****

**استخراج الگو های سفر از داده های کارت های هوشمند با خوشه بندی عمیق**

**Travel Pattern Mining from Smart Card Data with Deep Clustering**

دانشجو: شارون سارونیان

شماره دانشجویی: 98112095

گرایش: داده کاوی

استاد راهنما: دکتر مهدی قطعی

1. **مقدمه**

امروزه داده ها نقش مهمی در استخراج دانش به دست آورده­اند. در عصر حاضر، داده ها به وفور از منابع مختلفی جمع آوری می شوند و با آنالیز آن­ها می­توان الگوهای سفر انسان­ها و اطلاعات ارزشمندی در خصوص موارد مرتبط با شهر هوشمند استخراج کرد یا دید بهتری نسبت به آینده پیدا کرد [1]. یکی از مهمترین عناصر شهرهای هوشمند حمل­ونقل و در این بخش حمل­ونقل عمومی دارای مهمترین تاثیرها است [2]. برای تشویق مردم به استفاده­ی بیشتر از حمل و نقل عمومی، لازم است الگوهای سفری انسان­ها آنالیز و زیرساختهای حمل­ونقل عمومی مناسب آنها تعریف نمود [3]. بدین منظور، یکی از مهم­ترین منابع اطلاعات حمل­و­نقل عمومی کارت­های هوشمند می باشند که علاوه بر وسیله­ای برای پرداخت خودکار هزینه سفر، شامل داده های رفت و آمد کاربران مانند تراکنش ها، زمان و مکان سوار و پیاده شدن، هزینه ی سفر، و غیره هستند [4].

داده های کارت های هوشمند بسیار حجیم می­باشند و آنالیز الگوهای سفر به صورت دستی از روی آنها امکان پذیر نمی­باشد. از این رو، استفاده از ابزارهای قدرتمندی مانند یادگیری ماشین و علوم داده کاوی برای استخراج الگوهای با معنا و آنالیز کردن این حجم از اطلاعات نیاز است.

در این پایان نامه، می­کوشیم ابتدا داده های کارت هوشمند را با دیگر منابع داده­ای حمل و نقل عمومی مانند اطلاعات حرکتی (GPS) ترکیب می­کنیم و از روی آنها الگوهای زمانی و مکانی را استخراج می­کنیم و به اطلاعاتی مانند مسیرهای پرتردد، زمان های پررفت و آمد و غیره دست می­یابیم تا سازمان­های حمل و نقل بتوانند با استفاده از این الگوها، کیفیت و خدمات خود را افزایش داده و افراد بیشتری را به استفاده از حمل و نقل عمومی ترغیب سازند. بدین منظور از خوشه­بندی عمیق استفاده خواهیم کرد. در این رویکرد، اولا در نظر داریم که به جای استخراج ویژگی­های مورد نیاز برای الگوهای سفری به صورت دستی، از روشهای استخراج ویژگی خودکار استفاده کنیم. سپس بر اساس ویژگی­های خودکار، مساله خوشه­بندی برای شناسایی خوشه­های محتوی الگوهای مشابه تعریف می­کنیم. در نهایت هر خوشه محتوی الگوهایی است که نیاز به زیرساختهای حمل­ونقل عمومی مشابهی هستند.

ترکیب ساختار این پروپوزال به این ترتیب است. چالش های اصلی مسأله در بخش 2 آمده است. در بخش 3، صورت دقیق مسأله ی استخراج الگوهای سفر از کارت هوشمند بیان می شود. بخش 4 بـــه خلاصه سازی و بیان کارهای گذشتگان در این زمینه و روش های قبلی حل مسأله می پردازد. در بخش 5، روش حل مناسب که برای استخراج الگوها در این پایان نامه مد نظر است، شرح داده می شود. بخش 6 به شرح و توصیف پایگاه داده های انتخاب شده برای مطالعه پرداخته و در بخش 7، روش ارزیابی مدل و الگوریتم مورد مطالعه در این پایان نامه بیان می شود.

1. **چالش های مسأله**

چالش های زیادی در استخراج الگوهای سفر از کارت هوشمند وجود دارد، که ایده پردازی و پرداختن به همگی آن ها در مدت زمان تعیین شده این پایان نامه امکان پذیر نمی باشد. بنابراین، در ادامه چالش هایی آورده شده است که انتظار می رود بتوان در این پایان نامه راه حلی برای آن ها مطرح کرد.

