Potential Trend Discovery for Highway Drivers on Spatio-temporal Data

Weilong Ding[1,2,†, *],Zhe Wang[1,2], Jun Chen[1,2], Yanqing Xia[3], Jianwu Wang[4] and Zhuofeng Zhao[1,2]

[1] School of Information Science and Technology, North China University of Technology, Beijing, China
[2] Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data, Beijing, China
[3] Ministry of Information Technology, CITIC Bank, Beijing, China
[4] Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD, U.S.A.

[*] Corresponding authors' addresses: No. 5 Jinyuanzhuang Road, Shijingshan District, North China University of Technology, Beijing, 100144, China

[†] Corresponding author. dingweilong@ncut.edu.cn

ABSTRACT

Inter-city transportation plays an important role in modern cities, and has accumulated massive spatio-temporal data from various sensors by IoT (Internet of things) technologies. Travel characteristics and future trends of highway behind data are valuable for traffic guidance and personalized service. As a routine domain analysis, trend discovery for highway drivers faces challenges in processing efficiency and predictive accuracy. Insufficient profiles of those drivers are available directly, sensible executive latency on huge data is hard to guarantee, and inadequate features among spatio-temporal correlations hinder the analytical accuracy. In this paper, a travel-characteristic based method is proposed to discover the potential trend of payment identity for highway drivers. Considering time, space, subjective preference and objective property, travel characteristics are modeled on monthly data from highway toll stations, through which predictive errors can be reduced by gradient boosting classification. With real-world data of one Chinese provincial highway network, extensive experiments and case studies show that our method has second-level executive latency with more than 85% F1-score for trend discovery.

KEYWORDS

spatio-temporal data, travel characteristics, potential trend, ensemble learning, highway, Big Data.

1 Introduction

With the flourish of inter-city transportation, highway plays an important role in modern cities, and most urban drivers have participated in it unconsciously. It also brings traffic congestion issue, one of the most serious problems worldwide nowadays [1, 2]. Accordingly, highway IoT (Internet of Things) is built for official traffic management on various business data of extensive deployed sensors [3]. For instance, there are surveillance data from cameras, RFID (radio frequency identification devices) card data on cars, statistical data from inductive loops at stations, and weather data from monitors in meteorological stations. Big Data technology has been widely adopted in domain analytics recently [4]. As a typical one, *trend discovery* is to predict some attributes' updates for individual drivers or group drivers in recent future. Potential trend of citizens in smart cities is an active research area [5, 6]. That is useful for traffic guidance to alleviate congestion [7] and for personalized service to improve users' experience [8]. For trend prediction, highway drivers' travel characteristics are necessary and often employed from business data like toll data. Such toll data generated from toll stations keeps the timestamps and the locations when a vehicle enters or exits a station, and has the advantages of exact locality and higher quality [9].

However, it faces challenges to predict drivers' potential trends due to inherent limitations in practice. First, insufficient personal profiles of drivers are available directly for highway domain. For example, driver licenses are kept by police, and ETC (electronic toll collection) accounts are maintained by corresponding banks. Such information cannot be accessed externally for highway management due to privacy and security restrictions. The absence of profiles makes highway business analytics difficult to portrait drivers. Second, it is hard to hold low latency during analytical calculation when data grows into a huge size. Classic statistical models only perform well on limited samples at given spatial points, which are not suitable for trend prediction on massive accumulative data. For example, ARIMA regression treats sensory data as time series, and achieves results only for a single toll station [10]. It is inefficient for huge data from hundreds of stations in highway network. Third, spatio-temporal characteristics are not fully considered by traditional methods, which hinder predictive accuracy. For example, besides temporal patterns emphasized by classical time series models, spatial feature of localities (e.g., road network topology), personal preferences, and travel modes in history also influence travel trends of highway drivers. In brief, such problems have not been addressed properly yet, and it is not trivial to find potential trends for highway drivers.

In this paper, we take *payment identity* of highway drivers as an example, and propose a travel-characteristic based method to discover the potential trend of that attribute. Our contributions can be concluded as follows. (1) To describe business feature only on toll data, travel characteristics are modelled fully considering time, space, subjective preference, and objective property. Such characteristics are efficiently organized as dedicated drivers' trajectories. (2) To improve performance and accuracy of trend prediction, an ensemble-learning model through gradient boosting classification is built. With second-level executive latency on monthly data, our method can reduce predictive error about 2%-14% than traditional ways. (3) Evaluated quantitatively and qualitatively in a practical scene, our work shows convincing benefits on the real-world data through extensive experiments and case studies.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 presents the background including motivation and methodology. Section 4 elaborates trend discovery method for highway drivers' payment identity. Section 5 evaluates the effects in extensive experiments and case studies. Section 6 summarizes the conclusion.

2 Related work

Potential trends of urban citizens are significant nowadays, but their discovery still faces challenges in efficiency and accuracy [11]. Many works have adopted Big Data technologies in specific domains [12]. We categorize related studies into two technical perspectives: one is user profiling method to depict users' characteristics through offline processing, and the other is concept drift to find progressive attributes' change in online data condition.

2.1 User profiling method

Through offline collecting, organizing and inferring procedures on massive data, user profiling is to achieve users' interests, characteristics, behaviors, and preferences. To complete comprehensive information representation and efficient personalized service, it is often realized as tagging model. User profiling technologies can be described in various categories.

According to feature extraction manner, user profiling methods can be categorized into two types. (1) One is *explicit profiling*, in which users are depicted by explicitly defined features from data like demographics [13], social media check-in, and travel behaviors [14]. Such methods are straight-forward, but highly depended on detailed user-related profiles [15]. (2) The other is *implicit profiling*, which includes collaborative methods, latent factor models, network embedding and machine learning. Here, collaborative methods are commonly used in early periods, which assume that users in the same group behave analogously [16]. Such methods suffer from the data sparseness of multi-dimensional attributes. Latent factor models, such as matrix factorization or topic modelling, are employed to build profiles as latent factors or categorical distributions [17]. Such methods have to handle optimization problems in large parameter space, but are apt to be over-fitting on given data. Moreover, domain knowledge as the premise for data regularization is often lacked. Recent studies focus on machine learning methods to find users' hidden habitual behaviours [18]. In fact, those methods in either type have their own advisable aspects, and our solution in this paper considers the advantages in both ways. The travel characteristics we defined include explicit features (e.g., objective facts) and the implicit ones (e.g., subjective preferences). With the machine learning technology of gradient boosting classification, better predictive effects are achieved.

According to the nature of modeled profiles, user profiling methods can be categorized into other two types. One is for *static profile* on entire data set. Such profiles rely on steady interests in a long time period. For example, a learning method is proposed to predict volunteerism among social network users [19] through a binary classification. However, such methods fail to perform well on the mutable features like certain users' trends. The other is for *dynamic profile* on progressively accumulative data. Such profiles would vary in short-term or long-term. The short-term profile depicts users' current interests, while the long-term one represents sustainable feature. Considering behavioural or adaptive attributes, such profile is adopted on the data in high velocity. For example, dynamic user profiling is proposed by word embedding [20], to track the semantic representation of words and users dynamics over time in Twitter. Our work in this paper lies in dynamic profile generated on periodically accumulative data in highway domain, and aims to find potential long-term trend (i.e., monthly identity drift) among highway drivers. Without detailed personal profiles, the travel characteristics can be extracted and inferred through the model we proposed.

Domain-dependency is one of the shortcomings among current works [21], which often requires business rules and experts knowledge. In transportation related domains, user profiles are used to describe either group user or individual user. In most cases, large dataset is required for acceptable accuracy during modeling. On heterogeneous traffic data, a three-fold influence model is proposed on group users (i.e., crowd) for their traffic propagation [22], which finds cascading patterns through maximizing probability likelihood. On massive trajectory data, a peer and temporal-aware representation learning based framework is presented [23] for drivers' behaviour analysis through multi-view driving state transition graphs. On massive POI check-in data of passengers, adversarial sub-structured representation learning is introduced [18] for individual user profiling. Through machine learning technologies, those works have performed well in some domain scenes but still remains challenges in highway. Unlike the scenes above, personal detailed profile of highway drivers are not available due to cross-domain security restriction. Only dynamical behavioral data (i.e., toll data recording drivers' trips) is accessible. The travel characteristics are built through trajectory structure for efficient feature maintenance and further applications.

