Big Data Driven Vehicular Networks

Nan Cheng, Feng Lyu, Jiayin Chen, Wenchao Xu, Haibo Zhou, Shan Zhang, and Xuemin (Sherman) Shen

Abstract

VANETs enable information exchange among vehicles, other end devices and public networks, which plays a key role in road safety/infotainment, intelligent transportation systems, and self-driving systems. As vehicular connectivity soars, and new on-road mobile applications and technologies emerge, VANETs are generating an ever-increasing amount of data, requiring fast and reliable transmissions through VANETs. On the other hand, a variety of VANETs related data can be analyzed and utilized to improve the performance of VANETs. In this article, we first review VANETs technologies to efficiently and reliably transmit big data. Then, the methods employing big data for studying VANETs characteristics and improving VANETs performance are discussed. Furthermore, we present a case study where machine learning schemes are applied to analyze VANETs measurement data for efficiently detecting negative communication conditions.

Introduction

With the development of automobile technologies, vehicles are expected to be not only safer, but also greener, more comfortable and entertaining, while self-driving is also a defining requirement of future vehicles. As a promising technology to meet such expectations, vehicular communication networks (VANETs) enable automobiles to communicate with each other through vehicle-to-vehicle (V2V) communication and the network through vehicle-to-infrastructure (V2I) communication, and exchange information efficiently and reliably through the V2V and V2I communications, or more generally, vehicle-to-everything (V2X) communications. VANETs can facilitate a variety of useful applications, such as road safety enhancement, traffic management, vehicular mobile data services, and self-driving assistance [1, 2].

Due to the ever-increasing demand of mobile services and the fast development of self-driving technologies, the data volume required, generated, collected, and transmitted by VANETs has seen an exponential escalation, which is known as big data [3]. As explained in [4], the data in VANETs can well match the "5Vs" of big data characteristics, that is, volume, variety, velocity, value, and veracity, which justifies that VANETs data can be treated as big data and can be solved by big data techniques.

Relying on big data, future VANETs will enable a variety of promising applications and services, such as smart city and Intelligent Transportation System (ITS) applications, and significantly change many aspects of the society, including the transportation system, telecommunication, business, government, and human life styles. VANETs big data and enabled applications are shown in Fig. 1. For example, road traffic information can be collected by vehicles and road-side units, and reported to the ITS cloud server. Based on large-scale traffic information, real-time traffic prediction and management functions are conducted, so as to detect road anomalies, alleviate traffic jams, and reduce emissions and pollution. Self-driving vehicles will consume or generate multiple giga bytes (GB) of data per second, typically from outfitted high-quality cameras, LiDARs and radars [5]. Through data fusion, analysis and integration of the cloud data such as weather and road traffic information, and information from other vehicles, self-driving vehicles can make decisions on actuating the vehicle for driving autonomously, on a planned route, and eliminate traffic fatalities. As a potential impact of self-driving technologies, vehicles will be more like home or offices, and thus people will focus on the mobile applications and services that can better support in-vehicle activities, rather than driving the vehicle. Therefore, future VANETs will evolve to satisfy big mobile data demands, and support a wide variety of promising applications and services.

The trend toward big data can bring new challenges and opportunities for VANETs. On one hand, VANETs big data is with a significantly large amount, from heterogeneous sources, and having various requirements. To efficiently support big data, VANETs should be capable of providing extremely high data rate, large network capacity, heterogeneous network integration, and differentiated quality of service (QoS) guarantees. In addition, besides data communication, future VANETs are envisioned to play a critical role in data collection, storage, and computation. On the other hand, VANETs big data such as GPS, vehicle mobility trace, road traffic information, and network measurements, contains rich valuable network information. If properly utilized, such big data can reveal many network characterizations, evaluate network performance, and optimize network management, by applying advanced techniques such as big data mining, analysis, and machine learning mechanisms. The purpose of this article is to investigate the impacts of big data on VANETs, introduce the new chal-

Digital Object Identifier:
10.1109/MNET.2018.1700460

Nan Cheng, Jiayin Chen, Wenchao Xu, and Xuemin (Sherman) Shen are with the University of Waterloo; Feng Lyu is with Shanghai Jiao Tong University; Haibo Zhou (corresponding author) is with Nanjing University; Shan Zhang is with Beihang University.