یکی از چالش های این مسأله این است که اکثر سیستم های اتوماسیون پرداخت AFCفقط مبدأ سفر را ثبت می کنند و هنگام پیاده شدن نیازی به کارت زدن نیست، بنابراین مقصد سفر اغلب نامعلوم بوده و باید با روش هایی آن را تخمین زد.

از دیگر چالش های موجود یکتا نبودن کارت هاست، به این معنا که یک کاربر می تواند بیش از یک کارت داشته باشد و یا چندین نفر به طور مشترک از یک کارت استفاده کنند. لذا استخراج الگوهای صحیح این دو گروه چالش برانگیز است.

علاوه بر این، افرادی از زدن کارت هنگام سفر خودداری می کنند و دستیابی به الگوهای آنان با تکیه بر داده های کارتشان به تنهایی امکان پذیر نیست.

در این پایان نامه، کوشش می شود تا راهکارهای نوینی برای حل این چالش ها ارائه شود.

در انتها، تا به حال هیچ معیار ارزیابی عددی برای سنجش کیفیت الگوهای استخراج شده مطرح نشده است، و از تکنیک های مصورسازی به کمک نقشه های جغرافیایی استفاده شده است، که همین امر مقایسه ی راهکار پیشنهادی این پایان نامه از دیدگاه عددی و علوم کامپیوتری با کارهای پیشین و ارزیابی مناسب را دشوار می سازد. به همین جهت، یک معیار عددی در قسمت ارزیابی در نظر گرفته شده که انتظارات را برآورده سازد.

1. **صورت مسأله**

مسأله ی استخراج الگوهای سفر به فرآیندی گفته می شود که طی آن با داشتن داده های کارت هوشمند به عنوان ورودی، الگوهای مکانی و زمانی سفرهای کاربران از آن استخراج می شود. الگوی مکانی یعنی رفت و برگشتی بین چندین مکان که در سفرهای مختلف به طور مرتب تکرار شده است و به طور مشابه، الگوی زمانی یعنی چندین سفر طی چندین بازه ی زمانی که به طور مکرر انجام می شود.

فرض کنیم و به ترتیب نمایانگر تعداد نمونه ها و ویژگی ها باشند. ورودی و خروجی دقیق مسأله در زیر آورده شده است:

**ورودی:** ورودی ماتریس تشکیل شده از نمونه در سطرها و ویژگی در ستون ها است، بطوریکه:

مقدار نمونه ی ام در ویژگی ام

در این مسأله پایگاه داده ی شامل تراکنش های کارت های هوشمند است که در آن هر سطر، یک سفر انجام گرفته است، و ویژگی های معمول این مسأله یعنی ستون ها، شامل مبدأ و مقصد سفر، زمان آغاز و پایان سفر، و شماره ID کارت می باشد.

**خروجی:** خروجی دو ماتریس و است که به ترتیب حاوی الگوهای مکانی و زمانی استخراج شده از هستند و هر سطر این ماتریس ها شامل یک الگو می باشد. هر سطر ، یعنی هر بدین شکل تعریف می شود:

*و* امین ایستگاه مقصد و امین ایستگاه مبدأ

هر سطر ، یعنی هر هم به طور مشابه تعریف می شود:

*و* امین زمان پایان سفر و امین زمان شروع سفر

1. **کارهای گذشتگان**

الگوریتم های یادگیری با ناظر و بدون ناظر هر کدام کاربردهای متفاوتی دارند. در حوزه ی استخراج الگوها، تکنیک های خوشه بندی بدون ناظر کاربردی ترند، زیرا در اکثر داده ها مانند کارت های هوشمند برچسب نداریم، وکشف ساختار الگوهای نهانی مورد بحث است. در جدول 1، خلاصه ای از کارهای گذشتگان برای حل مسأله آمده است:

