Data Mining from Smart Card Data using Data Clustering

Abdul Khanadar, Bhanupriya Sharma, Shambhavi Srivastava
Guided By: Prof Gitanjali J

School Of Information Technology and Engineering
VIT University Vellore, Tamil Nadu, India.

Abstract:
The aim of this paper is to develop an effective methodology for the better understanding of the travelling patterns and evaluating behavioral attributes of traveller’s trip. Using smart card data, the data such as boarding location, boarding time, alighting location and alighting time of the traveler is collected and through this data, behavior of the traveler is understood. Methodologies used are pattern recognition and data clustering. This implementation would facilitate the transit authorities to improve the transport service.

Keywords: Pattern recognition, Clustering.

Introduction
More or less 74% of those studying in the VIT UNIVERSITY uses bus facility to reach their destination as the university spread 350 Acers of area building are construct in vast area hostellers use means of bus and information gathered for the daily travel. This example is presently getting to be evident in creating best university, for example, VIT University use a lot of student depend on bus facility, for instance, more than 32% of day scholars uses their own vehicle. Open travel has long been considered to give a powerful approach to decrease blockage, air contamination, and vitality utilization. To enhance travel most frequently of bus and sway more individuals to utilize open travel, travel offices have been striving to recognize the key components that draw in travel riders through mulling over their travel designs. With a superior understanding of the travel examples of travel riders, travel powers will have the capacity to assess their current administrations to uncover how best to change their advertising techniques to energize higher utilization. For sample, knowing why a few students are particularly reliable to travel can help travel organizations to figure out where and when they ought to best bus faculty. Taking into account distinguished travel examples and travel utilization regularities, travel powers have the capacity access the practical passage bundles for travel students, university organizers and explores can likewise use individual travel-conduct information for action based outing displaying and travel interest breaks down. Individual student conduct information can uncover how TOD inhabitants change their everyday driving practices and how student use shifts spatially and transiently. The utilization of smart card information to track student’s long haul travel exercises and examples, for example, the quantity of average day by day excursion chains, basic sheets/landing stops and excursion begin/end times, offers a significantly more advantageous and productive information source. Smart card information records both transient also spatial data for every student; individual based excursion conduct is a possibly exceptionally difficult theme. Used various day student smart card information to break down travel student travel variability and pointed out that creating a finer understanding of travel variability can help decrease operational expenses and oversee request. Again more excursions amid regular weekdays, and found that travel examples shifted via card sort. The greater part of the previously stated examination focused around smart card information contracted travel conduct data visibly instead of by investigating individual travel riders’ travel designs. Their dissection was focused around top notch information with complete data furthermore their method was not advanced for a substantial dataset. In all actuality, VIT University have embraced an exhaustive technique to store keen card information, giving strict approval and security instruments to ensure the individual data produced from shrewd card information. Delicate substance, for example, student age, name, sheets and landing areas are purposefully truncated to destination, so productive information mining methodologies are required to surmise student travel conduct data from these fragmented savvy card datasets. Concentrating student travel designs from shrewd card information can be especially difficult on the grounds that the Automatic Fare Gathering (AFC) framework was not initially intended to help travel arranging and travel execution measures accordingly, the brilliant card information gathered by the AFC framework fails to offer certain trek related data that influences information handling execution. To manage this information issue, this paper proposes a powerful and far reaching information mining strategy to concentrate individual travel riders’ travel examples and normality from a huge dataset with inadequate data. Particularly, two real issues are analysed in this study. Initially, the spatial and fleeting travel designs for a specific travel rider are researched. Here, “spatial travel example” implies that the student rider over and again visits the same or contiguous puts on a multi-day premise and “transient travel example” implies the student rider begins (and/or completes) his/her day by day trek amid the same time period. At that point we move onto focus the “consistency” of a student rider’s travel design, which alludes to “recurrence of the comparative excursions for this travel rider”, and the recurrence of the similar trips can be viewed as a successful estimation of travel consistency. The goals of this study are to aid both travel offices and transportation analysts by: (1) creating a novel information mining strategy to concentrate individual travellers’ travel examples and travel normality; and (2) guaranteeing these information mining calculations are equipped for
transforming gigantic savvy card datasets inside a fair slipped by time. The rest of this paper is sorted out as takes after.

