Trend Drift Discovery for Individual Highway Drivers through Ensemble Learning

Weilong Ding†, Zhe Wang, Yanbo Han
School of Information Science and Technology, North China University of Technology, Beijing, China
Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data, Beijing, China
dingweilong@ncut.edu.cn

Jianwu Wang
Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD, U.S.A.

ABSTRACT

Inter-city transportation plays an important role in modern smart cities, and has accumulated massive spatio-temporal data from various sensors in IoT (Internet of things). Current travel characteristics and future trends of highway traffic are valuable for traffic guidance and personalized service. As a routine domain analysis, trend drift discovery for highway drivers faces challenges in processing efficiency and predictive accuracy. Sensitive privacy of business data has to be considered, executive latency on huge data is hard to guarantee, and correlation among spatio-temporal characteristics cannot be fully employed. In this paper, a travel-characteristic based method is proposed to discover the potential drift of payment identity for individual highway drivers. Considering time, space, subjective preference and objective property, monthly travel characteristics are modeled on toll data from highway toll stations, and predictive error for those trends can be reduced dramatically through gradient boosting classification technology. With real-world data of one Chinese provincial highway, extensive experiments show that our method has second-level in executive latency with more than 85% F1-score for predictive accuracy.

KEYWORDS

Spatio-temporal data, Travel characteristics, Trends drift, Ensemble learning, Highway, Big Data.

† Corresponding author.

UrbComp2020, KDD 2020 workshop, August 23 – 27, 2020, San Diego, California USA

1 Introduction

With the flourish of inter-city transportation, highway plays an important role in modern smart cities, and most urban drivers have participated in it unconsciously. It also brings the traffic congestion issue, which is one of the most serious problems worldwide nowadays [8]. Accordingly, highway IoT (Internet of Things) environment is built for official traffic management on various business data of extensive deployed sensors. Big Data technology has been widely adopted in domain analytics recently. As a typical one, trend drift discovery is to predict the attributes and their update for individual driver or group drivers in next few days. Highway drivers' travel characteristics are necessary and often employed from business data. For example, with the advantages of exact locality and higher quality, the toll data from toll stations can be used [2]. It implies spatio-temporal correlation of certain drivers, because the timestamps and locations when a vehicle was entering or exiting a station were kept in the data.

However, it faces challenges to predict such potential trends due to inherent limitations in practice. First, exact personal profile of individual drivers is sensitive and not available directly. For example, driver names with license plates are kept by police officers; ETC (electronic toll collection) accounts are maintained by corresponding banks. Second, it is hard to hold low latency during analytical execution when data grows into a huge size. For example, ARIMA regression treats sensory data as time-series, and returns results once for a single location [3]. It is inefficient for hundreds of locations and huge data in practical highway network. Third, sufficient accuracy is hard to guarantee because traditional methods cannot fully cover key characteristics. For example, time restricted models only emphasize temporal feature using limited business factors. In fact, besides

temporal patterns, spatial feature of localities (e.g., road network topology), personal preferences and travel modes in historical accumulation would also directly influence the travel trends of highway drivers. All of them have been seldom considered synthetically in current works yet. In brief, it is not trivial to find drivers' potential trends in highway domain, in views of both feasibility and efficiency.

In this paper, we take the attribute of payment identity in highway driver as an example, and propose a travel-characteristic based method to discover the potential drift of such an attribute. Our contributions can be concluded as follows. (1) To adequately describe business feature on toll data without personal privacy, travel characteristics are modelled fully considering time, space, subjective preference, and objective property. Such characteristics are organized efficiently as dedicated drivers' trajectories. (2) To improve performance and accuracy of trends prediction, an ensemble-learning model through gradient boosting classification is built. With second-level executive latency on monthly data, it can reduce predictive error about 2%-14% than traditional ways. (3) Evaluated quantitatively and qualitatively in a practical scene, our work has convincing benefits on the real-world data through extensive experiments and case studies.

2 Related work

Potential trends of urban traffic are significant nowadays, but their discovery still faces challenges in efficiency and accuracy [7]. We divide related works into two technical perspectives as follows.