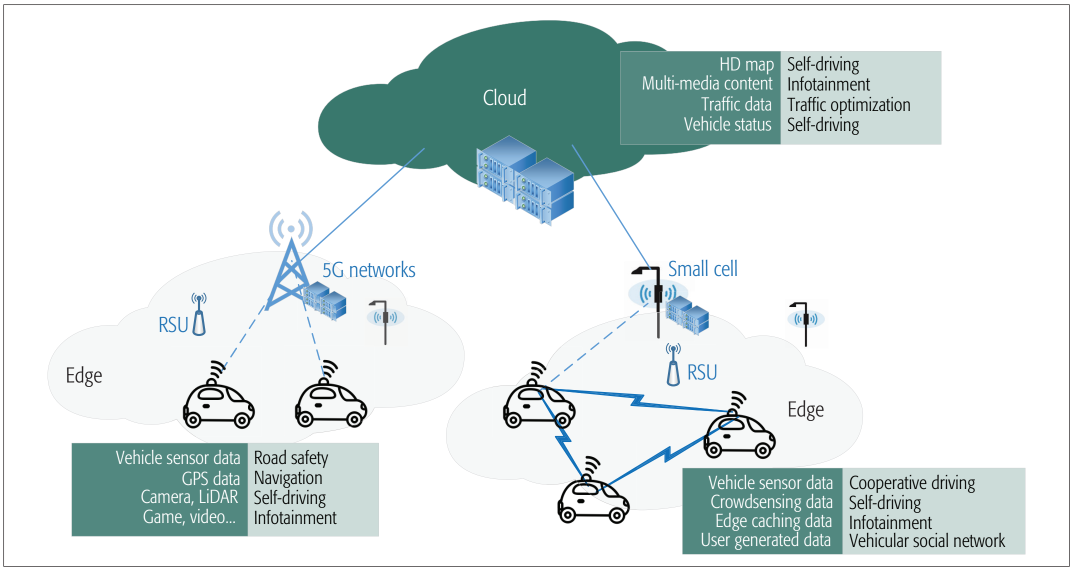

FIGURE 2. VANETs big data and applications.

lenges and opportunities, and discuss corresponding solutions. We focus on two related topics, that is, efficiently supporting big data in VANETs, and utilizing big data for better understanding and improving VANETs. We study a case where machine }learning schemes are applied to analyze VANETs measurement data for efficiently detecting negative communication conditions.

Big Data in VANETs

VANETs big data come from multiple heterogeneous sources, presenting diversified characteristics, such as volume, structure, value, requirements for processing delay, and so on. We classify the VANETs big data according to the sources of the data as follows.

Vehicle Sensing Data: Modern vehicles are equipped with various sensors (speedometer, tire pressure sensor, and so on.) to collect vehicle and environmental information. Rich information from such sensors can enable a wide range of applications, such as online vehicle diagnosis, road safety improvement, smart charging planning, accident detection, and so forth.

GPS Data: GPS devices can provide accurate and structured location-related information of vehicles, including longitude, latitude, altitude, and speed. GPS data can be used for diversified goals, such as navigation, traffic management, communication routing optimization, vehicular content caching and sharing, and so on. In addition, the datasets of large-scale vehicle trajectories, generated by tracing the long-term GPS data of vehicles in a geographical area, can be investigated to analyze VANETs characteristics, such as network connectivity, and design efficient mechanisms, such as routing protocols for delay-tolerant vehicular networks, and radio access network deployment.

Self-Driving Related Data: The autonomous vehicle will make big data even bigger. Self-driving technology requires the accurate perception and understanding of the environment to make proper decisions to control the vehicle. Since traditional sensors have limited capability, and cannot provide necessary information such as real-time road vision, accurate distance, and 3D maps, advanced devices like cameras and light detection and ranging (LiDAR) sensors are equipped for better perception. However, the high-definition cameras and LiDAR will produce a huge amount of data as they continuously collect high-definition data such as high-quality videos.

Traditional VANETs technologies can hardly satisfy the harsh requirements of big data applications due to the decentralized protocols and bandwidth limitations, which leads to the lack of network resources and flexibility to support big data with diversified QoS requirements.

Vehicular Mobile Service Data: In-vehicle infotainment is becoming more crucial for improving the experience of both drivers and passengers. Mobile applications such as video/audio streaming, online gaming, social networks, and user generated content (UGC) require or generate a huge amount of data.