Datadescription

VIT University has provided smart cards in form of ID cards to their students that could be used for tracking the student and improving the transit services. Transit riders are required to swipe their smart cards when checking-in and checking-out. They have to hold their smart cards near the card reader device to complete the process of entering or exiting buses. Due to which information regarding the boarding and alighting locations including the boarding and alighting time is stored. Key information stored in the database therefore includes smart card ID, route number, driver ID, boarding stop and time, alighting stop, and alighting time.

To show the temporal travel patterns and the pattern regularity for transit riders in VIT University, consider a typical travel week that is the week of exam (the week of Saturday October 4th to Sunday October 12th, 2014). The transaction data from 24,000 smart cards was collected for that week. According to this information it seems that most of the students began their first trip between 7:25AM and 7:55AM, and ended their travel for the day between 4:00PM and 7:00PM. This is probable to represent a typical commuting trip chain, where a student takes a bus from his or her hostel to their place of work and then returning back. The temporal distribution implies that strong temporal travel patterns exit in the multiday smart card data. However, the regular spatial travel pattern for a specific card holder remains uncertain and will be explored in the analysis.

Methodology

The two main objectives of this study travel patterns are pattern recognition and regularity mining. A flow diagram of the work performed for the study is illustrated

1. Extracting per day smart card data of each sitter from the database;
2. Using space and time relationship extract passenger’s trip chain.
3. Extracting the passenger’s travel pattern and travel regularity based on the generated trip chains by applying sequential data mining approach.

Extraction of trip chain

The prior step for examining the distance and time patterns of traveller is construction of trip chain information, i.e. for construction of trip chain information we need to study pattern recognition. A trip chain is nothing but day to day trips made by a traveller serially and can be used to estimate and study travellers daily travelling behaviours. Normally when people use smart cards while travelling they will swap their card only while boarding the vehicle. These smart card readers are not masterly in recording loading location, descending stops or when and where they get off. For studying and estimating behavioural attributes, Markov introduced a Markov chain based Bayesian decision tree algorithm which can be used concentrate and distil changes in the number of travellers loading at each stop with time. This information can be utilized by associating it with historical speed profiles. Finally, we can select the boarding stop by comparing probabilities so that the stop with the highest probability can be assumed as the boarding stop.

Table 1: Sample extracted trip chain information

<table>
<thead>
<tr>
<th>Chain ID</th>
<th>Card ID</th>
<th>Date</th>
<th>First Boarding Time</th>
<th>Last Alighting Time</th>
<th>Route Sequence</th>
<th>Step ID Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>140101</td>
<td>14MSN0086</td>
<td>4/10/2014</td>
<td>7:07:07</td>
<td>7:20:00</td>
<td>101-102-109</td>
<td>9999 992 983</td>
</tr>
<tr>
<td>140102</td>
<td>14MSN0086</td>
<td>5/10/2014</td>
<td>7:07:09</td>
<td>7:20:00</td>
<td>101-104-109</td>
<td>9999 890 983</td>
</tr>
<tr>
<td>140103</td>
<td>14MSN0086</td>
<td>6/10/2014</td>
<td>6:00:00</td>
<td>6:10:00</td>
<td>109-112-114</td>
<td>983 792 991</td>
</tr>
<tr>
<td>140104</td>
<td>14MSN0085</td>
<td>4/10/2014</td>
<td>7:30:00</td>
<td>7:59:14</td>
<td>111-113-114</td>
<td>791 793 991</td>
</tr>
<tr>
<td>140105</td>
<td>14MSN0024</td>
<td>4/10/2014</td>
<td>7:30:08</td>
<td>7:59:14</td>
<td>112-113-114</td>
<td>792 793 991</td>
</tr>
<tr>
<td>140106</td>
<td>14MSN0073</td>
<td>4/10/2014</td>
<td>7:30:09</td>
<td>7:59:15</td>
<td>112-113-114</td>
<td>792 793 991</td>
</tr>
<tr>
<td>140107</td>
<td>14MSN0114</td>
<td>4/10/2014</td>
<td>7:30:10</td>
<td>7:59:16</td>
<td>112-113-114</td>
<td>792 793 991</td>
</tr>
<tr>
<td>140108</td>
<td>14MSN0119</td>
<td>4/10/2014</td>
<td>7:30:11</td>
<td>7:59:17</td>
<td>112-113-114</td>
<td>792 793 991</td>
</tr>
<tr>
<td>140109</td>
<td>14MSN0098</td>
<td>4/10/2014</td>
<td>7:39:11</td>
<td>7:59:18</td>
<td>112-113-114</td>
<td>792 793 991</td>
</tr>
<tr>
<td>140110</td>
<td>14MSN0076</td>
<td>4/10/2014</td>
<td>7:59:14</td>
<td>7:59:19</td>
<td>112-113-114</td>
<td>792 793 991</td>
</tr>
</tbody>
</table>
Table shows the illustration of the survey done on VIT student’s trip in order to study each passengers travelling pattern. In this table, Chain ID can be described as a single identifier for each one excursion. Next is the card id which is the id of the travel rider. Initially first boarding time is the time in which voyager enters the transport and last alighting time is the time in which explorer gets down from the transport. For each one Card ID, the first excursions loading up time (First Boarding Time) and the last excursion’s landing time (Last Alighting Time) are connected with that Chain ID. Let us take the case of chain id 1401. This chain id has the same card id that chain id 1402 has that is 14msn0086. Route Sequence alludes to the routes the rider took and Stop ID Sequence alludes to the sheets and landing stop Ids for bus transports in VIT University. Recently no such type of study of travel behavior of particular person is proposed. This idea will provide a helping hand to administration management who can expand the frequency of transports considering some travelling examples of the travelers.

