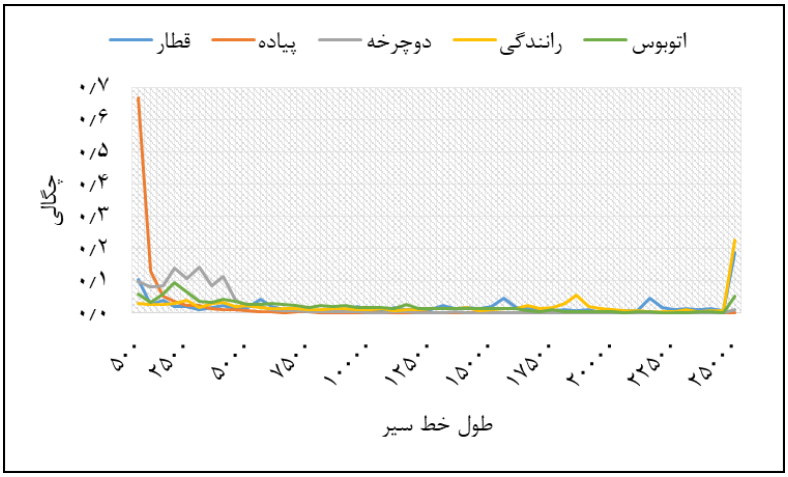
۵-۳- استخراج و تحلیل ویژگی‌ها

در این بخش ویژگی‌های خطوط سیر محاسبه و از طریق برخی آمار و نمودارها موردبررسی قرارگرفته و مقایسه می‌شوند. سرعت مهم‌ترین ویژگی خطوط سیر در تعیین نوع حمل‌ونقل است، شکل (۵) نمودار توزیع میانگین سرعت برای نوع‌های مختلف حمل‌ونقل را نشان می‌دهد. مطابق شکل نوع‌های حمل‌ونقل مختلف ازنظر ویژگی سرعت رفتار متفاوت و تعیین‌کننده‌ای دارند. فقط نوع حمل‌ونقل قطار در سرعت‌های بالای ۵۰ متر بر ثانیه فراوانی قابل‌توجهی دارد.



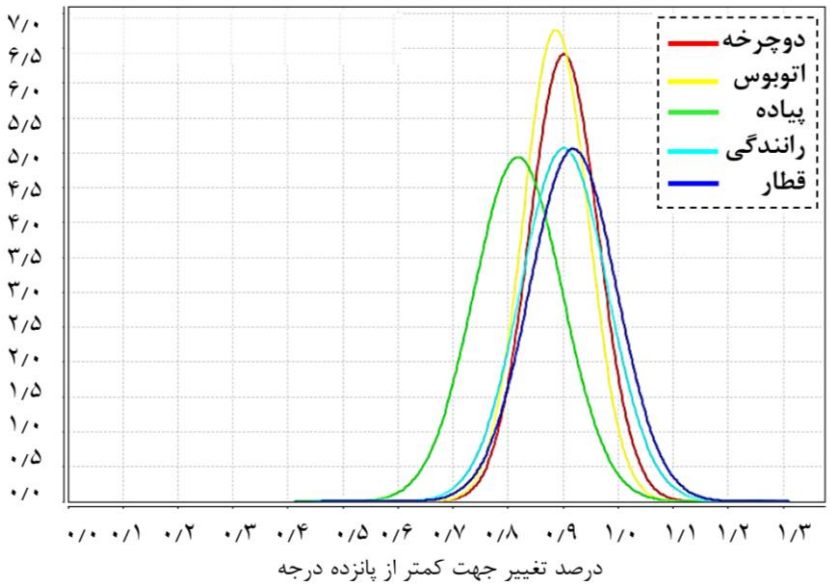
شکل ۵: نمودار توزیع میانگین سرعت برحسب *m/s*

شکل(۶) طول خطوط سیر را برای نوع‌های حمل‌ونقل مختلف نمایش می‌دهد. همان‌طور که انتظار می‌رود طول بخش عمده‌ای از خطوط سیر پیاده‌روی کمتر از دو کیلومتر است. همچنین طول کم خطوط سیر دوچرخه نسبت به سایر نوع‌ها مشهود است.