Supporting Big Data in Vehicular Networks

For big data systems to efficiently function, four essential parts need to be well supported, that is, data aggregation, storage, transmission, and computation. In VANETs, raw data can be gathered by vehicle sensors, and stored in on-board storage. Since the raw data contain redundancy, data processing is conducted to extract valuable information. After accumulating the data (either raw or processed), there is a demand to transmit the data to appropriate data storage systems (such as cloud/edge servers) for further analysis and processing. Therefore, VANETs should be capable of effectively supporting these big data functions.

Traditional VANETs employ IEEE 802.11p based dedicated short-range communication (DSRC) technologies, where data transmission mainly relies on distributed medium access control (MAC) and multi-hop routing protocols [6]. However, traditional VANETs technologies can hardly satisfy the harsh requirements of big data applications due to the decentralized protocols and bandwidth limitations, which leads to the lack of network resources and flexibility to support big

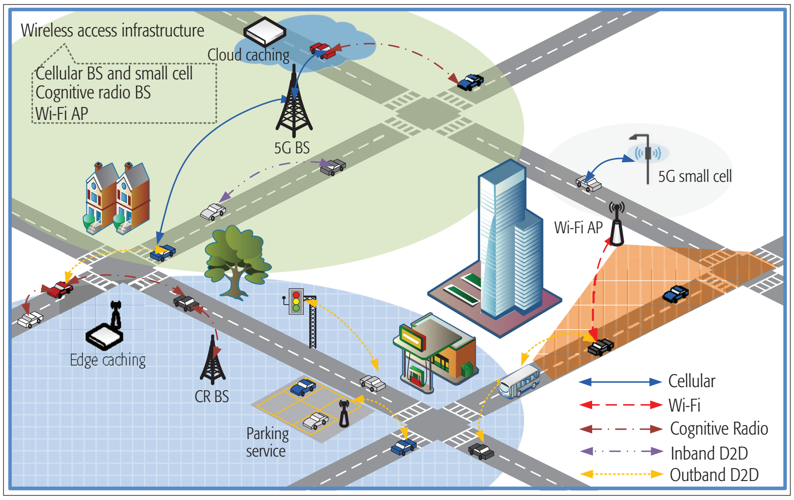

FIGURE 2. Supporting big data through VANETs.x

data with diversified QoS requirements. Moreover, issues such as energy efficiency, caching, and computation capabilities are not well considered in current VANETs, which are also essential in supporting big data. In this section, we discuss some promising VANETs technologies to better support big data, including 5G technologies and opportunistic data offloading mechanisms. As shown in Fig. 2, 5G macro cells can provide ubiquitous communication support, while 5G small cells, Wireless Local Area Networks (WLANs), cognitive radio networks (CRNs) and device-to-device (D2D) communications offer cost-effective data pipes for VANETs big data.

5G Technologies

An intuitive solution to support VANETs big data is the pervasive cellular network. As the 4G LTE network is struggling to support ever-increasing data volumes and the emerging mobile services with differentiated QoS requirements, 5G networks, the next-generation networks, are building a way to solve the issues. Based on software-defined network (SDN) related technologies, 5G networks are designed to serve as a platform to provide satisfying services for vertical fields, including telecommunication, transportation, agriculture, economics, government, education, etc. [7]. According to the key performance indicators, 5G networks are capable of offering a 10 Gb/s data rate with less than 1 ms end-to-end latency [8]. Moreover, machine-type communications with low power consumption and high reliability requirements are well supported for the emerging Internet of Things (IoT) applications.

To better characterize and support different services, 5G defines three categories of use cases, that is, enhanced mobile broadband (eMBB), ultra-reliable and low-latency communication (URLLC), and massive machine-type communication (mMTC), and the performance indicators of each categories. These three categories, together with the well-defined key technologies, can provide guaranteed performance to VANETs big data gathering and transmission tasks.

eMBB: In VANETs, the exponentially increasing big data demands of vehicular mobile data services requires a high-capacity network that can provide extremely high date rates. Enabled by promising network technologies, such as advanced channel coding, mmWAVE, and ultra-dense small cell networks, eMBB can provide a peak data rate of 10 Gb/s and mobile data volume of 10 Tb/s/km^2. Therefore, with 5G networks, emerging data-craving vehicular data applications can be better supported, and many more will come to reality.