Through offline collecting, organizing and inferring procedures on Big Data, user profiling is to achieve users' interests, characteristics, behaviors, and preferences. For comprehensive information representation and efficient personalized service, it is often used as tagging model in an interactive way. Domain-dependency is one of the shortcomings among current works [4], which often requires business rules and experts knowledge. In transportation related domains, user profiles can be employed for either group user or individual user. In most cases, large dataset are required for acceptable accuracy during user profiling. On heterogeneous traffic data, a three-fold influence model is proposed on group users (crowd) for their traffic propagation [9], which finds cascading patterns through maximizing probability likelihood. On massive trajectory data, a peer and temporal-aware representation learning based framework is presented [12] for drivers' behavior analysis through multi-view driving state transition graphs. On massive POI check-in data of passengers, adversarial sub-structured representation learning is introduced [11] for individual user profiling. Through machine learning technologies, those works have performed well in predictive accuracy but still remains challenges in highway domain. Unlike the scenes in above works, detailed basic profiles of highway drivers are not available because it is maintained in other security-sensitive domains like bank and police. Only highway business data at run-time (i.e., toll data recording drivers' trips at stations) is accessible. The travel characteristics are built accordingly through trajectory structure to discover potential user trends.

On continuous data, the representation of users with more than one dimension would be variable 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 of continuous data streams, named concept drift [6], often affect that data distribution and deteriorate the performance of existing classifiers[15]. Due to distinct criteria of changes, many 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 in this sub-section. Such detection approaches is not trivial and have to trade-off between performance and cost. Concept drift can be seen as a change in a probabilistic perspective as the joint probability distribution [5] among of data samples and their corresponding class labels. A basic assumption here is that the timing and distribution of concept drift are initially unpredictable. Accordingly, most learners first require detecting a concept drift occurs, then reacting to it by new learned data distribution, and an update classification model. In transportation related domains, concept drift detection is widely used in control centers to predict traffic conditions [16]. Concept Neurons framework [10] is proposed to empower the resistance of algorithms for concept drifts. It leverages on a combination of continuous inspection schemas and residual-based updates over the model parameters with output. To handle different drift types, it has been successfully applied on predicting highway traffic congestion in Porto. The domain scenes and technical problems focused in those works are different from that of ours. In this paper, the concerned problem can be regarded as a long-term single drift on labeled data of highway domain, but the monthly period here is too long to endure in an online stream processing. Inspired by the ensemble-learning methodology for concept drift detection, we introduce our model on accumulative data instead of the real-time one to find future trends in a periodical reaction manner. That is, our solution is

actually a prediction procedure on hybrid historical data.

In brief, novel classification model is still required to discover drivers' potential trends in specific highway domain.

3 Preliminaries

3.1 Motivation

Our research originates from Highway Big Data Analysis System of Henan which is 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. We focus on toll data in this paper. A record of toll data has the spatio-temporal structure as Table 1, which contains 12 attributes including six entity attributes, two temporal attributes and four spatial attributes.

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

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	2016/2/23 15:32:06	Time
exit_time	vehicle exit timestamp	2016/2/23 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 a routine business analysis, the trend discovery of payment identity focused in this paper is to periodically find the drift trends of identity for individual highway drivers in the next month. In traditional ways to discover the trends of payment identity, partial historical toll data of sensors would be loaded into a production data warehouse regularly (e.g., monthly even yearly); after ETL (Extract, Transform, Load) step with necessary pre-processing like [14], business OLAP (Online Analytical Processing) would be triggered to execute in that data warehouse; when completed, predicted drivers' payment identity categories or their probabilities can be accessed by business technicians for further interpretation. However in practice, such prediction brings long delays (e.g., one week) to release official reports due to complex processing procedure. 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.

3.2 Methodology

In highway domain, potential drift of payment identity can be evaluated by recent trends for any driver. In this paper, we focus on the monthly trends of individual highway drivers defined below.

Definition 1: Payment identity trend. The payment identity trend of any highway driver v∈V is presented as the predictive probability pair of category ETC and ETC in next month m. In such a pair, the ETC probability of driver v is P_v^+, and the MTC one is P_v^-.

Here, V is the set of drivers appeared in highway where any v could be distinguished by his vehicle's license plate.

Accordingly, potential drift of payment identity can be defined as follows.

Definition 2: Potential drift of payment identity. Based on Definition 1, the potential drift of payment identity for driver v in month m contains two facts: (1) given the predefined threshold h, the potential positive drift into ETC category if $P_v^+ - P_v^- \geq h$; (2) the potential negative drift into MTC if $P_v^+ - P_v^- \leq -h$. Moreover, due to the fact $P_v^+ + P_v^- = 1$, only P_v^+ is sufficient for the evaluation: positive drift implies $P_v^+ \geq (1+h)/2$, and negative drift means $P_v^+ \leq (1-h)/2$.