شکل ۶: نمودار توزیع طول خطوط سیر برحسب متر

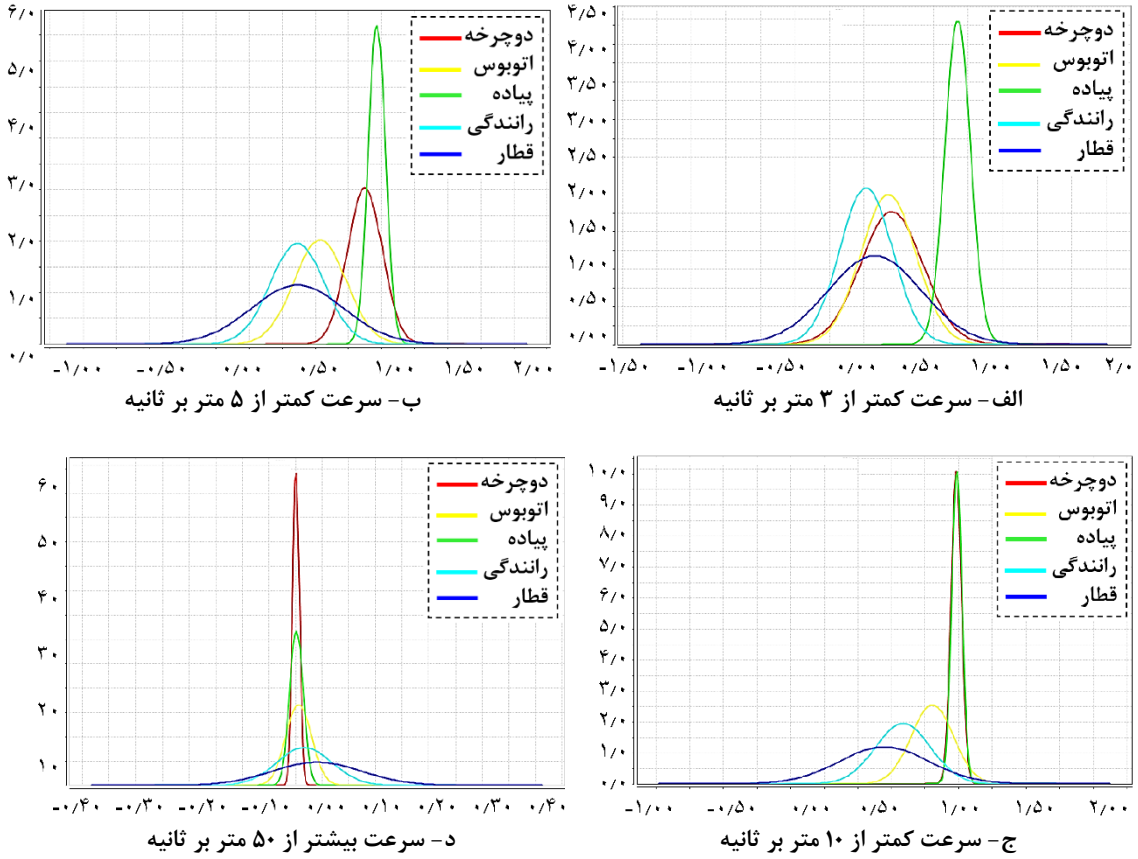
به‌منظور کشف تغییر جهت‌های محسوس از حد آستانه ۱۵ درجه استفاده شد [۲۰]. شکل (۷) توزیع آماری این ویژگی را برای نوع‌های حمل‌ونقل مختلف نشان می‌دهد. همچنین حد آستانه ۳ متر بر ثانیه برای محاسبه درصد نقاط توقف انتخاب شد [۲۱]. این نمودار نشان می‌دهد که نوع پیاده‌روی دارای بیشترین تغییر جهت بیشتر از ۱۵ درجه است.



شکل ۷: نمودار توزیع چگالی تغییر جهت

شکل (۸) توزیع داده‌های خط سیر در نوع‌های مختلف حمل‌ونقل را برای چهار بازه سرعت نمایش می‌دهد که به‌عنوان ویژگی‌های جدید در این تحقیق مورداستفاده قرار می‌گیرند. رفتار متفاوت هر نوع حمل‌ونقل در بازه‌های سرعتی مختلف در شکل (۸) مشهود است. مطابق شکل (۸-ج) نوع‌های پیاده‌روی و دوچرخه دارای

توزیع یکسانی برای سرعت‌های کمتر از ۱۰ متر بر ثانیه هستند، اما شکل(۸-ب) نشان می‌دهد که دو نوع پیاده‌روی و دوچرخه در سرعت‌های کمتر از ۵ متر بر ثانیه دارای توزیع متفاوتی هستند و در نوع پیاده‌روی درصد بالایی از نقاط دارای سرعت کمتر از ۵ ثانیه هستند. در سایر شکل‌ها نیز تفاوت رفتار نمودار برای نوع‌های مختلف حمل‌ونقل مشهود است.