In summary, it still exists great chances for us to improve user profiling studies with novel solution in highway domain.

2.2 Concept drift detection

On continuous data, multi-dimensional representation of users may vary over time. Especially in behavioral profiles, changes often come from users' mutative interests. Accordingly, online detection of concept drift emerges in recent years. Here, classification (including narrow classification on labeled data and clustering on unlabeled data) with underlying data distribution are termed as concept. The classification changes on data, termed as concept drift [24], reflect data distribution and often deteriorate the performance of pre-built classifiers [25]. Due to distinct criteria of changes, different types of concept drifts can be categorized. Considering the problem and data condition in this paper, we only demonstrate concept drift detection for narrow classification on labeled data here. Such detection approaches have to trade-off between performance and cost. Concept drift can be seen as the change of joint probability distribution [26] among data samples and their corresponding labels. A basic assumption here is that the timing and distribution of concept drift are initially unpredictable. Accordingly, most learners have to detect a concept drift first, and then react to it by new learned distribution with an updated classification model.

For the concept drift with constant features and classes, *single drift* on labeled data means that detection process only looks for current drift without previous ones. It is related to our work, and two main types of detection exist in current studies. (1) *Statistical-test based method* monitors the online trace of error rates and detects deviations. When significantly increased error rate appears, a concept drift is assumed to be. A framework is proposed with a hierarchical set of hypothesis testing [27] to detect concept drift through Linear Four Rates test. DDM [28] using Fisher's exact test is introduced to enable detection on data stream, and achieves well performance than counterparts. Multiscale Drift Detection [29] employs re-sampling and paired student's t-test, and emphasizes on lowering computation cost of concept drift detection. Those methods have advantageous performance on data samples through statistics based models, but are not suitable for high-velocity data due to infeasible sampling. (2) Ensemble-learning based method can improve predictive effects by composing multiple base classifiers. The performance of either integral ensemble model or individual base classifier has been studied for concept drift detection. Diversified dynamic-weighted majority (DDWM) [30] works interestingly: its base classifier would be removed if accuracy falls too low, and a new base one would be appended when accuracy is comparable or better than the global. EnsembleEDIST2 [31] approach utilizes three methods using EDIST2 to track ensemble's performance. An ensemble of partial least square model is proposed [32] for applications in Melamine resin production. A committee disagreement measurement is calculated, and the changes are detected using PageHinkley statistic on specific metrics. Such methods could achieve good predictive effects, while time consumption of base classifiers' training is no longer ignorable.

In transportation related domains, concept drift detection is widely used for control centers to predict traffic conditions [33]. Concept Neurons framework [34] empowers the resistance of algorithms for concept drifts. It leverages on a combination of continuous inspection schemas and residual-based updates over model parameters with output. To handle different drift types, it has been successfully applied on predicting highway traffic congestion in Porto. Their domain scenes and focused technical problems are different with ours. In this paper, the concerned problem can be regarded as a long-term single drift on labeled data in highway domain, but the monthly period here is too long to endure in an online stream processing. Inspired by the ensemble-learning methodology, we introduce our model on accumulative data instead of real-time one to find drivers' trends in a periodical reaction manner.

In summary, with the help of concept drift detection, novel classification model is required to discover drivers' potential trends in highway domain.

3 Background

3.1 Motivation

Our research originates from *Highway Big Data Analysis System* in Henan, the most populated province in China. The system we built has been in production since October 2017 and is expected to improve highway analytics through Big Data technologies. Operated by officers of *Henan Transport Department*, a billion records of heterogeneous data in recent two years have been loaded into the system. There are real-time toll data from toll stations, daily meteorological data from weather stations, solar and lunar calendric data from dedicated interfaces, and real-time license plate recognition data from surveillance cameras, etc. The toll data is our focus in this paper. As Table 1, a record of toll data is contains 12 attributes including six entity attributes, two temporal attributes and four spatial attributes.

Table 1: The structure of toll data.

Attribute	Notation	Example	Type
collector_id	toll collector identity	XXXX080169	
vehicle_license	vehicle identity	蓝豫 AA7R62	
vehicle_type	vehicle type	1	Entity
card_id	vehicle passing card identity	4101152822010XXXXXXX	
etc_id	vehicle ETC card identity	XXX7887	
etc_cpu_id	ETC card chip identity	XXX102	
entry_time	vehicle entry timestamp	2018/2/23 15:32:06	Time
exit_time	vehicle exit timestamp	2018/2/23 16:38:19	
entry_station	identity of entry station	33011	
entry_lane	lane number of entry station	2	Space
exit_station	identity of exit station	33012	
exit_lane	lane number of exit station	1	

As one significant business analytics in domain, highway drivers' trend discovery would predict individual drivers' attributes or categories in recent future through travel characteristics. The value of attribute or category would be variable in different perspectives, such as resident location of a vehicle, interested destinations, and payment identity, etc. The category *payment identity* for individual driver would be either “ETC (Electronic Toll Collection)” or “MTC (Manual Toll Collection)”, and is taken as an example to elaborate our method in this paper. In fact, ETC as a non-stopping payment technology is widely adopted in highway to promote drivers’ charging and passing efficiency. As we have discussed above, a driver has to create his ETC account with personal detailed profile in a corresponding bank, before he can be charged as this payment identity.

We find the following business observations from extensive domain research.

Observation 1. Although without detailed personal profile, drivers’ payment identity can be recognized by toll data. A record of Table 1 implies the payment behavior of a driver v in a trip: v is charged as ETC if *etc_id* or *etc_cpu_id* is not empty; otherwise he is as MTC.

Observation 2. The payment identity drift of a driver derives from multiple private reasons, while the trend among multiple ones is valuable for officers to react in their management and public services.

As a routine business analysis, the *trend discovery of payment identity* focused in this paper is to periodically detect the drift of this identity for individual highway drivers in the coming month. In traditional ways, historical toll data of sensors would be loaded into a production data warehouse regularly (e.g., monthly or even yearly); after ETL (Extract, Transform, Load) step with necessary pre-processing like [35], business OLAP (Online Analytical Processing) would be triggered to execute in that data warehouse; when completed, predicted results for drivers’ payment identity or their probabilities can be accessed by business technicians for further interpretation. However, such prediction brings long delays (e.g., one week or more) in practice to release official reports due to complex processing. Moreover, traditional models widely used like classic Logistic Regression and Decision Tree do not fit well on huge data from hundreds of stations, because their predictive errors are only qualified on limited samples at a single location. Accordingly, a novel method is required for trend discovery to improve both latency and accuracy on massive data without detailed personal profiles. It is just our original motivation.

3.2 Methodology

In highway domain, potential trend of payment identity can be evaluated by recent travel behaviors of individual drivers. Such trend could be depicted in different temporal or spatial granularity, such as short-term, long-term, single driver and group driver. In this paper, we focus on discovering the monthly trend of individual highway drivers defined below.

Definition 1: Payment identity trend. The payment identity trend of a highway driver $v \in V$ is presented as the predictive probability pair of category ETC and MTC in coming month m. In such a pair, the ETC probability of driver v is P_v^+, and the MTC one is P_v^- where $P_v^+ + P_v^- = 1$. Here, V is the set of drivers appeared in highway where any $v \in V$ could be distinguished by the license plate of a vehicle.

Accordingly, drift detection for trend discovery can be defined as follows.