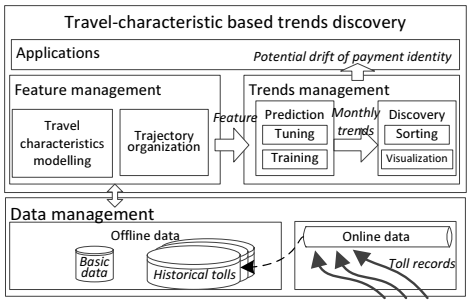

Figure 1: Overview of our method.

The overview of our method for potential drift discovery is illustrated as Figure 1, where four main parts are included. (1) Data management layer loads required data during trends prediction. Raw online toll records are received continuously through a dedicated message service, and then accumulated as historical tolls and aggregated as traffic volumes into No-SQL database. Necessary data cleaning and aggregative calculation are completed as pre-processing. Business basic data, such as profiles of highway station, section, line and region, has been imported in a relational database. (2) On such spatio-temporal data, the feature of travel characteristics is modeled and organized in feature management module. (3)With the features, trends management module adopts GBDT (Gradient Boost Decision Tree) technology to train an ensemble-learning model after algorithmic parameters tuning. Monthly trends of drivers' payment identity would be calculated by the trained model, written into the relational database, and employed for drift discovery after sorting procedures. (4)The visualized potential drift would be presented in online applications through dedicated API (application programming interface) for business management and custom service.

4 Potential drift discovery by classification prediction

4.1 Feature modelling and driver trajectory

Travel characteristics of drivers are key ingredient for predictive estimation and their feature has to be modelled properly.

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 discussed in observation 5; $l_v^m = (0..1)$ as a subjective preference dimension is a proportion that trips of v involve certain major cities S; $p_v^m = (0..1)$ as another subjective preference dimension is a proportion v travels in top-K OD (Origin-Destination) patterns.

The dimensions are illustrated in details below.

(1) The first two dimensions t_v^m and s_v^m imply the frequency of drivers' highway usage. The third dimension o_v^m implies the category of vehicle owned by a driver. Among the range, the first five [1...5] is the respective vehicle-model types of passenger-cars; the last four [6...9] is ones of freight-carriages.

(2) The fourth dimension $0 \leq l_v^m \leq 100\%$ reflects subjective preference influence about hot locations. A driver's trips involve a city c if they passed toll stations belonging to c. Such involvement facts 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.

(3) The fifth dimension $0 \leq p_v^m \leq 100\%$ reflects subjective preference influence about habitual travel patterns. OD is a domain technical term about travel demand in transportation related research. As a pair <origin, destination>, the OD of a trip can be extracted from the toll data in Table 1. As the frequent travels, the top-K OD in this definition reveals pattern typicality, where $1 \leq K \leq 3$ in common.

With the defined feature of travel characteristics, we propose a data structure driver trajectory to refactor original toll 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 noted as Definition 3, and each of multiple 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 as Table 1.

The trajectory building is an offline processing depicted as a MapReduce job. The map phase 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 groups the intermediate results by a composite key including vehicle license and month, counts mileage summary, trip size, OD proportion, city proportion, and Y_v^m.

4.2 Identity classification and potential drift detection

To discover potential drift through travel characteristics from driver trajectories, a classification model based on GBDT (Gradient Boosting Decision Tree) is designed in this section. 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 $\sum_{v\in V}\log(1+\exp(-Y_v^m F(X_v^m)))$, for input $(X_v^m, Y_v^m), v \in V$. At each stage d, 1≤d≤D, a weak model Fd would fit hd(x) to previous residuals by gradient boosting. That is, each Fd attempts to correct the errors of its predecessor model Fd-1. Besides the loss function Lf, optimized goal includes Ω(hd(x)) ~ (Fd_depth, Fd_shrinkage), which is the regularization part restricted by depth and shrinkage rate of base trees. For such regularization, we refer the concepts in XGBoost [1]: 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 training of classification prediction for drivers' payment identity can be implemented 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 $\in \mathbb{N}$, r $\in \mathbb{R}^+$, and $0 \le \eta \in \mathbb{R}^+ \le 1$.