الف- سرعت کمتر از ۳ متر بر ثانیه

ب- سرعت کمتر از ۵ متر بر ثانیه

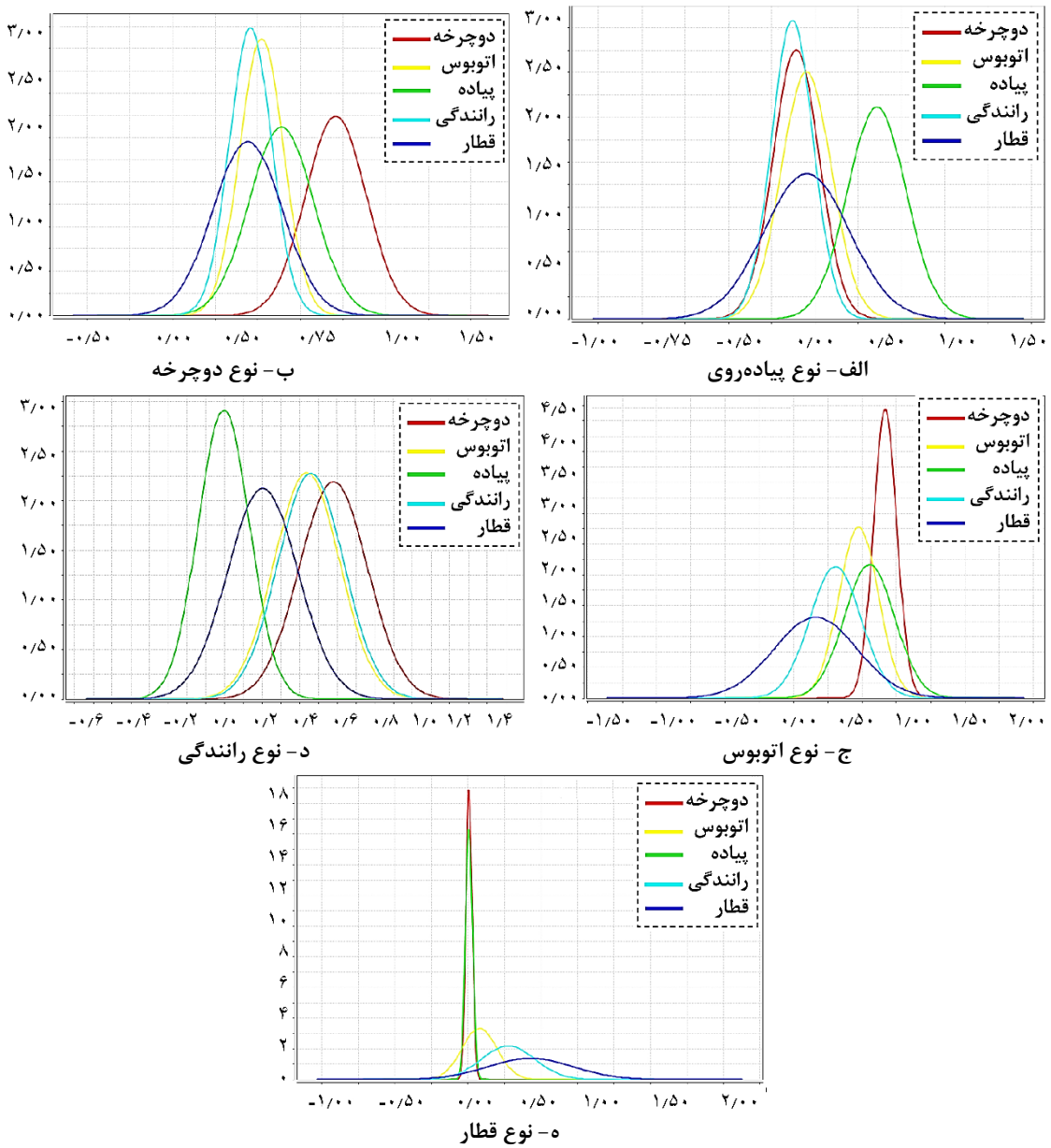
ج- سرعت کمتر از ۱۰ متر بر ثانیه

د- سرعت بیشتر از ۵۰ متر بر ثانیه

شکل ۸: نمودار توزیع چگالی درصد نقاط خطوط سیر در بازه‌های سرعتی مختلف

به‌عنوان یک ویژگی جدید، پس از محاسبه درصد حضور نقاط هر خط سیر در فاصله یک انحراف معیار از میانگین کل سرعت نوع‌های حمل‌ونقل، توزیع این ویژگی برای همه حالات در شکل (۹) نمایش داده‌شده است. نمودار زردرنگ شکل (۹-الف) نحوه توزیع سرعت خطوط سیر اتوبوس نسبت به میانگین سرعت پیاده‌روی را نشان می‌دهد. به‌طورکلی رفتار متفاوت داده‌های هر نوع حمل‌ونقل در ویژگی‌های محاسبه‌شده در این تحقیق، در شکل‌های (۹) مشهود است؛ بنابراین می‌توانند در بهبود عملکرد مدل پیاده‌سازی تحقیق تأثیرگذار باشند. به‌عنوان نمونه شکل (۹- ه) نشان می‌دهد که سرعت نقاط نوع‌های دوچرخه و پیاده‌روی به‌ندرت در فاصله یک انحراف معیار از میانگین سرعت نوع قطار قرار می‌گیرند. درحالی‌که مطابق انتظار نوع قطار در این نمودار دارای توزیع نرمال است و نوع‌های رانندگی و اتوبوس نیز به ترتیب نزدیک‌ترین توزیع را به توزیع نرمال دارند.

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]



شکل ۹: نمودار توزیع چگالی درصد نقاط خطوط سیر در بازه یک انحراف معیار از میانگین کل سرعت در نوع‌های مختلف

۴-۵- پیاده‌سازی مدل‌های طبقه‌بندی

در این مرحله چهار روش طبقه‌بندی *SVM*، *MLP*، درخت تصمیم و بیز ساده بر روی داده‌ها پیاده‌سازی شد. داده‌ها به دو بخش داده آموزشی و آزمایشی تقسیم شدند. به‌منظور مقایسه نتایج ۲۰ درصد ثابت از داده‌ها به‌عنوان داده آزمایشی در نظر گرفته شدند. داده‌های آزمایشی با توجه به حجم بالای مجموعه داده ژئولایف، دارای توزیع مکانی مناسبی بودند. شکل (۱۰) توزیع مکانی داده‌های آزمایشی را نشان می‌دهد. تمام روش‌ها به ازای پارامترهای ورودی مختلف پیاده‌سازی شد و بهینه‌ترین پارامترها به کمک جستجوی شبکه‌ای (*Grid Search*) تنظیم شد و نتایج زیر حاصل گردید. در جستجوی شبکه‌ای بر اساس بهترین مقادیر ارائه شده در جستجوی تصادفی، شبکه‌ای از مقادیر مختلف برای پارامترهای مدل‌ها ایجاد می‌گردد و جستجو آغاز می‌شود [۳۸].

شکل ۱۰: نقشه توزیع مکانی داده‌های آزمایشی

در پیاده‌سازی روش *SVM* چهار کرنل خطی، چندجمله‌ای، گوسین و تابع پایه شعاعی ((*RBF*) *Radial Basis Function*) به ازای پارامتر تنظیم در $C\in\{۰/۰۰۱, ۰/۰۱, ۰/۰۲, ۰/۰۳, ۰/۰۵, ۰/۱\}$ مورد جستجو قرار گرفتند. بهترین عملکرد توسط کرنل خطی به ازای مقدار پارامتر تنظیم برابر ۰/۰۲ شناسایی شد. کرنل خطی یا ضرب داخلی به شکل $K(x,y) = x.y$ تعریف می‌شود. جدول (۴) دقت کلی برآورد نوع حمل‌ونقل را به ازای پارامترهای متفاوت روش *SVM* نشان می‌دهد.