**جدول 1-خلاصه ای از مقالات مهم در حوزه ی استخراج الگوهای سفر**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **مقاله** | **سال** | **مفاهیم و نتایج** | **معیار شباهت** | **روش** | | **الگوریتم** |
| [2] | 2013 | شناسایی مسافرهای پرتکرار با خوشه بندی مسافرها | ایستگاه سوار شدن و سپس زمان سوار شدن | خوشه بندی Partition-based |  | DBSCAN,  K-means++, C4.5,  Naive Bayes, Neural Network |
|  |  |  |  | خوشه بندی Hierarchical-based |  |  |
|  |  |  |  | خوشه بندی Density-based |  |  |
|  |  |  |  | یادگیری بدون ناظر |  |  |
|  |  |  |  | یادگیری با ناظر |  |  |
| [5] | 2014 | تخمین هدف از سفر با خوشه بندی سفرها | زمان و مکان ایستگاه های سوار شدن و سپس پیاده شدن | خوشه بندی Partition-based |  | Naive Bayes |
|  |  |  |  | خوشه بندی Hierarchical-based |  |  |
|  |  |  |  | خوشه بندی Density-based |  |  |
|  |  |  |  | یادگیری بدون ناظر |  |  |
|  |  |  |  | یادگیری با ناظر |  |  |
| [6] | 2015 | شناسایی مراکز شهری و پررفت و آمد حول ایستگاه ها با خوشه بندی ایستگاه ها | زمان های سوار و پیاده شدن | خوشه بندی Partition-based |  | Agglomerative Hierarchical |
|  |  |  |  | خوشه بندی Hierarchical-based |  |  |
|  |  |  |  | خوشه بندی Density-based |  |  |
|  |  |  |  | یادگیری بدون ناظر |  |  |
|  |  |  |  | یادگیری با ناظر |  |  |
| [7] | 2016 | بدست آوردن 8 گروه متمایز از مسافرها با خوشه بندی مسافرها و تخمین فعالیت بعدی آن ها با دانستن فعالیت فعلی | زمان شروع و مدت فعالیت و land use حـول ایستگاه ها | خوشه بندی Partition-based |  | Continuous hidden Markov model |
|  |  |  |  | خوشه بندی Hierarchical-based |  |  |
|  |  |  |  | خوشه بندی Density-based |  |  |
|  |  |  |  | یادگیری بدون ناظر |  |  |
|  |  |  |  | یادگیری با ناظر |  |  |
| [8] | 2017 | بدست آوردن الگوهای زمانی پرتکرار مسافرها به صورت سفر از خانه به محل کار و برعکس | فعالیت های زمانی مسافرها | خوشه بندی Partition-based |  | Gaussian mixture generative-based |
|  |  |  |  | خوشه بندی Hierarchical-based |  |  |
|  |  |  |  | خوشه بندی Density-based |  |  |
|  |  |  |  | یادگیری بدون ناظر |  |  |
|  |  |  |  | یادگیری با ناظر |  |  |
| [9] | 2018 | تشخیص تغییرات ناگهانی در الگوهای بلندمدت سفر با تخمین احتمال تغییرات با متد بیزی | - | خوشه بندی Partition-based |  | Bayesian approach |
|  |  |  |  | خوشه بندی Hierarchical-based |  |  |
|  |  |  |  | خوشه بندی Density-based |  |  |
|  |  |  |  | یادگیری بدون ناظر |  |  |
|  |  |  |  | یادگیری با ناظر |  |  |
| [10] | 2019 | استخراج 4 دسته ویژگی از داده های اولیه و بدست آوردن الگوها با خوشه بندی | ویژگی های آماری بدست آمده از داده های اولیه | خوشه بندی Partition-based |  | Statistical Feature Extraction,Affinity Propagation (AP) clustering |
|  |  |  |  | خوشه بندی Hierarchical-based |  |  |
|  |  |  |  | خوشه بندی Density-based |  |  |
|  |  |  |  | یادگیری بدون ناظر |  |  |
|  |  |  |  | یادگیری با ناظر |  |  |
| [11] | 2020 | بدست آوردن الگوهای زمانی و مکانی مسافران با تجزیه تنسوری و آنالیز همبستگی | - | خوشه بندی Partition-based |  | NCP Decomposition, correlation analysis |
|  |  |  |  | خوشه بندی Hierarchical-based |  |  |
|  |  |  |  | خوشه بندی Density-based |  |  |
|  |  |  |  | یادگیری بدون ناظر |  |  |
|  |  |  |  | یادگیری با ناظر |  |  |

جدول بالا نشان می دهد که الگوریتم های متفاوتی برای استخراج الگوها به کار گرفته شده است و اکثرا روی شناسایی الگوهای مکانی مسافرها یا ایستگاه های پرتردد تمرکز شده است و عده کمی از سفرها به عنوان شئ خوشه بندی استفاده کرده اند که آن هم برای تخمین هدف از سفر بوده.