Table 2: Model training for the trends of drivers' payment identity.

Algorithm: *classification model training through GBDT*

Input: drivers' trajectories $TR_v^{m-1}, v \in V$, tree size *D*, maximal tree depth *H*, tree shrinkage *r*, and training ratio η.

Output: an ensemble model F_D to get payment identity trend of any highway driver in the next month.

1. randomly select $\eta * |V|$ ones from trajectories as $TR_v^{m-1}, v \in V', |V'| = \eta * |V|$;
2. request feature of travel characteristics (x_v, y_v) from selected trajectories $TR_v^{m-1}, v \in V'$.
3. initialize a model with constant values: $F_0(x) = \arg\min_\gamma \sum_{v=1}^{|V'|} Lf(y_v, \gamma)$;
4. for *d*=1 to *D*
5. 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 γ_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: $F_d(X) = F_{d\text{-}1}(X) + \gamma_d h_d(X)$;
11. endfor
12. return $F_D(X)$

With drivers' trajectories with series of parameters as input, the algorithm would output an ensemble model to get payment identity trend of any highway driver in the next month. According to the training ratio parameter η, training set is selected randomly as a subset among input trajectories in Line 1. Note that, only qualified trajectories with given flag value would be selected here. Considering space cost, the selection action here only records the indexes in NoSQL storage instead of actual trajectories. The feature of travel characteristics is achieved by index retrieval of the training set as Line2. To initiate the model, F0(X) is generated with constants γ as Line 3. After computing residuals via negative gradient direction in Lines 5-7, a base tree hd(X) as a weak model is built to fit xv with those residuals rid in each iteration as Line 8. To find

an approximation that minimizes average value of loss function Lf and regularization Ω, weight γd 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 γdhd(X) as Line 10. At last, the final model FD(X) as an ensemble of D base trees (i.e., in D iterations) is returned like Line 12.

Through the learned model FD, the potential drift discovery of payment identity can be found as follows. With the feature (x_v, y_v) of any driver v, the trend as a probability pair of Definition 1 can be achieved through FD(x_v). The potential 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 negative drift; when v is MTC identity, y_v=-1, $(P^+ - P^-) \geq h$ implies positive drift.

5 Evaluation

5.1 Setting

In the project mentioned in Section 3.1, our method is evaluated by experiments. 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. In the practical scene of Henan highway, toll data from more than 300 toll stations would be generated 1.5 million records a 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 modules of feature management, trends management and an online application 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 business analysis at 12:00 a.m. of 1st day in current month m on the data of previous month m-1, the classification prediction model would be triggered for re-training, and the trends of drivers' payment identity and potential drifts are achieved for month m. All those results (i.e., trends of any drivers and potential drifts) would be written to a dedicated table of HBase.

5.2 Experiment

Before the evaluation of prediction effects, three common metrics as follows are employed.

Definition 4: Metrics for classification prediction. According to terms in confusion matrix [13] 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; F1-score is defined as Equation 3. Here, with a coefficient β, the F1-score is the harmonic mean of the precision and recall, whose best value is at 1 and worst is at 0. To equally weight precision and recall here, we set $\beta = 1$ constantly.

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

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

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

Algorithmic parameters tuning is discussed first. The classification prediction model would be trained with proper algorithmic parameters. Among those trajectories with $\eta = 50\%$ and after multiple cross-validations, we found the optimal parameters: H=3, D=2000 and r=0.005.

Then, an experiment is conducted to show the results of trends prediction through the trained model above (i.e., M=2000, H=3, r=0.005). In order to quantitatively evaluate classification effects of our model (abbr., GBDT), three other models are also implemented for comparison in trends management module of Figure 1. They are non-linear model SVM (Support-Vector Machine), non-parametric model KNN (K- Nearest Neighbor) and binary Logistic model. All these counterparts are tuned respectively on the same training set and validation set as Experiment 2.

Experiment: Predictive effects. The driver trajectories in 12 months of 2016 are employed to evaluate prediction effect. Here, the parameter training ratio $\eta = 0$ which means the travel characteristics from all the trajectories here are used as test set, and the model is kept as the last trained one in Experiment 2. Through any of the four models to predict drivers' payment identity, all three metrics and the executive time are counted after the finish of classification prediction.