جدول ۴: تأثیر پارامترهای روش *SVM* بر دقت

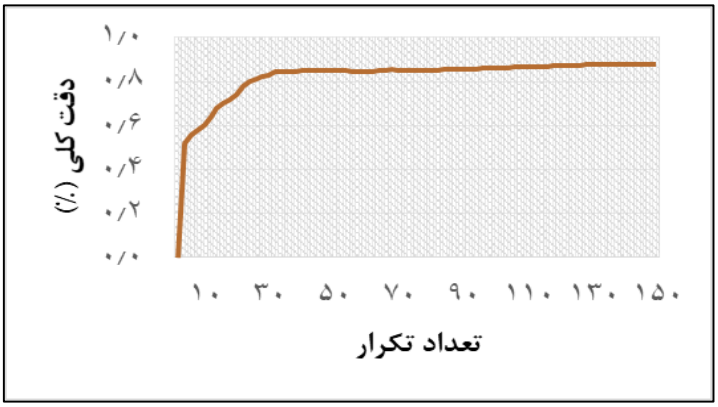
تابع کرنل	پارامتر تنظیم					
	۰/۰۰۱	۰/۰۱	۰/۰۲	۰/۰۳	۰/۰۵	۰/۱
کرنل خطی	۰/۸۳۱	۰/۸۵۸	**۰/۸۵۹**	۰/۸۵۷	۰/۸۵۱	۰/۸۴۵
کرنل چندجمله‌ای	۰/۸۳۹	۰/۸۵۵	۰/۸۵۵	۰/۸۵۵	۰/۸۴۷	۰/۸۴۱
تابع گوسین	۰/۸۰۱	۰/۸۲۲	۰/۸۲۵	۰/۸۲۵	۰/۸۱۹	۰/۸۰۶
تابع پایه شعاعی	۰/۸۳۷	۰/۸۵۰	۰/۸۵۲	۰/۸۵۰	۰/۸۴۲	۰/۸۴۰

برای روش *MLP* یک لایه میانی با ۲۵ گره انتخاب شد و نرخ یادگیری برابر ۰/۳ تنظیم شد. این روش به ازای یک تا پنج لایه میانی و تعداد گره‌ها $n =$ ۱۰, ۱۱, ... ,۴۰ پیاده‌سازی شد.جدول(۵) دقت کلی برآورد نوع حمل‌ونقل را به ازای معماری متفاوت در روش *MLP* نشان می‌دهد. شکل(۱۱) نمودار همگرایی روش *MLP* را نشان می‌دهد. درخت تصمیم با معیارهای بهره اطلاعاتی[۱]، نسبت بهره[۲] و شاخص جینی پیاده‌سازی شد. شاخص جینی بهترین عملکرد را از خود نشان داد. هم‌چنین حداکثر عمق درخت برابر ۱۰ و حداقل تعداد اعضای برگ دو انتخاب شد. جدول(۶) دقت کلی برآورد نوع حمل‌ونقل را به ازای پارامترهای متفاوت درخت تصمیم نشان می‌دهد.

در پیاده‌سازی روش بیز ساده با رویکرد برآورد حریصانه از چهار کرنل با حداقل باند ۰/۰۸ استفاده شد. در این روش تعداد کرنل‌ها بین ۱ تا ۱۰ و حداقل باند در $B\in\{۰/۰۱, ۰/۰۵, ۰/۰۸, ۰/۱, ۰/۲, ۰/۵\}$ مورد جستجو قرار گرفتند. جدول (۷) دقت کلی برآورد نوع حمل‌ونقل را به ازای پارامترهای متفاوت روش بیز ساده نشان می‌دهد.

[۱] *Information Gain*

[۲] *Gain Ratio*



شکل ۱۱: نمودار همگرایی روش *MLP*

جدول ۵: تأثیر پارامترهای روش *MLP* بر دقت

تعداد لایه‌ها	تعداد گره‌ها					
	۱۰	۱۵	۲۰	۲۵	۳۰	۳۵
۱	۰٫۸۴۱	۰٫۸۵۲	۰٫۸۷۰	**۰٫۸۷۸**	۰٫۸۷۸	۰٫۸۷۸
۲	۰٫۸۴۴	۰٫۸۵۵	۰٫۸۷۰	۰٫۸۷۷	۰٫۸۷۷	۰٫۸۷۷
۳	۰٫۸۴۵	۰٫۸۵۸	۰٫۸۷۰	۰٫۸۷۳	۰٫۸۷۳	۰٫۸۷۳

جدول ۶: تأثیر پارامترهای درخت تصمیم بر دقت

معیار	حداکثر عمق درخت					
	۴	۶	۸	۱۰	۱۲	۱۴
بهره اطلاعاتی	۰٫۷۹۵	۰٫۸۲۹	۰٫۸۴۶	۰٫۸۴۲	۰٫۸۴۰	۰٫۸۳۷
نسبت بهره	۰٫۶۹۸	۰٫۷۴۳	۰٫۷۴۳	۰٫۷۴۸	۰٫۷۵۱	۰٫۷۵۶
شاخص جینی	۰٫۷۸۷	۰٫۸۳۸	۰٫۸۴۹	**۰٫۸۵۳**	۰٫۸۵۱	۰٫۸۵۰