Definition 2: Drift detection of payment identity. Based on Definition 1, the drift detection of payment identity for a driver v in month m contains two facts according to a predefined threshold h. (1) If $P_v^+ - P_v^- \geq h$, a potential positive drift into ETC identity appears for a MTC driver v; (2) If $P_v^+ - P_v^- \leq -h$, a potential negative drift into MTC emerges for a ETC driver v. Moreover, due to the fact $P_v^+ + P_v^- = 1$, such drifts are simplified only with P_v^+: $P_v^+ \geq (1+h)/2$ implies positive drift and $P_v^+ \leq (1-h)/2$ means negative drift.

The overview of our method for trend discovery is illustrated as Figure 1, where four main parts are included. (1) **Data management** layer loads required data. Raw online *toll records* are received continuously through a dedicated message service, and then accumulated as *historical tolls* into NoSQL database. Necessary data cleaning and aggregative calculation are completed as pre-processing, which can be found in our previous works [9, 35]. Business basic data, such as highway station, section, line and region, has been imported into a relational database. (2) On spatio-temporal toll data above, the feature of travel characteristics is modeled and organized in **feature management** module, which would be discussed respectively in Section 4.1 and 4.2. (3) With the features, **trend management** module adopts GBDT (Gradient Boost Decision Tree) technology to train an ensemble-learning model after algorithmic parameters tuning by cross-validation. Monthly trend of drivers' payment identity would be calculated by the trained model, written into the relational database, and employed for drift detection after sorting and visualization. (4) The visualized results would be presented in multiple online applications of **business application** layer through dedicated API (application programming interface).

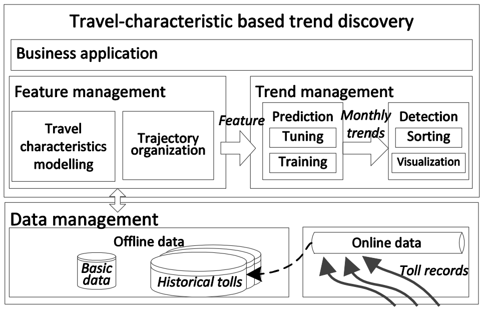


Figure 1: Overview of our method.

In practice, our method as a routine analysis would run at 12:00 a.m. of 1st day in month *m* on the data of month *m*-1, and output drivers' trends. The drift of payment identity of month *m* can be detected then. In fact, although discussed in a specific highway domain, our method would be general to find attributes' drift (e.g., payment identity here) on spatio-temporal data (e.g., toll data here) in many other analogous fields.

4 Trend discovery by classification prediction

4.1 Feature modelling

In the framework as Figure 1, the trend of payment identity for individual drivers is found by classification prediction. Travel characteristics of driver are key ingredient for predictive estimation and their feature has to be modelled properly. Our consideration is based on the following business statistical observations.

Observation 3: Compared with MTC ones, ETC drivers consume more time in highway and own longer accumulative mileage in one month.

Observation 4: Compared with MTC ones, ETC drivers have common travel preferences. In January 2018, 38% of ETC drivers passed the toll stations surrounding provincial city *Zhengzhou*; nearly 10% of them traveled along the line *Zhengzhou Southwest Circle*.

Observation 5: The drivers of private cars more likely belongs to ETC identity. According to seat size in a vehicle, five vehicle-model types exist for passenger-car; according to the weight capacity of a vehicle, four vehicle-model types are for freight-carriage. In January 2018, more than 72% of ETC vehicles lie in 1st passenger-car model, which is dominant among all the nine types.

Considering observations above, the travel characteristics of highway drivers can be defined as follows.

Definition 3: Feature of highway drivers' travel characteristics. In a month m, the feature of travel characteristics of drivers V is a vector $\{(X_v^m, Y_v^m) | v \in V\}$. For a driver $v \in V$, $Y_v^m \in \{-1, 1\}$ presents his payment identity: $Y_v^m = 1$ implies ETC; $Y_v^m = -1$ implies MTC. $X_v^m = (t_v^m, s_v^m, o_v^m, l_v^m, p_v^m)$ as historical characteristics of v includes five dimensions: t_v^m as a temporal dimension is his trip times; s_v^m as a spatial dimension is his accumulative mileage; $o_v^m \in \{1..9\}$ as an objective attribute dimension is the vehicle-model type; $l_v^m = (0..1)$ as a subjective preference dimension is a proportion that trips of v involving certain major cities S; $p_v^m = (0..1)$ as another subjective preference dimension is a proportion that top-K OD (Origin-Destination) patterns in v's trips, $\mathrm{K} \in \mathbb{Z}^+$.

The dimensions are illustrated in details below.

(1) The first two dimensions $t_v^m$and s_v^m fit the spatio-temporal influences in Observation 3. Both imply the travel frequency of a driver in highway. The third dimension o_v^m refers the objective factor in

Observation 5. It implies the types of vehicle owned by a driver: the label-encoding values 1~5 of o_v^m are passenger-car models; the others 6~9 are for freight-carriage model. All of those three dimensions are straight forward from attributes of toll data in Table 1.

(2) The fourth dimension $0 \leq l_v^m \leq 100\%$ reflects subjective preference about hot locations in Observation 4. A driver's trips involve a city c if they passed any toll station belonging to c. Such involvement can be extracted from either *entry_station* or *exit_station* of the toll data in Table 1, because any toll station in China belongs to a certain prefecture-level city. Around flourishing major cities, more travels appear, higher probability of traffic congestions exist, and larger requirements of ETC emerge for highway drivers to improve their travel efficiency. In *Henan* highway discussed in Section 3.1, the set of major cities *S* only contains the pivot *Zhengzhou* according to extensive domain surveys.

(3) The fifth dimension $0 \leq p_v^m \leq 100\%$ reflects subjective preference about habitual travel patterns in Observation 4. OD is a technical term about travel demand in transportation related research [12, 36]. As a pair <*origin*, *destination*>, the OD of a trip can be extracted from the toll data in Table 1: *origin* is from *entry_station*, and *destination* is from *exit_station*. For a certain driver, top-K OD reveals his routine travels, where $1 \leq \mathrm{K} \leq 3$ in common. More such routines are in his trips, more significant the highway would be in his urban life, and more likely he would to be an ETC driver. In fact, a driver commuting every workday has higher probability in ETC identity than the ones owning casual long journeys only on holidays.

4.2 Feature organization through driver trajectory

The feature of travel characteristics would be built monthly on massive historical data for any individual highway driver, but the huge size of drivers brings managerial difficulty and retrieval inefficiency. Accordingly, we propose a data structure *driver trajectory* to rebuild original data for feature management.

Definition 4: Driver trajectory. A driver trajectory TR_v^m is a link structure to organize the feature of travel characteristics for a driver *v* in month *m*, which contains two types of components: head and node. In a driver trajectory TR_v^m, a single *head*$= < v, m\ , Y_v^m, X_v^m >$ keeps aggregative feature as Definition 3, and each of nodes represents a trip of *v*. Here, a node=< l_e, l_x, t_e, t_x >, where its component is the attributes *entry_station*, *exit_station*, *entry_time*, *exit_time* in order from toll data of Table 1.

In order to build such a data structure for individual drivers, we design a distributed computing procedure on massive toll data.

The trajectory building is an offline processing depicted as a MapReduce job in Figure 2. The map phase in the left part parses each record of toll data, and exacts required attributes in a trip. Then, it counts the mileage of this trip Δs_r by the shortest cartographic distance instead of the direct Euler length. The reduce phase in the right part groups intermediate results by a composite key including vehicle license and month, and counts mileage summary, trip time, OD proportion, city proportion, and Y_v^m.