Predictive effects are illustrated as Figure 2, where first threes are dimensionless and the unit of the fourth is second. We found our method performs well in precision and F1-score with relatively low executive time. (1) All four models can achieve trends in months of that year, and our GBDT has the advantage in precision. Among the fours in 12 months, the precisions are more than 70%, the recalls are around 80%, and F1-scores are more than 84%. Even with the lowest recall and F1-score, KNN is still sufficient enough in practice. The high accuracy in metrics is owed to the travel characteristics we defined which include multiple spatio-temporal dimensions. The practical feasibility of our proposed feature is proved. Our method GBDT performs best in metric precision and second-best in metric F1-score, which comes from

the predictive capacity by our ensemble-learning model. The excellent predictive effect of our proposed classification model is proved either. (2) Our GBDT has evident advantages in executive latency. It cost steady one second due to fast convergence of the algorithm of Table 2, while that of Logistic appears a little better with much more fluctuations. Although KNN consumes the least time sometimes, but achieves the worst predictive effects in metric recall and metric F1-score. Note that, the executive time of SVM is longer by an order of magnitude than others, and it is measured in additional vertical coordinates as Figure 2(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 seven-day holiday (nearly one-quarter in a month) exists when the vehicles in highway would be free of charge due to national regulations. It makes vague payment identity of drivers and would confuse factual characteristics. (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 no re-training is done for these models in this experiment. In fact, as we have mentioned in Section 3.2, our method would update the prediction model once a month to fit the recent trends better, and performs more accurate results than these results. However, even in the valley of the results, our method still holds the precision larger than 76% and the F1-score larger than 84%. It comes from the travel characteristics we defined in driver trajectories, and high accuracy and feasibility are proved. All those make our method more practical and efficient in highway domain.

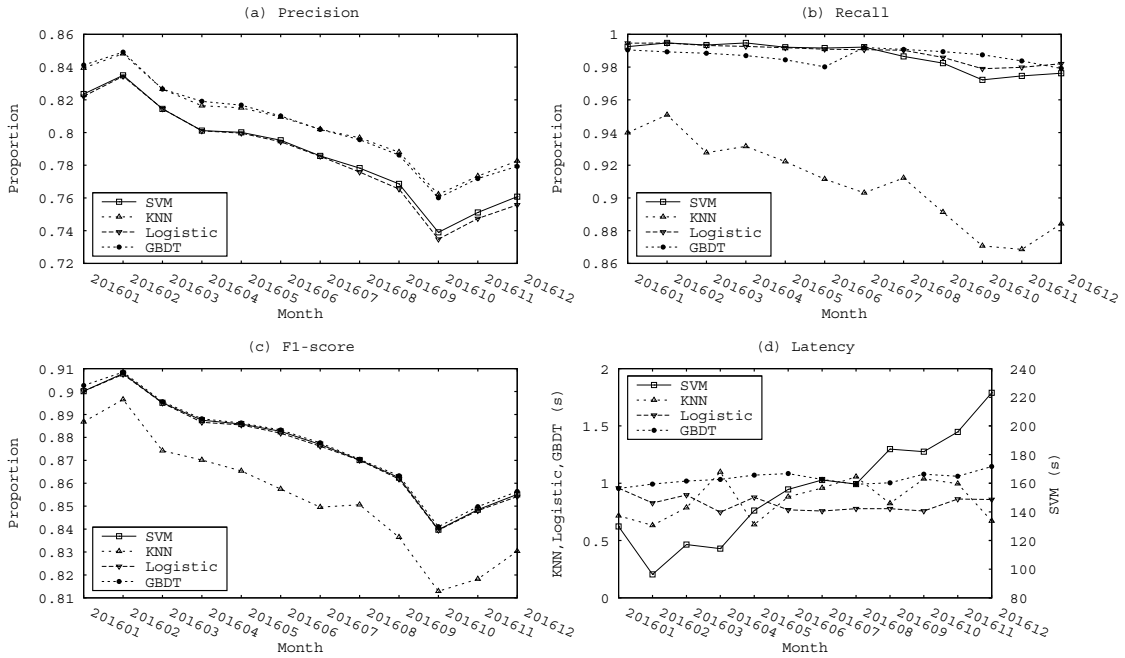

Figure 2: Predictive effects.

In summary, our method performs well in executive performance and holds high predictive accuracy.