از گپ های مقالات می توان به در نظر نگرفتن الگو های بلند مدت اشاره کرد. اکثرا الگوها را در مدت یک روز، هفته یا ماه استخراج کرده اند در حالی که الگوهای بلند مدت با کوتاه مدت متفاوتند و الگوهای کوتاه مدت توانایی نمایش روند واقعی حمل و نقل را ندارند. تعداد کمی از مقالات، داده هایی به جز کارت هوشمند را در نظر گرفته اند در حالی که عوامل مختلفی مثل آب و هوا، شرایط بحرانی و غیره در استخراج الگوها دخیل است. از دیگر گپ های مقالات، عدم توانایی در شناسایی مسافرهایی هستند که بدون کارت زدن سوار و پیاده می شوند یا از کارت های دیگران استفاده می کنند و به این دلیل، الگوی آن ها به درستی ثبت نمی شوند.

1. **کارهای آینده**

راه حل پیشنهادی این پایان نامه جهت حل مسأله به ترتیب شامل پیش پردازش، پاکسازی و آماده سازی داده های کارت های هوشمند، سپس استفاده از یک تکنیک ادغام داده ها (data fusion) جهت ترکیب این داده ها با پایگاه داده های مناسب دیگر مانند داده های آب و هوایی و فصلی، داده های GPS، و غیره و آماده سازی یک پایگاه داده جامع و تلفیق شده­ی حمل و نقل و سپس استفاده از خوشه بندی عمیق برای استخراج ویژگی های قدرتمند و بهینه، گروه بندی داده ها و استخراج الگوهای سفر زمانی و مکانی کاربران می باشد.

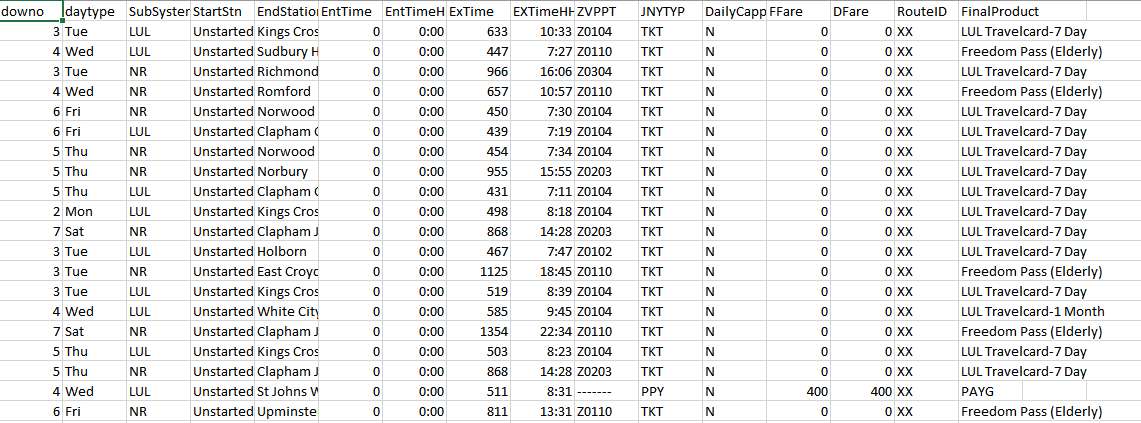
تکنیک خوشه­بندی عمیق عبارت است از بهره گیری از شبکه های عصبی عمیق جهت یادگیری ویژگی های مناسب مورد استفاده برای اهداف خوشه بندی. در سال های اخیر، زمینه های مختلفی مانند پردازش تصویر [12]از خوشه بندی عمیق استفاده کرده اند و نتایج حاصل در مقایسه با خوشه بندی بدون استخراج ویژگی یا با ویژگی های آماری کلاسیک بهتر بوده است.