During the building procedure, feature calculation is done with these following details. (1) On monthly historical toll data in NoSQL storage as input, the procedure would scan data only once to build features for all the drivers, and output them as trajectories into dedicated table of the same NoSQL storage. Here, column-based NoSQL like HBase is proper to maintain trajectories where the head and each node of a trajectory are respective columns. It makes the access of travel characteristics efficient because only the column of trajectories' heads is adequate. (2) Spatio-temporal dimensions of travel characteristic are completed by cumulative addition. In reduce phase, t_v^m is current count plus one; s_v^m is current summary plus a trip mileage Δs_r achieved in map phase. (3) Subjective preference dimensions of travel characteristics are calculated by division operation. For the correctness of this operation, the parameter "mapred.reduce.slowstart.completed.maps" must be set as 1 in Hadoop to get complete intermediate results before any reduce phase. To achieve l_v^m and p_v^m in the reduce phase for any driver, the denominator is an intermediate accumulative variable, and the numerator comes from current status in an iteration. (4) Objective attribute dimension of travel characteristics and Y_v^m are assigned by the following rule. For a driver *v* in month *m*, if vehicle-model *r.o* (or payment identity *r.y*) in his trips is invariant, o_v^m (or Y_v^m) in feature is the very value; otherwise, it would be set as the value in his last trip. For the latter case, this trajectory should not be employed for model training in Section 4.3 due to this alteration, and a dedicated flag variable is labeled. (5) Because the components of X_v^m have different ranges, normalization is necessary for X_v^m to reduce predictive error of machine-learning models. Z-score [37] is adopted before the prediction in Section 4.3, which is a widely used for dimensionless quantity in mathematical statistics, and would not be explained further due to its straight forward manner.

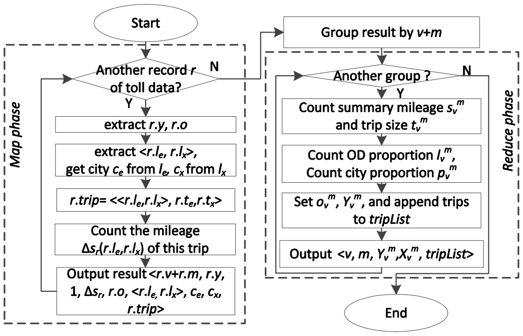

Start
Map phase
Another record r of toll data?
N
Y
extract $r.y$, $r.o$
extract $<r.l_e, r.l_x>$, get city c_e from l_e, c_x from l_x
$r.trip$= $<<r.l_e,r.l_x>, r.t_e,r.t_x>$
Count the mileage $\Delta s_r(r.l_e,r.l_x)$ of this trip
Output result $<r.v+r.m$, $r.y$, 1, Δs_r, $r.o$, $<r.l_e, r.l_x>$, c_e, c_x, $r.trip>$
Group result by $v+m$
Reduce phase
Another group ?
N
Y
Count summary mileage s_v^m and trip size t_v^m
Count OD proportion l_v^m, Count city proportion p_v^m
Set o_v^m, Y_v^m, and append trips to $tripList$
Output $<v, m, Y_v^m, X_v^m, tripList>$
End

Figure 2: Trajectory building with feature calculation.

4.3 Classification prediction and drift detection

To discover potential trend through travel characteristics from driver trajectories, a classification model based on GBDT (Gradient Boosting Decision Tree) is designed here. GBDT can combine weak models into a stronger one in iterative stages to improve predictive accuracy with an arbitrary loss function.

According to business habits for evaluation, logistic regression is adopted as loss function *Lf*. The goal is to find a model *F* by minimizing logistic error $\sum_{v \in V} \log\left(1 + \exp\left(-Y_v^m F(X_v^m)\right)\right)$, for input $(X_v^m, Y_v^m), v \in V$. At each stage *d*, 1⩽*d*⩽*D*, a weak model F_d would fit $h_d(x)$ to previous residuals by gradient boosting. That is, each F_d attempts to correct the errors of its predecessor model F_{d-1}. Therefore, an ensemble-learning model can be described as the optimization problem as follows.

$$\min_F \sum_{v \in V} Lf\left(y_v, F_{d-1}(x_v) + h_d(x_v; \varphi_d)\right) + \Omega(h_d)$$

$$\text{s.t.} \quad F_d(X) = F_{d-1}(X) + h_d(X; \varphi_d)$$

$$F_D(X) = \sum_{d=1}^{D} h_d(X; \varphi_d)$$

As Definition 3, $x_v \in X$ is X_v^m, y_v is Y_v^m. Here, φ_d is a parameter of the weak model F_d in d^{th} iteration, and F_D is the final ensemble model. Referring the concepts in XGBoost [38], we extend the optimized goal as the loss function *Lf* with $\Omega(h_d(x)) \sim (F_d\_\text{depth}, F_d\_\text{shrinkage})$. The appended $\Omega(h_d(x))$ is a regularization part restricted by base trees' depth and shrinkage rate: tree depth controls model complexity (i.e., the degree of model can fit); shrinkage rate is a small extent to slow down the re-enforce of generating a new base tree.

Therefore, the model of classification prediction for drivers' payment identity can be trained as the procedure in Table 2 with multiple algorithmic parameters: tree size (i.e., iterative number) *D*, maximal tree depth *H*, tree shrinkage *r*, training ratio η, and drivers' trajectories $TR_v^{m-1}, v \in V$. Here, *D*, $H \in \mathbb{Z}^+$, $r \in \mathbb{R}^+$, and $0 \leq \eta \in \mathbb{R}^+ \leq 1$.

Table 2: Model training to find trends of drivers' payment identity.

```
Algorithm: classification model training through GBDT
Input: trajectories $TR^{m-1}$ of all the drivers V, tree size D, maximal tree depth H, tree shrinkage r, and training ratio $\eta$.
Output: an ensemble model $F_D$ to get payment identity trends of highway drivers in month m.
randomly select $\eta * |V|$ ones from $TR^{m-1}$ as $TR_v^{m-1}, v \in V', |V'| = \eta * |V|$;
request feature of travel characteristics $(x_v, y_v)$ from selected trajectories $TR_v^{m-1}, v \in V'$.
initialize a model with constant values: $F_0(x) = \arg\min_\gamma \sum_{v=1}^{|V'|} Lf(y_v, \gamma)$ ;
for d=1 to D
  for i=1 to |V'|
6.          compute residuals: $r_i^d = -\left[\frac{\partial Lf(y_v, F(x_v))}{\partial F(x_v)}\right]_{F(X)=F_{d-1}(X)}$;
7.      endfor
8.      find a base tree $h_d(X)$ to fit those residuals on $\{(x_v, r_i^d)\}_{v=1}^{|V'|}$ ;
9.      get weight $\gamma_d$ by one-dimensional optimization: $\gamma_d = \arg\min_\gamma \sum_{v=1}^{|V'|} Lf(y_v, F_{d-1}(x_v) + \gamma h_d(x_v)) + \Omega(h_d)$ ;
10.     update the model: $F_d(X) = F_{d-1}(X) + \gamma_d h_d(X)$;
11.  endfor
12.  return $F_D(X)$
```

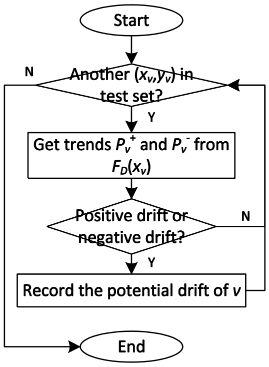

Start
Another (x_v,y_v) in test set?
N
Y
Get trends P_v^+ and P_v^- from $F_D(x_v)$
Positive drift or negative drift?
N
Y
Record the potential drift of v
End

Figure 3: Drift detection procedure.

With driver trajectories in month m-1 and multiple parameters as input, the algorithm would output an ensemble model to get payment identity trend for any individual highway driver in this coming month m. According to the training ratio parameter η, training set is selected randomly from input trajectories in Line 1. Note that, only qualified trajectories with given labeled flag would be selected, which is discussed in Section 4.2. Considering space cost, the selection action here only records the indexes in NoSQL storage instead of actual trajectories. The feature of travel characteristics is acquired by retrieval index of the training set as Line 2. To initiate the model, $F_0(X)$ is generated with constants γ as Line 3. After computing residuals via negative gradient direction in Lines 5-7, a base tree $h_d(X)$ as a weak model is built to fit x_v with those residuals r_i^d in each iteration as Line 8. To find an approximation that minimizes average value of loss function Lf and regularization Ω, weight γ_d is got in Line 9. Here, the regularization of base trees is controlled by parameters H and r. The model then incrementally expands itself by adding new weighted tree $\gamma_d h_d(X)$ as Line 10. At last, the final model $F_D(X)$ as an ensemble of D base trees (i.e., in D iterations) is returned like Line 12.