URLLC: Mission-critical data services in VANETs, such as safety message transmission, require very low latency and very high reliability. The requirements fall into the category of URLLC in 5G, which can provide less than 5 ms latency and higher than 99.999 percent reliability.

mMTC: Relying on potential technologies such as machine-to-machine communication and narrow-band IoT (NB-IoT), mMTC aims to support ubiquitous machine-type connections with low energy consumption and low latency. A large amount of VANETs big data is generated by densely deployed lightweight devices, such as sensors equipped in vehicles or deployed along the roads. 5G technologies can accommodate such massive concurrent connectivity, provide reliable data transmission, and prolong device battery life, and therefore facilitate big data gathering services.

5G also defines enhanced vehicle-to-everything (eV2X) use cases for supporting the vertical field of vehicular communication and data services [9]. The requirements for typical V2X scenarios are defined, including vehicle platooning, advanced driving, extended sensors, and remote driving.

Opportunistic Data Pipes

Although 5G networks can significantly improve network capacity, the ever-increasing big data will still put a severe burden on the network, resulting in possible network congestion. In addition, the commercialization and deployment of 5G networks will start in the year 2020, and will be a long-term process. Therefore, in the near future, the 4G LTE networks with relatively small capacity will be straining to accommodate big data. Moreover, usually using the cellular network to transmit a large amount of data will incur prohibitive costs. As a result, alternative data pipes for supporting big data are required. WLANs, CRNs and D2D communications can be employed to offload VANETs big data from the cellular network in a cost-effective way.

WiFi Offloading: WiFi, operating on unlicensed spectrum, is a popular solution to deliver data content at low cost. The feasibility of WiFi for outdoor Internet access at vehicular mobility, referred to as drive-thru Internet, has been demonstrated in [10]. Different from the fully covered cellular network, WiFi only provides intermittent small coverage areas along the road. Therefore, although WiFi operates on unlicensed spectrum, it is spatially/temporally opportunistic for vehicles to employ due to vehicle mobility. Therefore, employing mobility feature is an important issue in vehicular WiFi offloading. One example is prediction-based delayed offloading. Based on mobility prediction and priori knowledge of WiFi deployment, future opportunities of WiFi access and corresponding throughput can be predicted. Then, according to the delay tolerance of different users or applications, offloading decision can be made whether to wait for WiFi offloading or directly transmit through cellular networks.

Cognitive Radio Technology: Cognitive radio is envisioned as a promising spectrum-sharing technology which enables unlicensed users to opportunistically exploit spatially and/or temporally vacant licensed radio spectrum bands that are allocated to licensed systems. CR technology can employ vast underutilized spectrum resources to support big data transmissions. However, in VANETs, the high mobility of vehicles may require excessively frequent spectrum sensing to protect the primary transmissions [11]. TV white spaces (TVWS) have been suggested for wireless broadband access due to the abundant and currently underutilized spectrum resources at VHF/UHF bands and its superb penetration property. Unlike other licensed system, the spectrum usage of TV broadcasting systems is highly stable and predictable, and can be inquired from a database. Therefore, TVWS is envisioned as a potential solution to CR-enabled VANETs [12].

Device-to-Device Communication: By utilizing proximity, mobile users can communicate directly with each other using the cellular spectrum (or other spectrum bands) without traversing the base station or the backhaul networks, named device-to-device D2D communications. Therefore, D2D communications can increase overall spectral efficiency and reduce communication delay for mobile users, which may be applied to many VANETs applications such as video streaming, location-aware advertisement, safety related applications, and so forth. However, incorporating D2D communication in vehicular environments introduces several new challenges. For example, full channel state information, which is usually needed in resource allocation schemes for D2D communication, is hard to track and easy to be outdated in VANETs. In addition, the topology of VANETs makes the interference pattern more difficult to model than a general cellular network where a Poison point process (PPP) can be applied to model the user spatial distribution.

Employing Big Data in Vehicular Networks

As mentioned above, big data in VANETs can provide valuable insights into VANETs, which can be employed to characterize and evaluate the performance of VANETs, and design new protocols with big data intelligence. In this section, we show the utilization of two typical data sets in VANETs, that is, vehicle mobility trace data and VANETs measurements data. An overview of big data employment in VANETs is shown in Fig. 3. The two data sets can be employed to extract a practical channel model and mobility model, and predict vehicle movement. With such knowledge, VANETs characterization and intelligent protocol design can be achieved.