برای نمونه، مقاله ی [12] این تکنیک را روی تصاویر پایگاه داده های ImageNet و YFCC100M پیاده سازی کرده است، بطوریکه ابتدا با شبکه ی عمیق convnet استخراج ویژگی انجام داده است، و سپس به صورت استقرایی این ویژگی ها را با الگوریتم k-means خوشه بندی کرده و وزن های شبکه را با پیشگویی تعلق اشیا به خوشه ها به عنوان شبه برچسب بروزرسانی کرده است. نتایج حاصل از این روش روی داده ها از تمامی روش های قبلی بهتر بوده است و از طرفی نیاز به دانش خبرگانی مانند تعداد خوشه ها نبوده است بنابراین انتظار می رود که استفاده از تکنیک های خوشه بندی عمیق در حوزه استخراج الگوهای سفر، همانند این حوزه و سایر حوزه ها نتایج مطلوبی به دنبال داشته باشد.

سامانه ی پیشنهادشده روی پایگاه داده ی ذکرشده در قسمت 6 پیاده سازی و اجرا شده و با معیارهای ارزیابی اشاره شده در قسمت 8 تست و ارزیابی خواهد شد.

1. **پایگاه داده**

در این پایان نامه از پایگاه داده ی Oyster Card Journey Information که شامل داده های کارت هوشمند یک هفته ای Oyster شهر لندن می باشد، استفاده خواهد شد. این پایگاه داده از سایت رسمی حمل و نقل لندن بــــــه نشانی <https://api-portal.tfl.gov.uk/docs> دانلود شده است، شامل 5% از داده های یک هفته ای Oyster در ماه نوامبر 2009 می باشد، و دارای 576,048,1 سطر و 16 ستون می باشد. نمونه ای از این پایگاه داده در تصویر 1 مشاهده می شود:



**تصویر 1-بخشی از پایگاه داده ی Oyster Card Journey Information**

در جدول 2، ویژگی های این پایگاه داده به همراه توضیحات هر ویژگی آورده شده است:

**جدول 2-پایگاه داده ی Oyster Card Journey Information**

|  |  |
| --- | --- |
| عددی بین 1 تا 7، 1 نشان دهنده یکشنبه، 2 دوشنبه و غیره | **downto** |
| روز هفته – یکشنبه تا شنبه (Sun to Sat) | **daytype** |
| mode های سفر شاملLUL-London Underground (متروی معمولی) ، قطار شرکتNational Rail، اتوبوس های LTB-London Buses، قطار DLR-Docklands Light Railway، و تراموای TRAM-Croydon Tram | **SubSystem** |
| ایستگاهی که سفر در آن شروع شده | **StartStn** |
| ایستگاهی که سفر در آن پایان یافته | **EndStation** |
| زمان ورود سفر به فرمت دقایق بعد از نیمه شب | **EntTime** |
| زمان ورود به فرمت زمانی HH:MM | **EntTime HHMM** |
| زمان خروج سفر به فرمت دقایق بعد از نیمه شب | **ExTime** |
| زمان خروج با فرمت زمانی HH:MM | **ExTime HHMM** |
| نواحی (zones) بلیت فصلی Oyster، در صورت استفاده | **ZVPPT** |
| نوع محصول سفر، شامل PPY-Pure PAYG، TKT-Pure Oyster Season، و MIXED-Combined Pay Card Oyster Station | **JNYTYP** |
| در صورتی که سفر PAYG محدود شده باشد، مقدار Y به خود می گیرد. | **Daily Cupping** |
| هزینه کامل PAYG قبل از هر گونه تخفیف | **FFare** |
| هزینه PAYG بعد از تخفیف های بر پایه ی مصرف | **DFare** |
| شماره مسیر اتوبوس، در صورت سوار شدن به اتوبوس | **RouteID** |
| مشخصات محصول ترکیبی استفاده شده برای سفر | **Final Product** |

1. **روش ارزیابی**

برای ارزیابی الگوریتم ها و روند پیشنهادی برای حل مسأله، دو رویکرد مختلف در پیش گرفته و نتایج بدست آمده را با هم مقایسه می کنیم.

در رویکرد اول، ابتدا به صورت دستی و با بهره گیری از روش های آماری، ویژگی هایی از روی داده های آماده شده بدست آورده و با الگوریتم های کلاسیک خوشه بندی و یادگیری بدون ناظر که تعدادی از آن ها در قسمت 4 ذکر شد، داده ها را خوشه بندی کرده و الگوها را استخراج می کنیم.