Through the learned model F_D, the drift detection of payment identity can be depicted as the procedure in Figure 3. The test set is all the drivers' trajectories in month m. With the feature (x_v, y_v) of a driver v in test set, the trend as a probability pair of Definition 1 can be achieved through $F_D(x_v)$. The drift is found by evaluating P_v^+ with given threshold h according to Definition 2. That is, when v is ETC identity, y_v=1, $(P^+ - P^-) \leq -h$ implies v's negative drift; when v is MTC identity, y_v=-1, $(P^+ - P^-) \geq h$ implies v's positive drift. Potential trend for all the individual drivers would be discovered, which could be employed for further usage in domain applications.

5 Evaluation

In the practical scene mentioned in Section 3.1, we conduct three experiments and two case studies to evaluate our method quantitatively and qualitatively.

5.1 Setting

Five Acer AR580 F2 rack servers via Citrix XenServer 6.2 are utilized to build a private Cloud, each of which own 8 processors (Intel Xeon E5-4607 2.20GHz), 64 GB RAM and 80 TB storage. To maintain historical toll data in data management layer as Figure 1, three virtual machines of the Cloud form a HBase 1.6.0 cluster, each of which owns 4 cores CPU, 22 GB RAM and 700 GB storage. Toll data from more than

300 toll stations in Henan highway is used, which was generated as the speed of 1 million records per day. Another one of virtual machines (4 cores CPU, 8 GB RAM and 200 GB storage installing CentOS 6.6 x86_64 operating system) is used to install MySQL 5.6.17 as a relational database for business basic data (station, section and highway line). The feature management module, trend management module and application layer are also implemented on that machine with Oracle JDK 1.7.0, Apache Tomcat 7.0.103, and scikit-learn 0.21.3.

As a routine analysis running at 12:00 a.m. of 1st day in coming month *m*, the classification prediction model would be re-trained, and the trend of drivers would be achieved. All those results would be written to a dedicated table of HBase.

5.2 Experiment for feature management and trend prediction

With such given configured settings, three experiments are implemented for quantitative evaluation. The first one is to estimate travel characteristic management, the second is to tune proper algorithmic parameters, and the last is to compare predictive effects with other models in four perspectives.

We first evaluate trajectory building and feature retrieval of feature management module.

Experiment 1: Feature management. Among toll data in January 2016, the data of certain days is appended to the input in each test, and the executive times to build driver trajectories are noted. Then, the average executive time on fixed one million records in each test can be deduced. Moreover, the time to access feature by our method (abbr. trajectory) is counted and compared with traditional methods to query independent characteristics (abbr. trip, mileage, location%, and OD% respectively). The access time of our method is the summary of executive time to build trajectories and query time for travel characteristics from trajectories; while that of others are independent query times on NoSQL storage.

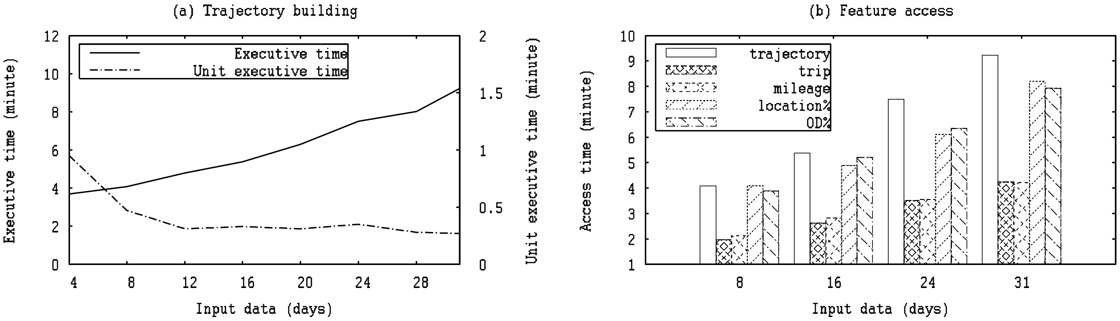


Figure 4: Feature building and access.

The results are presented in Figure 4. (1) Our method for trajectories building has well horizontal scalability in two perspectives of Figure 4 (a). When input scales, the increments of executive time excel the linearity. That time holds minute-level and not doubled even when the input size grows nearly eight folds. It can also be demonstrated clearly in unit executive time with the second vertical axis, where average executive time on one million records declines to steady one-quarter minute. The result shows that processing capacity of our method is stable and horizontally scalable. (2) For feature access, our method performs much better than traditional ways as shown in Figure 4 (b). All the characteristics except vehicle-model in Definition 3 are considered in this experiment, because objective attribute can be found by simple direct query. In our method through driver trajectories, travel characteristics can be gained at a time, which have been maintained in a data structure for efficient query. While in traditional ways, these characteristics have to be queried independently from NoSQL storage. The access time of our method includes trajectory building time and query time, in which the former one is dominant due to its much I/O cost. Accordingly as Figure 4 (b), when volume grows, the access time of our method presents similar tendency as Figure 4 (a), while those of others increase in distinct extents due to their different nature of aggregative operations. Furthermore, on the data of the same volume, the access time of our method is much shorter than the sum of that of others, which proves driver trajectory low latency to access features.

Therefore, our method is efficient for feature management during trajectory building and feature access. Then, before the evaluation of prediction effects, three common metrics are employed.

Definition 4: Accuracy metrics for classification prediction. According to confusion matrix [39] as true positive (*TP*), false positive (*FP*), true negative (*TN*) and false negative (*FN*), *precision* is defined as Equation 1, *recall* is defined as Equation 2, and *F1-score* is defined as Equation 3. Here, with a coefficient β, the F1-score is a harmonic mean of precision and recall, whose value is between 0 and 1. To equally weight precision and recall here, we set $\beta = 1$ constantly.

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

$$\text{F1} - \text{score} = \frac{(1+\beta^2)*\text{precision}*\text{recall}}{(\beta^2)*\text{precision}+\text{recall}} \quad (3)$$

The tuning for algorithmic parameters is discussed then. As the design of Section 4.3, the classification prediction model would be trained with proper algorithmic parameters: tree size (i.e., iterative number) *D*, maximal tree depth *H*, tree shrinkage rate (i.e., learning rate) *r*, and training ratio η. Here, *D*, $H \in \mathbb{Z}^+$, $r \in \mathbb{R}^+$, and $0 \leq \eta \in \mathbb{R}^+ \leq 1$. In practice, the model has to be re-trained periodically (e.g., once a month) to fit recent data better. Depth parameter *H* has upper bound 6 [40], and is set median 3 constantly here to trade-off model complexity and tree-structure split efficiency. Moreover, we set $\eta = 50\%$ by default due to business habit. Accordingly, an experiment is designed to find the optimal combination of algorithmic parameters *D* and *r*.

Experiment 2: Parameters tuning. The toll data in December 2015 is employed to build driver trajectories. Among those trajectories, training ratio $\eta = 50\%$ makes half of them as training set and the others would be validation set. When *H*=3 as mentioned before, we want to find the optimal pair of *D* and *r* for the model. Tree size parameter *D* is set as 50, 100, 500, 1000, 2000 and 3000 respectively, and *r* is set as 0.005, 0.01, 0.1 and 0.5 then. Under each combination of *D* and *r* on training set, three metrics of Definition 4 would be counted in average on validation set.