Take Chain ID 140101 as a sample. The travel rider sheets the separation built passage transport in light of Route at Stop ID 101, and landed at Stop ID 109.

**Student Travel Pattern Identification**

As soon as the sitter’s travel chain table has been built, information can be extracted from the given data. The obtained data can be utilized in generating the travel design of every student by clustering the given trip chain data. As shown in table 1 entries some sitter’s have a certain way of travelling in bus so they parade some specific design. These masked patterns of the student travelling can be extracted with the help of density based spatial clustering algorithm with noise (DBSCAN). Since more number of records have to be grouped together in the clusters so we need to add some more steps to the DBSCAN calculation. Basically 2 key parameters do, in any case, need to be characterized in the DBSCAN algorithm: the distance (e) separation and the (Min pts) minimum number of points. The e variable symbolizes the separation or the distance between two points. Let us take an example that if 2 records fall within some particular distance then that record can be clustered together. Min pts symbolizes the minimum number of records in each cluster. For example; if in some cluster the number of records exceeds min pts then that will be considered as clamor. On the off chance that the records are near one another (i.e. more thick), these records are more inclined to be clustered by DBSCAN.

A student may start their outing in both spacious and timely domains i.e. with respect to space /distance or time respectively. So for clustering each students time and distance are considered. On the basis of these students can be clustered in different groups. In our study we have taken minimum 50 students that can be clustered in 1 group. There may be chances that some students getting off stops may be adjacent to one another then these stops are considered as same or identical stops. To perform all these calculation. Along these lines, an extra calculation was utilized to distinguish the relationship between numerous courses and connected at the present time DBSCAN grouping, as takes after

Step 1: randomly recover the unvisited record from the database.

Step 2: if the time contrast between unvisited and last visited record is more than 1 hr rehash step 1.

Step 3: if the relationship between the visited and unvisited record exist inside (200 meters) then add into cluster made in step 1.

Step 4: for each 1 cluster, if the quantity of aggregate records is short of what 50, then no more records can be entered.

Step 5: continue to process those unvisited records from step 1 through step 4 until all the records are hailed as visited.

Step 6: the quantity of total clusters is the quantity of average trip chains for every day.

The repeating course, sheets/landing stops and timings can be obtained by numbering the most frequent pattern inside each bunch.

Take the outing chain information from table 1 as a sample. Taking into account the DBSCAN bunching calculation, a few examples can be gathered:

1) This travel rider routinely begins his or her first trek around 07:07:07 am.

2) Recurring courses happen for generally weekdays. Sometimes the unusual travel example is caught, which is hailed as the noise by the DBSCAN algorithm.

3) As explained before riders may have taken different routes to the same stop then those routes will be considered as “identical stops”.

**Regularity clustering**

Regularity clustering is for extracting an individual travel pattern, clustering the identical travel pattern and locates them into different regularity levels according to their temporal and spatial characteristics.

In this reference regularity intimates “recurrence of the indistinguishable trips for the students.” In the trip chain data multifarious qualities were so picked as features for clustering as takes after:

1. Number of travel days. The more days a transit rider voyages, the more probable it is that he or she is a regular transit rider.