در رویکرد دوم، از تکنیک پیشنهادی خوشه بندی عمیق بهره می گیریم بطوریکه با استفاده از شبکه های عصبی عمیق، عمل استخراج ویژگی از روی داده ها را به صورت اتوماتیک درون شبکه انجام می دهیم و شبکه خود ویژگی هایی قدرتمند استخراج می کند که انتظار می رود از ویژگی های دستی و کلاسیک بهتر باشد. سپس، با همان روش های کلاسیک خوشه بندی که در بالا اشاره شد، مجددا اقدام به استخراج الگوهای سفر می کنیم.

در گام آخر، کیفیت هر گروه از الگوهای بدست آمده به وسیله ی خوشه بندی را با معیار نیم رخ (silhouette coefficient) می سنجیم، که یک روش ضمنی (implicit) است و هنگامی استفاده می شود که هیچ دانش خبرگانی مانند تعداد واقعی خوشه ها در دست نداریم. این معیار خوشه ها را براساس میزان فشردگی درون-خوشه ای و میزان دور بودن یک خوشه از سایر خوشه ها می سنجد. [1]

برای ، با خوشه بندی فرضی ، معیار نیم رخ برای هر نمونه ی عبارت است از:

معیار نیم رخ مقداری بین 1و1- می گیرد. وقتی این مقدار به 1 نزدیک است، یعنی خوشه ی شامل فشرده است و به حد کافی از بقیه خوشه ها دور است که ایده آل است. برعکس، مقدار نزدیک 1- یعنی به اشیای داخل خوشه های دیگر نزدیک تر از اشیای درون خوشه خودش است، که مطلوب نیست. [1]

برای ارزیابی کیفیت یک خوشه، میانگین مقدار نیم رخ همه اشیای آن خوشه را محاسبه می کنند. برای ارزیابی کیفیت کلی خوشه بندی، میانگین مقدار نیم رخ همه اشیای پایگاه داده استفاده می شود. [1]

# **منابع**

|  |  |
| --- | --- |
| [1] | J. Han, J. Pei and M. Kamber, Data mining: concepts and techniques, Elsevier, 2011. |
| [2] | X. Ma, Y. J. Wu, Y. Wang, F. Chen and J. Liu, "Mining smart card data for transit riders’ travel patterns," *Transportation Research Part C: Emerging Technologies,* vol. 36, pp. 1-12, 2013. |
| [3] | H. Faroqi, M. Mesbah and J. Kim, "Applications of transit smart cards beyond a fare collection tool: a literature review," *Advances in Transportation Studies,* vol. 45, 2018. |
| [4] | M. Utsunomiya, J. Attanucci and N. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," *Transportation research record,* vol. 1971, no. 1, pp. 118-126, 2006. |
| [5] | T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: A data fusion approach," *Transportation Research Part C: Emerging Technologies,* vol. 46, pp. 179-191, 2014. |
| [6] | O. Cats, Q. Wang and Y. Zhao, "The identification and classification of urban centres using public transport passenger flows data," *Journal of Transport Geography,* vol. 48, pp. 10-22, 2015. |
| [7] | G. Han and K. Sohn, "Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model," *Transportation Research Part B: Methodological,* vol. 83, pp. 121-135, 2016. |
| [8] | A. S. Briand, E. Côme, M. Trépanier and L. Oukhellou, "Analyzing year-to-year changes in public transport passenger behaviour using smart card data," *Transportation Research Part C: Emerging Technologies,* vol. 79, pp. 274-289, 2017. |
| [9] | Z. Zhao, H. N. Koutsopoulos and J. Zhao, "Detecting pattern changes in individual travel behavior: A Bayesian approach," *Transportation research part B: methodological,* vol. 112, pp. 73-88, 2018. |
| [10] | Y. Zhang, T. Cheng and N. S. Aslam, "Exploring the relationship between travel pattern and social-demographics using smart card data and household survey," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives,* pp. 1375-1382, 2019. |
| [11] | J. Tang, X. Wang, F. Zong and Z. Hu, "Uncovering Spatio-temporal Travel Patterns Using a Tensor-based Model from Metro Smart Card Data in Shenzhen, China," *Sustainability,* vol. 12, no. 4, p. 1475, 2020. |
| [12] | M. Caron, P. Bojanowski, A. Joulin and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. |