The results in three perspectives are presented in Figure 5, and some conclusions can be drawn from it. First, from Figure 5 (a) and (b), metric precision and recall are in negative correlation. When *D* grows with the same *r*, metric precision increases slowly, but metric recall decreases. It can be interpreted from the inherent semantic of Definition 4. For parameter *r*, smaller one performs better in metric recall, while few differences appear in metric precision. Second, from Figure 5 (c), the harmonic mean F1-score shows similar form as metric recall. It is the reason that metric recall presents more violent fluctuations than precision. Moreover, due to the nature of ensemble learning, too large *D* requiring much training time would lead over-fitting. Accordingly, under comprehensive consideration, the optimal combination of *D*=2000 and *r*=0.005 is chosen for our model. It just reflects the trade-off predictive effects among multiple influential factors.

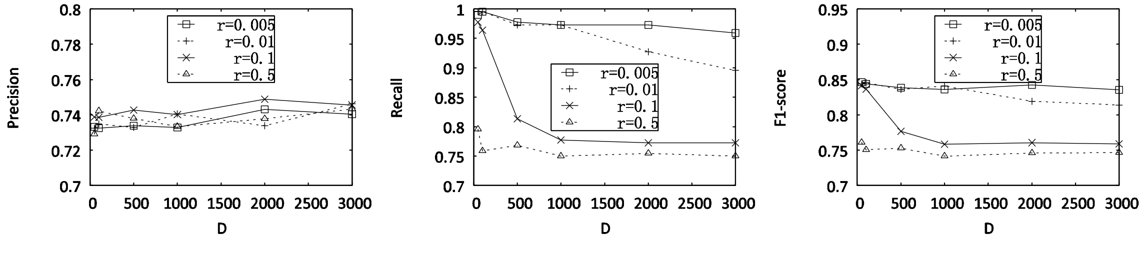


Figure 5: Parameter tuning.

Then, through the trained model above (i.e., *D*=2000, *H*=3, *r*=0.005), another experiment is conducted to compare the results of trend discovery. In order to quantitatively evaluate classification effects of our model (abbr., GBDT), three other models are also implemented in trend management module of Figure 1. They are non-linear model SVM (Support-Vector Machine, kernel=’*rbf*’ here), non-parametric model KNN

(K- Nearest Neighbor, K=3 here) and binary Logistic model. All those are tuned respectively on the same training set and validation set as Experiment 2. The following experiment is to compare our work with these counterparts in accuracy metrics and executive times.

Experiment 3: Predictive effects. The driver trajectories of 12 months in 2016 are employed to evaluate prediction effect. There are about 2.5 million trajectories (i.e., driver size) within a month in average. The model of our method is the one trained by Experiment 2. Through four comparative models to predict drivers' payment identity, three metrics in Definition 4 and executive time are counted after the finish of classification prediction.

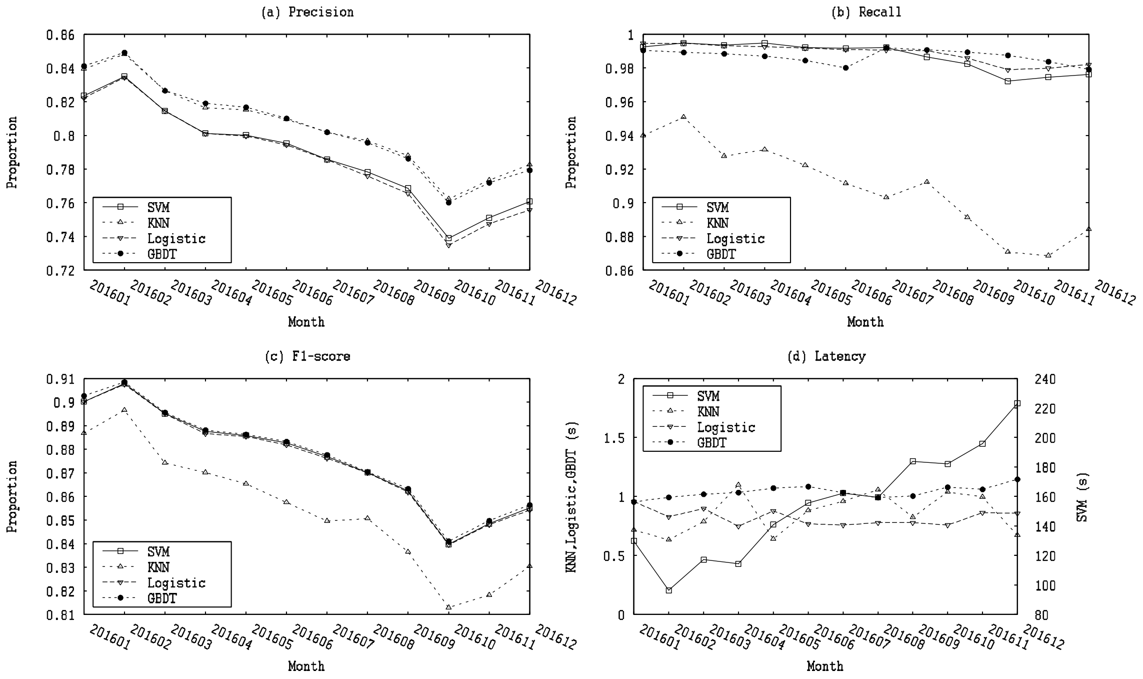


Figure 6: Benchmarking on predictive effects.

Predictive effects are illustrated as Figure 6. The first threes show dimensionless metrics and the fourth one exhibits executive time with the unit of second. We found our method performs well in accuracy metrics with relatively low executive time. (1) All four models can depict trend of 12 months in one year, and our GBDT has the advantage in precision. Among these fours, the precisions are larger than 70%, the recalls are around 80%, and F1-scores are more than 84%. With the worst here in recall and F1-score, KNN is still sufficient enough in practice. High accuracy metrics are owed to the travel characteristics we defined which include multiple spatio-temporal dimensions. The proposed feature is proved its practical feasibility. GBDT performs the best in metric precision and second-best in metric F1-score, which comes from the predictive capacity through ensemble-learning. The excellent predictive effect of our model is proved either. (2) Our GBDT has evident advantages in executive latency. It consumes one second steadily due to fast convergence of the algorithm in Table 2. Logistic appears a little better with more fluctuations in time but has the worst metric precision. Although KNN costs the least time sometimes, but achieves the worst metric recall and metric F1-score. Note that, the executive time of SVM is much longer by an order of magnitude than others, and it is measured in secondary vertical coordinates as Figure 6(d). (3) From the results in one year through four models, we found common facts among different perspectives. (i) In views of precision and F1-socre, a peak emerges in February and a valley appears in October. In both months, a 7-day holiday exists when vehicles in highway would be free of charge due to national regulations. It makes vague payment identity of drivers and would confuse characteristics building. (ii) In views of precision and F1-socre, the accuracy drops roughly when month elapses. It is interpretable because we focus on the comparison among different models and re-training is absent for these models here. In fact, as we have mentioned in Section 3.2, our method would update the classification model once a month to fit the recent data better, and would perform more accurately than these results. However, even in the valley of the results, our method still holds metric precision larger than 76% and metric F1-score larger than 84%. It also comes from the travel characteristics we defined. High accuracy and practical feasibility are proved again. In summary, our method performs well in executive performance and holds high predictive accuracy.

5.3 Case studies for trend discovery

Then, two case studies are introduced to qualitatively evaluate our method for practical usage in real-world scenario. With uncovered drivers' trend, one is an individual profiling application for any specific driver, and the other is a group profiling application for entire drivers.

Case 1: Individual driver profiling. With threshold *h* in Definition 2 as 30%, Figure 7 is the individual profiling application of our system in January 2017 for a driver whose vehicle license is "豫 A6XXXX". Two main parts are included: the top is a tag-based user model and the bottom is the charts of statistics and predictive trend.