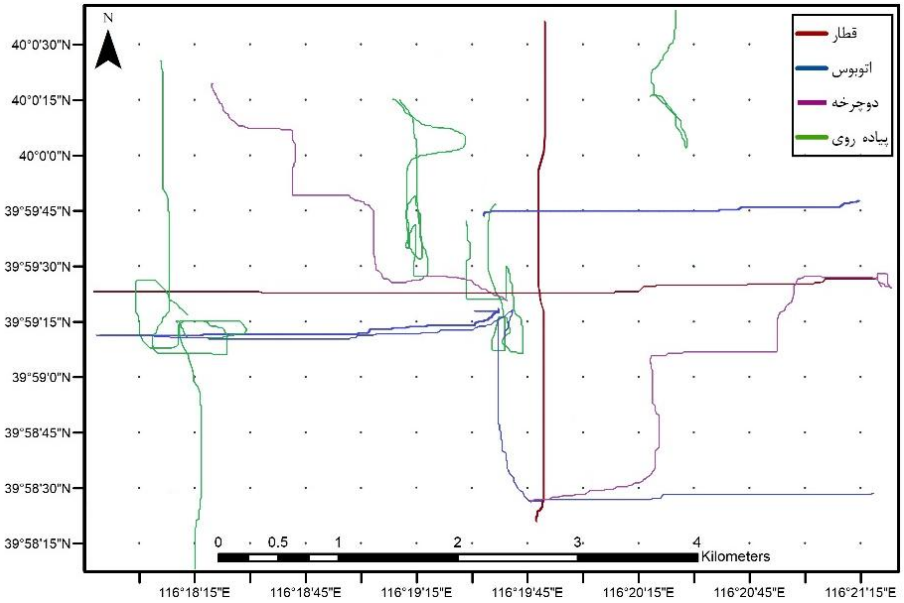
جدول ۷: تأثیر پارامترهای بیز ساده بر دقت

تعداد کرنل‌ها	حداقل باند					
	۰٫۰۱	۰٫۰۵	۰٫۰۸	۰٫۱	۰٫۲	۰٫۵
۱	۰٫۷۳۳	۰٫۷۴۰	۰٫۷۳۵	۰٫۷۳۴	۰٫۷۳۳	۰٫۷۳۰
۳	۰٫۷۶۱	۰٫۷۶۴	۰٫۷۶۵	۰٫۷۶۵	۰٫۷۶۱	۰٫۷۵۸
۴	۰٫۷۶۹	۰٫۷۷۷	**۰٫۷۸۲**	۰٫۷۸۱	۰٫۷۷۱	۰٫۷۶۸
۵	۰٫۷۶۱	۰٫۷۷۴	۰٫۷۸۲	۰٫۷۸۰	۰٫۷۷۰	۰٫۷۶۸
۸	۰٫۷۵۱	۰٫۷۶۴	۰٫۷۷۳	۰٫۷۶۹	۰٫۷۶۳	۰٫۷۶۰

۶- نتایج و بحث

در این بخش نتایج حاصل از پیاده‌سازی تحقیق بیان می‌شود و به مقایسه دقت عملکرد روش‌های مختلف طبقه‌بندی در تعیین نوع حمل‌ونقل پرداخته خواهد شد. هم‌چنین دقت طبقه‌بندی نوع‌های حمل‌ونقل در این تحقیق با دقت مشابه‌ترین تحقیقات گذشته مقایسه می‌شود.

شکل(۱۲) نتایج حاصل از روش *MLP* بر روی بخش بسیار کوچکی از داده‌های موردمطالعه را نشان می‌دهد. در این شکل تمام نوع‌های حمل‌ونقل به‌درستی شناسایی شده‌اند.



شکل ۱۲: بخشی از نتایج تحقیق به روش *MLP*

۶-۱- مقایسه نتایج روش‌های طبقه‌بندی

به‌منظور مقایسه نتایج حاصل از پیاده‌سازی روش‌های مختلف طبقه‌بندی دو معیار دقت کلی[۱] و شاخص کاپا[۲] محاسبه گردید. جدول‌های (۸) الی (۱۱) ماتریس ابهام[۳] را برای چهار روش طبقه‌بندی نشان می‌دهند.

در میان روش‌های پیاده‌سازی این تحقیق روش *MLP* در هر دو معیار دقت کلی و شاخص کاپا توانست دقت بهتری نسبت به سایر روش‌ها از خود نشان دهد. همچنین روش *SVM* عملکرد بهتری نسبت به روش‌های بیز ساده و درخت تصمیم داشت. جدول(۱۲) نیز مقدار معیارهای دقت را برای روش‌های مختلف شرح می‌دهد.

[۱] *Overall accuracy*

[۲] *Kappa index*

[۳] *Confusion matrix*

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]

جدول ۸: ماتریس ابهام روش درخت تصمیم

پیش‌بینی	واقعی: قطار	رانندگی	پیاده	اتوبوس	دوچرخه
قطار	۱۰۷	۱۷	۴	۶	۰
رانندگی	۱۱	۱۹۴	۱۰	۳۷	۳
پیاده	۳۲	۱۰	۸۱۸	۲۹	۳۳
اتوبوس	۵	۵۴	۱	۳۰۱	۲۲
دوچرخه	۳	۴	۲	۵	۲۵۶

جدول ۹: ماتریس ابهام روش *MLP*

پیش‌بینی	واقعی: قطار	رانندگی	پیاده	اتوبوس	دوچرخه
قطار	۱۲۵	۶	۰	۹	۰
رانندگی	۲	۲۱۶	۳	۵۱	۱
پیاده	۲۵	۷	۸۲۴	۲۲	۱۸
اتوبوس	۴	۴۶	۳	۲۸۵	۲۱
دوچرخه	۲	۴	۵	۱۱	۲۷۴

جدول ۱۰: ماتریس ابهام روش *SVM*

پیش‌بینی	واقعی: قطار	رانندگی	پیاده	اتوبوس	دوچرخه
قطار	۱۰۶	۱۰	۱	۶	۰
رانندگی	۱۶	۱۸۸	۲	۳۱	۱
پیاده	۳۱	۱۲	۸۲۴	۲۹	۳۱
اتوبوس	۳	۶۳	۳	۳۰۶	۱۹
دوچرخه	۲	۶	۵	۶	۲۶۳