2. Number of identical First Boarding Times. Boarding time indicates a rider’s temporal characteristics. If a rider begins their trip at the same time of a day ever weekday, then this rider is more prone to be a frequent transit rider.

3. Number of identical Route Sequences. A general spatial pattern for a rider was depicted by the route sequence. The number of same route sequences emulated during the week may show a monotonous travel design.

4. Number of identical Stop ID Sequences. The Stop ID sequence may comprise of elaborated spatial similarity information. In several situations, two diverse stop Ids may be spatially adjacent, which can be recognized by GIS buffer processing.
There may be correlation to some extent between the selected features that is number of identical Root Sequence and Stop ID sequence. In spite of this, the elimination of the correlated features should not be done as there may be some missing values and redundancy to some level which help in improving the algorithm accuracy, as redundant features lead to more accurate clustering results.

For efficient and effective cluster regularity, an appropriate clustering algorithm is requiring. The K-means is a well known clustering algorithm in which n records are grouped into k clusters by minimizing the within-cluster sum of squares. Therefore by consecutively updating the record value means, each observation is allocated into the

$$Z = (v - \min(v))/ (\max(v) - \min(v))$$

(1)

K-Means++ was therefore opt for the cluster transit riders with identical travel patterns, and each standardized variable can then be consolidated amid the travel pattern clustering process. Clusters of regularity formed here are in five form, they are: Very High(VH), High(H), Medium(M), Low(L), Very Low(VL). The cluster centres are:

$$C_1 = (V_{11}, V_{12}, V_{13}, V_{14})$$

$$C_2 = (V_{21}, V_{22}, V_{23}, V_{24})$$

$$C_3 = (V_{31}, V_{32}, V_{33}, V_{34})$$

Where $V_{ij}$ indicates the jth feature and i indicates the ith cluster.

Cluster centre distance is the Euclidean distance which is calculated between $C_i$ and zero point

$$D_1 = \sqrt{(V_{11} - 0)^2 + (V_{12} - 0)^2 + (V_{13} - 0)^2 + (V_{14} - 0)^2}$$

$$D_2 = \sqrt{(V_{21} - 0)^2 + (V_{22} - 0)^2 + (V_{23} - 0)^2 + (V_{24} - 0)^2}$$

$$D_3 = \sqrt{(V_{31} - 0)^2 + (V_{32} - 0)^2 + (V_{33} - 0)^2 + (V_{34} - 0)^2}$$

Now $D_i$ is sorted in descending order and the nearest regularity is allocated to clusters based on their order. Then finally for each transit rider corresponding regularity level is estimated through evaluating and examining the minimum distance to the centre of each cluster. According to this clustered result the transit riders are grouped differently, such as regular transit rider have Very High(VH) and High(H) regularity levels.

So in the VIT University having 23000 students uses smart card for their bus facility through which they get some frequency of buses to reach their destination. This is done using DBSCAN algorithm and K-Means ++ clustering algorithm.

**Performance augmentation**

In VIT University more than 30,000 understudies are available. So we need to consider the travelling data for these understudies utilizing smart card. Keen card transaction information focuses are created constantly. Preparing and bunching of information is possible by K-Means ++ algorithm. Accordingly, the K-Means ++ approach is a consistent PC, a calculation taking into account the rough set hypothesis was accordingly connected to enhance bunching execution. Through k-means grouping calculation 5 sort of bunches {C1, C2, C3, C4, C5} named as {VL(very low), L(low), M(medium), H(high), VH(very high)}.
day excursion chains, basic boarding and alighting time and excursion begin/end times, offers a significantly more advantageous and productive information source. Smart card information records both transient also spatial data for every student, making it plausible to lead individual student design examination through longitudinal breaks down Most past savvy card-based exploration into travel voyager practices has concentrated student travel utilization and access separation, reporting that travel use information can give helpful data to travel arranging and statistical surveying, stressed the vitality of building reliability and developed a few measures (counting fulfilment and nature of administration) to measure travel devotion, despite the fact that these discoveries were focused around information accumulated.