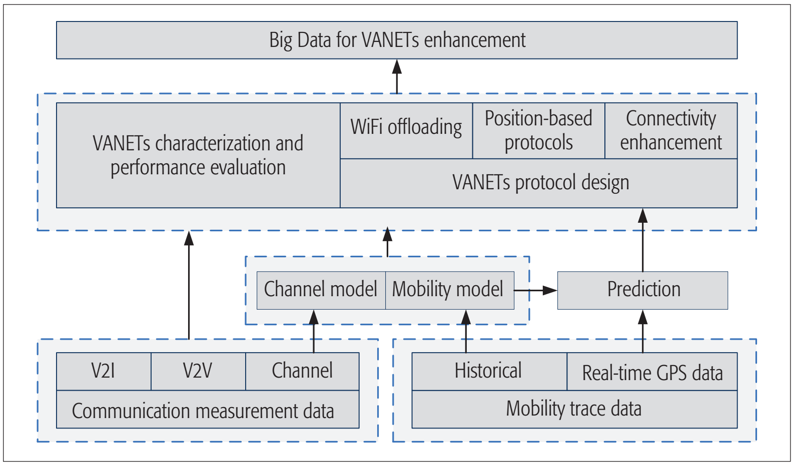

FIGURE 3. Employing big data in VANETs.

Vehicle Mobility Trace Data

Admittedly, the high mobility of vehicles leads to challenges for VANETs. However, mobility can also provide benefits on the network, for example, mobility-aware protocols and delay-tolerant data dissemination. Through the analysis of the datasets of vehicle mobility, an amount of valuable information can be obtained, such as the practical mobility model, network connectivity, spatial and temporal density distribution, and so on. There are several database that stores real and large-scale taxi mobility trace data from different cities, including San Francisco, Shanghai, and Shenzhen [13]. The main content of the trace data includes time stamp, vehicle velocity, driving direction and vehicle location, which can be used for further study of VANETs.

Mobility models are widely used in VANETs location-based protocol design and performance evaluation. Due to the time intervals of vehicles reporting their trace, the trace data is always error-prone and has gaps between locations in two consecutive records. Therefore, some data preprocessing mechanism is needed. For instance, due to the predictability property of vehicle mobility, it is possible to fill the gap by predicting the route through analyzing road maps, traffic signs and the past vehicle traces. Then, a realistic mobility model can be generated from the modified trace data.

Position-based routing schemes and MAC protocols are designed to adapt to the high mobility and frequently changing topology of VANETs. The mobility model and network characteristics can be obtained by analyzing the mobility trace data and network measurement data, which are taken into consideration in the design of routing schemes and MAC protocols. For instance, position-based routing schemes can exploit the real-time position and predict vehicle movement to improve transmission performance. Position-based MAC protocols can predict potential packet collisions due to vehicle mobility and make efforts to avoid them. Historical mobility trace data can also be used in simulations to evaluate the designed MAC and routing protocols.

Furthermore, mobility trace data is also useful in analyzing and improving the connectivity of VANETs. Network connectivity metrics can be evaluated from the mobility trace data, including link duration, average hops, number of connected vehicle pairs, and interconnect time distribution. Improvement of connectivity can also be achieved with the aid of trace data. The prediction methods of vehicle movement can be developed to make seamless handoff possible for communication between vehicles and infrastructures. In