جدول ۱۱: ماتریس ابهام روش بیز ساده

پیش‌بینی	واقعی: قطار	رانندگی	پیاده	اتوبوس	دوچرخه
قطار	۷۳	۱۴	۱۱	۱۳	۱
رانندگی	۴۲	۲۰۲	۷	۶۶	۱
پیاده	۲۹	۱۲	۷۸۶	۲۸	۳۱
اتوبوس	۱۱	۴۱	۱۴	۲۱۸	۲۴
دوچرخه	۳	۱۰	۱۷	۵۳	۲۵۷

جدول ۱۲: مقایسه دقت روش‌های طبقه‌بندی

مدل	دقت کلی	شاخص کاپا
درخت تصمیم	۰٫۸۵۳	۰٫۷۹۵
MLP	۰٫۸۷۸	۰٫۸۵۹
SVM	۰٫۸۵۹	۰٫۸۰۳
بیز ساده	۰٫۷۸۲	۰٫۶۹۸

۶-۲- بررسی تأثیر پیش‌پردازش داده‌ها و ویژگی جدید بر دقت نتایج

به‌منظور بررسی میزان تأثیرگذاری مراحل پیش‌پردازش بر دقت نتایج تحقیق، چهار روش مورداستفاده بر روی داده‌های خام نیز پیاده‌سازی شد. مطابق جدول (۱۳) معیار دقت کلی برای مقایسه نتایج روش‌ها با پیاده‌سازی بر روی داده‌های خام و داده‌های پالایش‌شده محاسبه شده است. مقایسه دقت کلی نتایج نشان‌دهنده تأثیر قابل‌توجه پیش‌پردازش داده‌ها بر روی دقت نتایج است، به‌طوری‌که در همه روش‌ها دقت نتایج پس از انجام مراحل پیش‌پردازش بیش از ۴ درصد بهبود یافته است. همچنین در این تحقیق از درصد حضور نقاط هر خط سیر در فاصله یک انحراف معیار از میانگین کل سرعت نوع‌های حمل‌ونقل به‌عنوان یک ویژگی جدید استفاده شده است. معیار دقت کلی برای چهار روش مورداستفاده در تحقیق با حضور و عدم حضور ویژگی جدید نیز در جدول(۱۴) مورد مقایسه قرارگرفته است. نتایج نشان می‌دهد

افزودن این ویژگی به مدل‌های طبقه‌بندی موجب بهبود دقت کلی نتایج شده است.

جدول ۱۳: مقایسه دقت کلی نتایج پس از پیش‌پردازش داده‌ها

مدل	بدون پیش‌پردازش	با پیش‌پردازش
درخت تصمیم	۰٫۸۰۸	۰٫۸۵۳
MLP	۰٫۸۳۷	۰٫۸۷۸
SVM	۰٫۸۱۲	۰٫۸۵۹
بیز ساده	۰٫۷۴۲	۰٫۷۸۲

جدول ۱۴: مقایسه دقت کلی نتایج با افزودن ویژگی جدید

مدل	بدون حضور ویژگی جدید	با حضور ویژگی جدید
درخت تصمیم	۰٫۸۴۲	۰٫۸۵۳
MLP	۰٫۸۶۹	۰٫۸۷۸
SVM	۰٫۸۵۱	۰٫۸۵۹
بیز ساده	۰٫۷۷۲	۰٫۷۸۲

۶-۳- مقایسه با تحقیقات پیشین

به‌منظور سنجیدن دقت روش‌های پیاده‌سازی این تحقیق در پیش‌بینی نوع حمل‌ونقل خطوط سیر، دقت این تحقیق با مطالعاتی که پیش‌ازاین بر روی داده‌های پروژه ژئولایف انجام‌شده بود، مقایسه شد. مطابق با آنچه در جدول(۱۵) آورده شده است، نتایج این تحقیق با استفاده از مدل *MLP* دقت بهتری نسبت به مطالعات مشابه از خود نشان داده است. اگرچه مدل‌های دیگر این تحقیق نیز توانسته‌اند دقت بهتری نسبت به اکثر مطالعات قبلی از خود نشان دهند. این بهبود دقت می‌تواند ناشی از پیش‌پردازش مناسب داده‌ها، استخراج ویژگی‌های جدید و تنظیم دقیق پارامترهای ورودی مدل‌ها باشد. در این تحقیق برخلاف تحقیق ژنگ و همکاران [۲۰] از مسئله قطعه‌بندی خودکار خطوط سیر توسط مدل چشم‌پوشی شده است و خطوط سیر به شکل قطعه‌بندی شده به مدل معرفی شده است و همین مسئله بخشی از این اختلاف دقت را به وجود آورده است.