6 Conclusion

In this paper, a novel travel-characteristics based method is proposed to discover highway drivers' potential trends. On massive historical spatio-temporal data, the feature of travel characteristics is built efficiently through defined trajectory structure, considering about time, space, objective facts, and subjective preference of highway drivers. Through gradient boosting classification technology, a prediction model is trained with low executive latency and well-performed accuracy than traditional models during trends prediction. In extensive experiments and case studies on real data of one Chinese province, the time consumption is improved to less than 3 seconds, metric precision is near 75%, metric recall is around 95%, and metric F1-score is more than 85%. Potential drifts about highway driver's payment identity can be

represented intuitively as user profiling views in various online visualizations.

Due to confusing recognition of payment identity on some special holidays, trends and potential drifts discovery in those periods is a big challenge. In our future work, dedicated and fine-grained characteristics are planned to employ in feature accordingly.

REFERENCES

[1] CHEN, T. AND GUESTRIN, C. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA2016 ACM, 2939785, 785-794.

[2] DING, W., WANG, X. AND ZHAO, Z. 2020. CO-STAR: A collaborative prediction service for short-term trends on continuous spatio-temporal data. Future Generation Computer Systems 102, 481-493.

[3] DING, W. AND ZHAO, Z. 2018. DS-Harmonizer: A Harmonization Service on Spatio-Temporal Data Stream in Edge Computing Environment. Wireless Communications and Mobile Computing 2018, 12.

[4] EKE, C.I., NORMAN, A.A., SHUIB, L. AND NWEKE, H.F. 2019. A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions. IEEE Access 7, 144907-144924.

[5] GAO, J., FAN, W. AND HAN, J. 2007. On Appropriate Assumptions to Mine Data Streams: Analysis and Practice. In Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM 2007), 28-31 Oct. 2007 2007, 143-152.

[6] HU, H., KANTARDZIC, M. AND SETHI, T.S. 2020. No Free Lunch Theorem for concept drift detection in streaming data classification: A review. WIREs Data Mining and Knowledge Discovery 10, e1327.

[7] KOLAJO, T., DARAMOLA, O. AND ADEBIYI, A. 2019. Big data stream analysis: a systematic literature review. Journal of Big Data 6, 47.

[8] LA A, I., LOBO, J.L., CAPECCI, E., DEL SER, J. AND KASABOV, N. 2019. Adaptive long-term traffic state estimation with evolving spiking neural networks. Transportation Research Part C: Emerging Technologies 101, 126-144.

[9] LIANG, Y., JIANG, Z. AND ZHENG, Y. 2017. Inferring Traffic Cascading Patterns. In Proceedings of the Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA2017 ACM, 3139960, 1-10.

[10] MOREIRA-MATIAS, L., GAMA, J. AND MENDES-MOREIRA, J. 2016. Concept Neurons – Handling Drift Issues for Real-Time Industrial Data Mining. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases(ECML PKDD 2016), Riva del Garda, Italy2016 Springer International Publishing, 96-111.

[11] WANG, P., FU, Y., XIONG, H. AND LI, X. 2019. Adversarial Substructured Representation Learning for Mobile User Profiling. In Proceedings of the Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA2019 ACM, 3330869, 130-138.

[12] WANG, P., FU, Y., ZHANG, J., WANG, P., ZHENG, Y. AND AGGARWAL, C. 2018. You Are How You Drive: Peer and Temporal-Aware Representation Learning for Driving Behavior Analysis. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2018), London, United Kingdom2018 ACM, 3219985, 2457-2466.

[13] WIKIPEDIA 2020. Confusion matrix.https://en.wikipedia.org/wiki/Confusion_matrix

[14] XIA, Y., WANG, X. AND DING, W. 2018. A Data Cleaning Service on Massive Spatio-Temporal Data in Highway Domain. In Proceedings of the Service-Oriented Computing – ICSOC 2018 Workshops, Hangzhou, China2018 Springer International Publishing, 229-240.

[15] ZHANG, W. AND WANG, J. 2017. A Hybrid Learning Framework for Imbalanced Stream Classification. In Proceedings of the IEEE International Congress on Big Data (BigData Congress 2017), Honolulu, HI, USA, 25-30 June 2017 2017 IEEE, 480-487.

[16] ŽLIOBAITĖ, I., PECHENIZKIY, M. AND GAMA, J. 2016. An overview of concept drift applications. In Big data analysis: new algorithms for a new society Springer, 91-114.