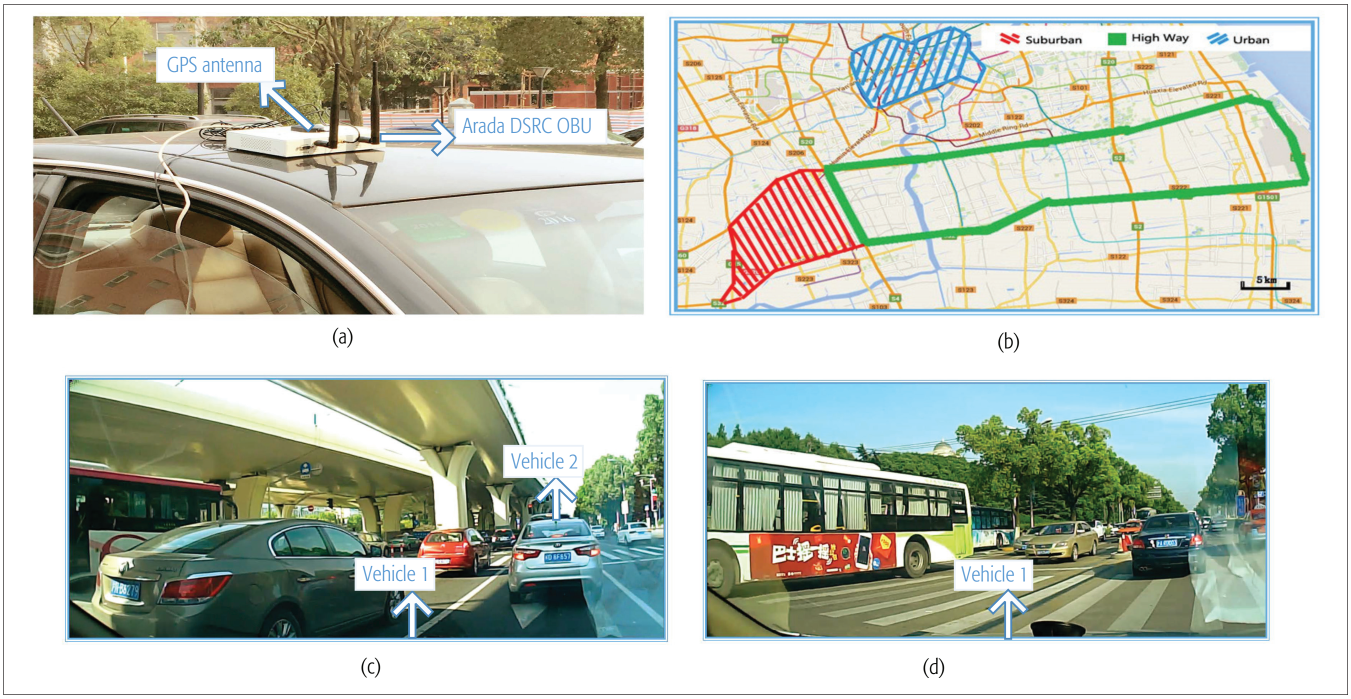

FIGURE 4. Illustration of the data collection campaign: a) data collection devices; b) data collection scenarios; c) LoS condition; d) NLoS condition.

addition, by investigating the real-time trace data generated in VANETs, information about vehicle traffic flow can be obtained. Then, unmanned aerial vehicles (UAVs) can be deployed in order to improve network connectivity.

VANETs Measurement Data

Measurement of VANETs communication plays a vital role in VANETs characterization, since in VANETs, many influencing factors are difficult to model, such as mobile channels, pedestrians, terrain, and obstacles. In order to obtain realistic measurement data, communication devices using the IEEE 802.11p protocol are deployed on vehicles and road-side units (RSUs) during experiments. These experiments are conducted in various environments such as urban, suburban, rural, open fields and freeways, and different measurement data is collected depending on the characteristics of interest.

WiFi offloading is envisioned as a potential solution to the data explosion problem in cellular networks. However, the high mobility of vehicles makes WiFi offloading in VANETs different from static WiFi offloading. Measurement data like connection establishment time, connection time, interconnection time, max rate and transferable data volume in once drive-thru is collected to analyze WiFi offloading performance. Then, a three-phase feature is observed as an important WiFi offloading characteristic, including entry, production, and exit phases. It shows that in the entry and exit phases, the connection quality is weaker and data rate is lower than in the production phase, which provides guidance to researchers about how to improve the offloading performance, for example, reducing the association and authentication time in order to maximize data transfer in the production phase.

Unlike static or low-mobility wireless channels, the vehicular channel is more complicated due to shadowing by nearby vehicles, high Doppler shifts, and inherent nonstationary [14]. Therefore, building an accurate and practical channel model is crucial for VANETs performance analysis, protocol design, and simulation experiments. This can be done by studying real communication measurement data, including both V2V measurements and V2I measurements in different important environments. The resulting channel models characterize the vehicular channel from different channel metrics, including pathloss, signal fading, delay spread, Doppler spread, and angular spread.

Case Study

In this section, we study a case where big data and machine learning schemes are employed to support efficient protocol design in VANETs communications.