Some explanations should be addressed. (1) Among tag-based user model in the top part of Figure 7, three types of tags exist here for a highway driver. The first is *factual tag* which can be extracted from data directly. The vehicle license, vehicle-model and payment identity are such tags from attributes of toll data. The second is *statistic tag* which is achieved from aggregative calculations on data. The monthly trip time, mileage and typical OD pattern proportion are such tags. The third is *predictive tag* which is built by predictive calculation. For example, the tag "potential ETC churn" is generated by our method which implies the driver may change his payment identity from ETC to MTC in coming month. Another tag "probable locale" is built by a clustering prediction which infers this driver's resident city is *Zhengzhou*. (2) More information can be drilled down in the bottom part of Figure 7. The "hot time" as funnel chart and "hot station" as pie chart are temporal and spatial preference of the driver, which are achieved by statistical aggregations. The "payment identity trend" is the result from our method, in which the probability-pair supports a potential negative drift of this driver for his payment identity. (3) Business strategies would be followed up by highway commercial department on these results. After a potential negative drift is recognized, this driver may be persuaded to hold his ETC identity by some customized preferential policies, such as discount coupons about refueling.

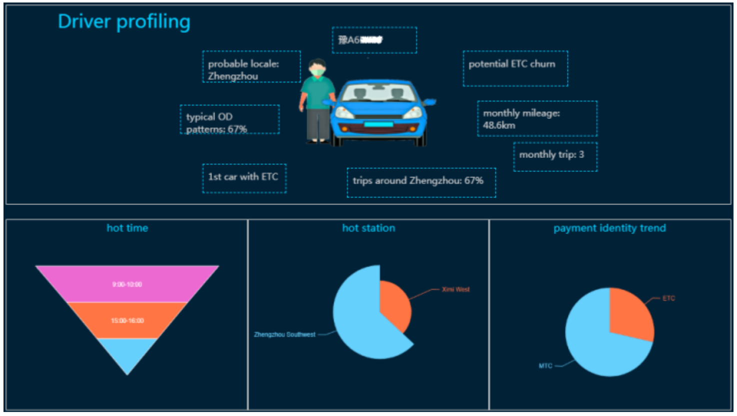
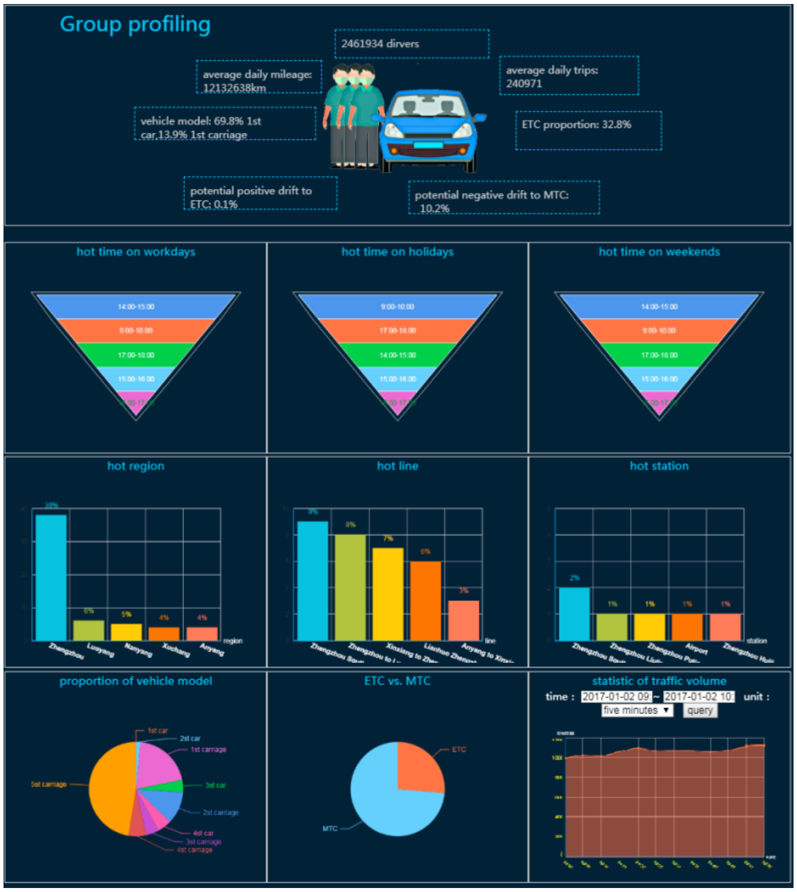


Figure 7: Individual driver profiling.

In summary, the trend of payment identity has been well employed in an online application for individual highway driver. Such trend can be intuitively employed for further business strategies.

Case 2: Group driver profiling. With threshold *h* in Definition 2 as 30%, Figure 8 is the group driver profiling application of our system in January 2017. For this "virtual" driver, it represents the entire highway drivers in Henan (i.e., universal-set group). Two main parts are included either: the top is a tag-based user model and the bottom is various charts about statistics and predictive trend.



Group profiling

2461934 dirvers

average daily mileage: 12132638km

average daily trips: 240971

vehicle model: 69.8% 1st car,13.9% 1st carriage

ETC proportion: 32.8%

potential positive drift to ETC: 0.1%

potential negative drift to MTC: 10.2%

time : 2017-01-02 09 ~ 2017-01-02 10 unit : five minutes ▾ query

Figure 8: Group driver profiling.

We address some key points in details. (1) Among tag-based user model in the top part of Figure 8, statistic tags and predictive tags are presented for this group. Generated from aggregative calculations, the statistic tags show driver size, average daily trip, average daily mileage, ETC proportion, and the proportion of vehicle-models in passenger-cars (and freight-carriages). Two predictive tags "potential positive drift to ETC" and "potential negative drift to MTC" are built by our method which implies the proportion among all the drivers who would change his payment identity in coming month. (2) More information can be drilled either in the bottom part of Figure 8. In views of funnel charts, hot times on workdays / weekends / holidays as temporal preference are presented. In views of bar charts, hot regions / lines / stations as spatial preference are provided. In views of pie charts, vehicle-model proportion and current ETC vs. MTC are expressed. In the last view of line chart, traffic volume trend in the next day are showed in multiple temporal granularities, which come from corresponding predictive calculations in our previous work [9]. (3) On these results, business strategies would also be made by highway commercial department. When negative drift or positive drift bursts, domain officers would rethink their current management policies and prompt strategic reform, such as revising ETC service quality.

In summary, the trend of payment identity is well adopted in an online application for a group driver. Such trend can be studied further by extensive business strategies.

6 Conclusion

In this paper, a novel method is proposed to discover highway drivers' trend. On massive historical spatio-temporal data, the feature of travel characteristics is built efficiently through driver trajectories, considering time, space, objective facts, and subjective preferences of highway drivers. Through gradient boosting classification, our model is proved low executive latency with well-performed accuracy than traditional ones. In extensive experiments and case studies on real data of one Chinese province, the latency is around 1 second, metric precision is near 75%, metric recall is around 95%, and metric F1-score is more than 85%. Potential trend of highway driver's payment identity has been well adopted in business applications with intuitive benefits.

Due to confusing recognition of payment identity on some special holidays, trend discovery for highway drivers in those periods still faces a big challenge. In our future work, other heterogeneous data and fine-grained characteristics are planned to employ to improve prediction effects.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 61702014), Beijing Municipal Natural Science Foundation (No. 4192020), Top Young Innovative Talents of North China University of Technology (No. XN018022), and "Yuyou" Talents of North China University of Technology (No. XN115013).