جدول ۱۵: مقایسه نتایج تحقیق با تحقیقات مشابه

تحقیق	دقت کلی	تحقیق	دقت کلی
هوانگ و همکاران (۲۰۲۰) [۲۷]	۰٫۸۱	ونگ و همکاران (۲۰۱۷) [۲۲]	۰٫۷۴
جیمز و همکاران (۲۰۲۰) [۲۶]	۰٫۹۳	اندو و همکاران (۲۰۱۶) [۱۳]	۰٫۶۹
نواز و همکاران (۲۰۲۰) [۲۵]	۰٫۸۴	ژنگ و همکاران (۲۰۱۰) [۱۲, ۲۱]	۰٫۷۶
دبیری و همکاران (۲۰۱۸) [۱۰]	۰٫۸۵	بهترین مدل این تحقیق (*MLP*)	**۰٫۸۸**

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]

۷- نتیجه‌گیری

در این تحقیق، برای ایجاد مدل شناسایی نوع حمل‌ونقل کاربران بر مبنای داده‌های خطوط سیر روش‌های مختلف طبقه‌بندی به کار گرفته شد. مطالعه این تحقیق بر روی داده‌های خام *GPS* صورت گرفت و هیچ داده خارجی بر آن اضافه نشد. در مرحله پیش‌پردازش داده‌های غیر معتبر حذف شدند و به‌منظور کاهش خطای اتفاقی یک کرنل نرم‌کننده بر روی خطوط سیر اعمال شد. سپس یک مجموعه از ویژگی‌های خطوط سیر شامل سرعت، طول، شتاب و غیره استخراج شد و تعدادی ویژگی آماری نیز بر آن‌ها افزوده شد. سپس مدل‌های طبقه‌بندی با چهار روش درخت تصمیم، بیز ساده، *SVM* و *MLP* ایجاد شد و عملکرد آن‌ها موردبررسی و مقایسه قرار گرفت.

نتایج حاصل از این تحقیق نشان داد که شبکه عصبی پرسپترون چندلایه در بین روش‌های مختلف بهترین کارایی را در شناسایی نوع حمل‌ونقل کاربران دارد. ارزیابی نتایج نشان داد که مدل‌های پیاده‌سازی این تحقیق نسبت به تحقیقات پیشین عملکرد بهتری از خود نشان می‌دهد. عملکرد مناسب مدل‌های پیاده‌سازی این تحقیق می‌تواند ناشی از پیش‌پردازش دقیق و جامع داده‌ها و استخراج ویژگی‌ها مناسب باشد. همچنین جستجوی گسترده صورت گرفته برای تنظیم پارامترهای ورودی مدل‌ها منجر به نتایج بهتر شد.

خطوط سیر در این تحقیق به شکل بخش‌بندی شده بر مبنای نوع حمل‌ونقل به مدل‌های طبقه‌بندی معرفی شدند. درحالی‌که داده‌های خام ممکن است به شکل دنباله پیوسته‌ای از نوع‌های مختلف حمل‌ونقل ایجاد شوند؛ بنابراین بخش‌بندی خطوط سیر بر مبنای تغییر نوع حمل‌ونقل می‌تواند به‌عنوان یک موضوع تحقیقاتی در آینده موردتوجه قرار گیرد. همچنین برای تحقیقات آینده می‌توان از منابع داده دیگری مانند داده‌های شبکه‌های اجتماعی مکان‌مبنا استفاده کرد و نتایج حاصل را مقایسه نمود. همچنین تعمیم روش پیاده‌سازی این تحقیق به داده‌هایی که شامل برچسب نوع حمل‌ونقل نیستند و به‌کارگیری روش‌های نظارت نشده می‌تواند یک زمینه تحقیقاتی در آینده باشد. استفاده از بعد زمانی داده‌ها و استخراج ویژگی‌های دیگر و مطالعه دقیق آن‌ها نیز می‌تواند در تحقیقات آینده موجب بهبود دقت شوند.

مراجع

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 5, pp. 1-55, 2014.

[2] Y. Zheng, "Trajectory data mining: an overview," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 6, pp. 1-41, 2015.

[3] E. Camossi, P. Villa, and L. Mazzola, "Semantic-based anomalous pattern discovery in moving object trajectories," arXiv preprint arXiv:1305.1946, 2013.

[4] O. Wolfson, B. Xu, S. Chamberlain, and L. Jiang, "Moving objects databases: Issues and solutions," in Proceedings. Tenth International Conference on Scientific and Statistical Database Management (Cat. No. 98TB100243), 1998, pp. 111-122.

[5] A. Elragal and N. El-Gendy, "Trajectory data mining: integrating semantics," Journal of Enterprise Information Management, vol. 26, pp. 516-535, 2013.

[6] S. Chen, C. S. Jensen, and D. Lin, "A benchmark for evaluating moving object indexes," Proceedings of the VLDB Endowment, vol. 1, pp. 1574-1585, 2008.

[7] R. H. Güting, T. Behr, and C. Düntgen, "SECONDO: A Platform for Moving Objects Database Research and for Publishing and Integrating Research Implementations," IEEE Data Eng. Bull., vol. 33, pp. 56-63, 2010.

[8] N. Pelekis, Y. Theodoridis, S. Vosinakis, and T. Panayiotopoulos, "Hermes–a

framework for location-based data management," in International Conference on Extending Database Technology, 2006, pp. 1130-1134.

[9] O. Wolfson, P. Sistla, B. Xu, J. Zhou, and S. Chamberlain, "DOMINO: Databases for moving objects tracking," ACM SIGMOD Record, vol. 28, pp. 547-549, 1999.

[10] S. Dabiri and K. Heaslip, "Inferring transportation modes from GPS trajectories using a convolutional neural network," Transportation research part C: emerging technologies, vol. 86, pp. 360-371, 2018.

[11] N. Eluru, V. Chakour, and A. M. El-Geneidy, "Travel mode choice and transit route choice behavior in Montreal: insights from McGill University members commute patterns," Public Transport, vol. 4, pp. 129-149, 2012.

[12] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma, "Understanding transportation modes based on GPS data for web applications," ACM Transactions on the Web (TWEB), vol. 4, pp. 1-36, 2010.

[13] Y. Endo, H. Toda, K. Nishida, and A. Kawanobe, "Deep feature extraction from trajectories for transportation mode estimation," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2016, pp. 54-66.