The greater part of the previously stated examination focused around smart card information concentrated travel conduct data visibly instead of by investigating individual student travel designs. Connected the affiliation standard also bunching calculations to measure student record normality, and led an individual travel conduct examination utilizing both transient and spatial routines. Nonetheless, their dissection was focused around top notch information with complete data furthermore their method was not advanced for a substantial dataset. In all actuality, most travel offices have embraced an exhaustive technique to store keen card information, giving strict approval and security instruments to ensure the individual data produced from shrewd card information. Delicate substance, for example, student age, name, boarding and alighting are purposefully truncated to address protection concerns, so productive information mining methodologies are required to surmise traveller travel conduct data from these fragmented savvy card datasets. Concentrating travel student travel designs from shrewd card information can be especially difficult on the grounds that the Automatic Fare Gathering (AFC) framework was not initially intended to help travel arranging and travel execution measures. To manage this information issue, this paper proposes a powerful and far reaching information mining strategy to concentrate individual travel student's travel examples and normality from a huge dataset with inadequate data.

Particularly, two real issues are analysed in this study. Initially, the spatial and fleeting travel designs for a specific travel student are researched. Here, "spatial travel example" implies that the travel rider over and again visits the same or contiguous puts on a multi-day premise and "transient travel example" implies the travel rider begins (and/or completes) his/her day by day trek amid the same time period. At that point we move onto focus the "consistency" of a travel student travel design, which alludes to "recurrence of the comparative excursions for this travel student", and the recurrence of the similar trips can be viewed as a successful estimation of travel consistency. The goals of this study are to aid both travel offices and transportation analysts by:

1) Creating a novel information mining strategy to concentrate individual student travel examples and travel normality

Comparison of data mining Algorithms

Practical passage bundles for student see how student practices are prone to change because of a few charge structures, and accordingly select a passage strategy that attains the ideal harmony between upgrading the engaging quality of the student framework and boosting passage income. University organizers and explores can likewise use individual student-conduct information for action based outing displaying and travel interest breaks down. Data on the travel designs for individual student can likewise be used to evaluate the adequacy of distance travel. Specifically, individual student conduct information can uncover how TOD inhabitants change their everyday driving practices and how student use shifts spatially and transiently. In any case, getting individual record customary travel design examination to a great extent depends on rider fulfilment reviews or travel journals, which is unreasonable and troublesome to actualize at a multiday level because of the low reaction rate and accuracy. The utilization of, smart card information to track student long haul travel exercises and examples, for example, the quantity of average day by

After clustering average time for each cluster can be calculated on the basis of which peak hours for student travelling can be evaluated.

Figure 3: Regular day's students travel trip

Figure 4: On the days of examination students travel trip
2) Guaranteeing these mining information calculations are equipped for transforming gigantic savvy card datasets inside a fair slipped by time.

The rest of this paper is sorted out as takes after. The information utilized as a part of this study is initially presented.

Discussion

This study opens up fascinating new open doors for leveraging brilliant card information to make a superior understanding of travel of student conduct and therefore possibly enhance open travel frameworks. Particularly, three significant potential applications that could profit from this study can be speculated, as follows:

• Travel conduct research

In the recent decades, travel interest examination has moved from an outing based go methodology to a movement based travel standard. Movement based travel models oblige a generous measure of peak hours data for every explorer stretching out over a moderately long stretch. Customarily, obliging a gigantic measure of assets to process and build groupings of spatiotemporal exercises. The individual-level travel example digging calculations created for this study offer an option and novel methodology for measuring the comparability and variability of travel student through an examination of their multi-day shrewd card transactions. this will significantly encourage travel conduct demonstrating improvement.

• Transit OD estimation

An alternate potential utilization of the proposed travel example and travel consistency mining calculations is to enhance the precision of the travel OD estimation system. Each one travel rider's redundant authentic courses and stops can be utilized as former data for destination stop induction.

Conclusions

The study has proposed a series of efficient and effective data-mining approaches with which to model student’s travel the patterns used is smart card data of the type collected in the DBSCAN algorithm was utilized to successful detect each students travel records using the identified trip chains. The K-Means++ clustering algorithm and the rough set theory were then jointly applied to classify the travel pattern regularities. The performance of the resulting rough-set-based algorithm was compared with four other classification algorithms outperformed all the other data mining algorithms in terms of accuracy and efficiency such as distance the studies is focus group discussions and travel diaries, in order to improve the algorithm accuracy. It would also be interesting to integrate the student travel pattern information obtained through this study with map-based transportation systems.

As we hope by the means of smart card recording VIT University committee I will react the facts of concepts we use and increase the facility of the students to reach their destination by number of buses

References