REFERENCES

1. Laña, I., Lobo, J.L., Capecci, E., Del Ser, J., Kasabov, N.: Adaptive long-term traffic state estimation with evolving spiking neural networks. Transportation Research Part C: Emerging Technologies 101, 126-144 (2019)
2. Yang, X., Zhou, S., Cao, M.: An Approach to Alleviate the Sparsity Problem of Hybrid Collaborative Filtering Based Recommendations: The Product-Attribute Perspective from User Reviews. Mobile Networks and Applications 25, 376-390 (2020)
3. Gao, H., Liu, C., Li, Y., Yang, X.: V2VR: Reliable Hybrid-Network-Oriented V2V Data Transmission and Routing Considering RSUs and Connectivity Probability. IEEE Transactions on Intelligent Transportation Systems 1-14 (2020)
4. Gao, H., Qin, X., Barroso, R.J.D., Hussain, W., Xu, Y., Yin, Y.: Collaborative Learning-Based Industrial IoT API Recommendation for Software-Defined Devices: The Implicit Knowledge Discovery Perspective. IEEE Transactions on Emerging Topics in Computational Intelligence 1-11 (2020)
5. Mannering, F.: Temporal instability and the analysis of highway accident data. Analytic Methods in Accident Research 17, 1-13 (2018)
6. Curry, A.E., Kim, K.H., Pfeiffer, M.R.: Inaccuracy of Federal Highway Administration's Licensed Driver Data: Implications on Young Driver Trends. Journal of Adolescent Health 55, 452-454 (2014)
7. Zhu, F., Lv, Y., Chen, Y., Wang, X., Xiong, G., Wang, F.: Parallel Transportation Systems: Toward IoT-Enabled Smart Urban Traffic Control and Management. IEEE Transactions on Intelligent Transportation Systems 1-9 (2019)
8. Park, J., Iagnemma, K., Reimer, B.: A User Study of Semi-Autonomous and Autonomous Highway Driving: An Interactive Simulation Study. IEEE Pervasive Computing 18, 49-58 (2019)
9. Ding, W., Wang, X., Zhao, Z.: CO-STAR: A collaborative prediction service for short-term trends on continuous spatio-temporal data. Future Generation Computer Systems 102, 481-493 (2020)
10. Ding, W., Zhao, Z.: DS-Harmonizer: A Harmonization Service on Spatio-Temporal Data Stream in Edge Computing Environment. Wireless Communications and Mobile Computing 2018, 12 (2018)
11. Kolajo, T., Daramola, O., Adebiyi, A.: Big data stream analysis: a systematic literature review. Journal of Big Data 6, 47 (2019)
12. Zhu, L., Yu, F.R., Wang, Y., Ning, B., Tang, T.: Big Data Analytics in Intelligent Transportation Systems: A Survey. IEEE Transactions on Intelligent Transportation Systems 20, 383-398 (2019)
13. Zhao, W.X., Li, S., He, Y., Wang, L., Wen, J.-R., Li, X.: Exploring demographic information in social media for product recommendation. Knowledge and Information Systems 49, 61-89 (2016)
14. Ding, W., Wang, Z., Zhao, Z.: A Platform Service for Passenger Volume Analysis on Massive Smart Carad Data in Public Transportation Domain. In: 15th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2019), pp. 681-697. Springer International Publishing, Cham, (Year)
15. Chen, J., Liu, Y., Zou, M.: Home location profiling for users in social media. Information & Management 53, 135-143 (2016)

16. Gao, H., Kuang, L., Yin, Y., Guo, B., Dou, K.: Mining consuming Behaviors with Temporal Evolution for Personalized Recommendation in Mobile Marketing Apps. Mobile Networks and Applications 25, 1233-1248 (2020)
17. He, X., Zhang, H., Kan, M.-Y., Chua, T.-S.: Fast Matrix Factorization for Online Recommendation with Implicit Feedback. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 549–558. Association for Computing Machinery, Pisa, Italy (2016)
18. Wang, P., Fu, Y., Xiong, H., Li, X.: Adversarial Substructured Representation Learning for Mobile User Profiling. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 130-138. ACM, Anchorage, AK, USA (2019)
19. Song, X., Ming, Z.-Y., Nie, L., Zhao, Y.-L., Chua, T.-S.: Volunteerism Tendency Prediction via Harvesting Multiple Social Networks. ACM Trans. Inf. Syst. 34, Article 10 (2016)
20. Liang, S., Zhang, X., Ren, Z., Kanoulas, E.: Dynamic Embeddings for User Profiling in Twitter. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1764–1773. Association for Computing Machinery, London, United Kingdom (2018)
21. Eke, C.I., Norman, A.A., Shuib, L., Nweke, H.F.: A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions. IEEE Access 7, 144907-144924 (2019)
22. Liang, Y., Jiang, Z., Zheng, Y.: Inferring Traffic Cascading Patterns. Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 1-10. ACM, Redondo Beach, CA, USA (2017)
23. Wang, P., Fu, Y., Zhang, J., Wang, P., Zheng, Y., Aggarwal, C.: You Are How You Drive: Peer and Temporal-Aware Representation Learning for Driving Behavior Analysis. 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2018), pp. 2457-2466. ACM, London, United Kingdom (2018)
24. Hu, H., Kantardzic, M., Sethi, T.S.: No Free Lunch Theorem for concept drift detection in streaming data classification: A review. WIREs Data Mining and Knowledge Discovery 10, e1327 (2020)
25. Zhang, W., Wang, J.: A Hybrid Learning Framework for Imbalanced Stream Classification. In: IEEE International Congress on Big Data (BigData Congress 2017), pp. 480-487. IEEE, (Year)
26. Gao, J., Fan, W., Han, J.: On Appropriate Assumptions to Mine Data Streams: Analysis and Practice. In: Seventh IEEE International Conference on Data Mining (ICDM 2007), pp. 143-152. (Year)
27. Yu, S., Abraham, Z.: Concept Drift Detection with Hierarchical Hypothesis Testing. In: Proceedings of the 2017 SIAM International Conference on Data Mining, pp. 768-776. American Statistical Association., (Year)
28. Cabral, D.R.d.L., Barros, R.S.M.d.: Concept drift detection based on Fisher’s Exact test. Information Sciences 442-443, 220-234 (2018)
29. Wang, X., Kang, Q., Zhou, M., Yao, S.: A Multiscale Concept Drift Detection Method for Learning from Data Streams. In: 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE), pp. 786-790. (Year)
30. Sidhu, P., Bhatia, M.P.S.: A novel online ensemble approach to handle concept drifting data streams: diversified dynamic weighted majority. International Journal of Machine Learning and Cybernetics 9, 37-61 (2018)
31. Khamassi, I., Sayed-Mouchaweh, M., Hammami, M., Ghédira, K.: A new combination of diversity techniques in ensemble classifiers for handling complex concept drift. Learning from data streams in evolving environments, pp. 39-61. Springer (2019)
32. Nikzad-Langerodi, R., Lughofer, E., Cernuda, C., Reischer, T., Kantner, W., Pawliczek, M., Brandstetter, M.: Calibration model maintenance in melamine resin production: Integrating drift detection, smart sample selection and model adaptation. Analytica Chimica Acta 1013, 1-12 (2018)
33. Žliobaitė, I., Pechenizkiy, M., Gama, J.: An overview of concept drift applications. Big data analysis: new algorithms for a new society, pp. 91-114. Springer (2016)
34. Moreira-Matias, L., Gama, J., Mendes-Moreira, J.: Concept Neurons – Handling Drift Issues for Real-Time Industrial Data Mining. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases(ECML PKDD 2016), pp. 96-111. Springer International Publishing, (Year)
35. Xia, Y., Wang, X., Ding, W.: A Data Cleaning Service on Massive Spatio-Temporal Data in Highway Domain. In: Service-Oriented Computing – ICSOC 2018 Workshops, pp. 229-240. Springer International Publishing, (Year)

36. Wang, S., Li, L., Ma, W., Chen, X.: Trajectory analysis for on-demand services: A survey focusing on spatial-temporal demand and supply patterns. Transportation Research Part C: Emerging Technologies 108, 74-99 (2019)
37. https://en.wikipedia.org/wiki/Standard_score
38. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. ACM, San Francisco, California, USA (2016)
39. https://en.wikipedia.org/wiki/Confusion_matrix
40. Steadman, M.: Gradient Boosted Regression Trees. DataroRot, vol. 2019, (2014)