[14] X. Kong, M. Li, K. Ma, K. Tian, M. Wang, Z. Ning, et al., "Big trajectory data: A survey of applications and services," IEEE Access, vol. 6, pp. 58295-58306, 2018.

[15] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 1027-1036.

[16] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring social ties between users with human location history," Journal of Ambient Intelligence and Humanized Computing, vol. 5, pp. 3-19, 2014.

[17] R. C. Shah, C.-y. Wan, H. Lu, and L. Nachman, "Classifying the mode of transportation on mobile phones using GIS information," in Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing, 2014, pp. 225-229.

[18] D. J. Patterson, L. Liao, D. Fox, and H. Kautz, "Inferring high-level behavior from low-level sensors," in International Conference on Ubiquitous Computing, 2003, pp. 73-89.

[19] Y. Zheng, Y. Chen, X. Xie, and W.-Y. Ma, "GeoLife2. 0: a location-based social networking service," in 2009 tenth international conference on mobile data management: systems, services and middleware, 2009, pp. 357-358.

[20] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw gps data for geographic applications on the web," in Proceedings of the 17th international conference on World Wide Web, 2008, pp. 247-256.

[21] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on GPS data," in Proceedings of the 10th international conference on Ubiquitous computing, 2008, pp. 312-321.

[22] H. Wang, G. Liu, J. Duan, and L. Zhang, "Detecting transportation modes using deep neural network," IEICE TRANSACTIONS on Information and Systems, vol. 100, pp. 1132-1135, 2017.

[23] Z. Xiao, Y. Wang, K. Fu, and F. Wu, "Identifying different transportation modes from trajectory data using tree-based ensemble classifiers," ISPRS International Journal of Geo-Information, vol. 6, p. 57, 2017.

[24] S. Dabiri, C.-T. Lu, K. Heaslip, and C. K. Reddy, "Semi-supervised deep learning approach for transportation mode identification using GPS trajectory data," IEEE Transactions on Knowledge and Data

Engineering, vol. 32, pp. 1010-1023, 2019.

[25] A. Nawaz, H. Zhiqiu, W. Senzhang, Y. Hussain, I. Khan, and Z. Khan, "Convolutional LSTM based transportation mode learning from raw GPS trajectories," IET Intelligent Transport Systems, vol. 14, pp. 570-577, 2020.

[26] J. James, "Travel Mode Identification With GPS Trajectories Using Wavelet Transform and Deep Learning," IEEE Transactions on Intelligent Transportation Systems, 2020.

[27] Z. Huang, P. Wang, and Y. Liu, "Statistical characteristics and transportation mode identification of individual trajectories," International Journal of Modern Physics B, vol. 34, p. 2050092, 2020.

[28] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144-152.

[29] Y. Ma and G. Guo, Support vector machines applications vol. 649: Springer, 2014.

[30] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE transactions on systems, man, and cybernetics, vol. 21, pp. 660-674, 1991.

[31] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques: Elsevier, 2011.

[32] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in Neural networks for perception, ed: Elsevier, 1992, pp. 65-93.

[33] T. Vincenty, "Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations," Survey review, vol. 23, pp. 88-93, 1975.

[34] R. W. Schafer, "What is a Savitzky-Golay filter?[lecture notes]," IEEE Signal processing magazine, vol. 28, pp. 111-117, 2011.

[35] Y. Zheng, H. Fu, X. Xie, W.-Y. Ma, and Q. Li, "Geolife GPS trajectory dataset-user guide," Microsoft Research, 2011.

[36] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in Proceedings of the 18th international conference on World wide web, 2009, pp. 791-800.

[37] Y. Zheng, X. Xie, and W.-Y. Ma, "GeoLife: A collaborative social networking service among user, location and trajectory," IEEE Data Eng. Bull., vol. 33, pp. 32-39, 2010.

[38] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," Telkomnika, vol. 14, p. 1502, 2016.

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]

K. N. TOOSI UNIVERSITY OF TECHNOLOGY
FACULTY OF GEODESY AND GEOMATICS ENGINEERING

Journal of Geospatial Information Technology
Vol.10, No.3, Autumn 2022

Research Paper

Analyzing the performance of different machine learning methods in determining the transportation mode using trajectory data

Morteza Tayebi [1], Parham Pahlavani [2*]

1- GIS PhD student, School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran
2- Associate Professor, School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran

Abstract

With the widespread advent of the smart phones equipping with Global Positioning System (GPS), a huge volume of users' trajectory data was generated. To facilitate the urban management and present appropriate services to users, studying these data was raised as a widespread research field and has been developing since then. In this research, the transportation mode of users' trajectories was identified based on their raw GPS data. These data are often associated with errors, it was attempted to minimize them by applying a comprehensive pre-processing procedure in this research. Accordingly, various features were extracted to identify the transportation modes including walk, bike, train, bus, and driving. In this regard, four classification methods including decision tree, multilayer perceptron neural network, Naïve Bayes, and support vector machine were used to build a predictive model. In order to improve the performance of the implementation methods, the percentage of the points of each trajectory on the distance of one standard deviation from the total speed average of transportation modes has been used as a new feature. The above-mentioned four models were implemented with different regularization parameters and their values were set to the optimal values by applying a comprehensive grid search. Then, Kappa and the overall accuracy indices were employed to evaluate different methods. The results of this study show that the multilayer perceptron neural network with overall accuracy of 0.88 has the best results compared to the other models.

Key words: *Trajectory data, Determining the transportation mode, Classification, Machine learning.*

Correspondence Address*: GIS Group, School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran.*
Tel*: +98 21 61114524.*
Email: *pahlavani@ut.ac.ir*

[DOR: 20.1001.1.20089635.1401.10.3.4.2]
[DOI: 10.52547/jgit.10.3.71]