Online NLoS Detection

In VANETs, packets related to safety information should be delivered perfectly (with transmission chance and without packet loss). However, it is found that a non-line-of-sight (NLoS) condition is a key factor of V2V link performance degradation [15]. Inspired by the intuition that blindly sending more packets in harsh NLoS conditions can hardly succeed but incur resource wasting and increase interference to other neighboring vehicles, we propose an innovative scheme to *detect NLoS conditions online* by learning the V2V measurement data. Given that the NLoS condition can be detected, more robust protocols can be devised, for example, allocating scarce wireless channel resources to those vehicles under line-of-sight (LoS) conditions or seeking helper vehicles to relay packets for those vehicles under NLoS conditions. In the sequel, we will describe the scheme in two parts, that is, measurement data collection and building a detection model using machine learning methods.

Collecting V2V Communication Measurement Data Sets

We collect V2V communication trace data using two experimental vehicles each mounted with an Arada LocoMate^{TM} OBU (DSRC module) on the roof. The transmitter vehicle sends a 300-bytes packet every 100 ms to the receiver vehicle, which consists of a sequence number, the latitude, the longitude, the altitude and the speed information of the transmitter. Meanwhile, both the transmitter and the receiver log all the packets transmitted and received. In addition, we deploy two cameras on each vehicle with one mounted on the front glass and the other fixed on the rear glass, which record the whole process for off-line analysis. Figure 4a shows the data collection devices.

We conduct data collection campaigns including three major road types in a city, that is, highway, suburban, and urban. Each data set contains the following three types information:

- Communication trace: by comparing a packet's sequence number at the sender and receiver, each packet can be marked as received or dropped and we can compute the packet delivery ratio (PDR) throughout the entire experiment time.
- GPS trace: both vehicles have logged a GPS trace, which can provide speed, altitude and distance information.
- Recorded videos: these can be utilized to check the communicating environments, for example, types of road, traffic conditions, surrounding obstacles and so on.

Three types of data are within time synchronization for better observation and comparison. The overall campaign lasts for over two months with an accumulated distance of over 1,500 kilometers and a total size up to 110 Gb. We run our testbed within areas of the above three road types in Shanghai as shown in Fig. 4b. We denote three data sets by $\mathcal{H}$ (highway), $\mathcal{S}$ (suburban), and $\mathcal{U}$ (urban).

Supervised Machine Learning

In this subsection, we use two classic supervised machine learning methods, that is, Naive Bayes (NB) and Support Vector Machines (SVM), to detect NLoS conditions.

Labeling NLoS Conditions: Before using machine learning techniques, we first label out all NLoS conditions. Since the all the data collection campaigns are recorded by cameras, we mark all NLoS situations when two vehicles cannot visually see each other. Although NLoS conditions found by cameras are not necessarily to be NLoS for RF radios, those visually NLoS conditions are still good approximations of real radio NLoS conditions and valuable for learning. Figures 4c and 4d show examples of a LoS condition and a NLoS condition, where vehicle 1 and vehicle 2 are communicating vehicles, but vehicle 2 is blocked by obstacles in the NLoS condition and cannot be found in Fig. 4d.

Feature Extraction: When machine learning algorithms are processed, a representative tuple of features rather than raw data is a more effective input. Thus, it is necessary to extract effective features from a raw data set. According to the analysis in the work [15], PDRs are heavily influenced by LoS/NLoS conditions and LoS/NLoS durations are with memories due to the power law distributions. Therefore, we can use history PDR values as features for training. At this point, we select three features, that is, PDR value of the previous 1 second, PDR value of the previous 5 seconds, and PDR value of the previous 10 seconds.

> Building an accurate and practical channel model is crucial for VANETs performance analysis, protocol design, and simulation experiments. This can be done by studying real communication measurement data, including both V2V measurements and V2I measurements in different important environments.

Machine Learning with NB and SVM: After feature extraction, we obtain samples in the form of <3-dimensional features, label>. We then use parts of samples to train NB and SVM models. NB methods are a set of supervised learning algorithms based on applying Bayes' theorem. Given a label variable y and a tuple of feature vectors x_1 to x_n, Maximum A Posteriori (MAP) estimation is used to estimate $P(y)$ and $P(x_i|y)$. NB learners and classifiers can be extremely fast compared to some sophisticated methods. The cores in SVM are the kernel and the similarity function. A kernel is a landmark, and the similarity function computes the similarity between an input example and the kernels.

Performance Evaluation

To evaluate the performance of machine learning methods, we define the following metrics based on the true positive (TP), true negative (TN), false positive (FP) and false negative (FN):

- Accuracy: the probability that the identification of a condition is the same as the ground truth.
- Precision: the probability that the identifications for NLoS conditions are exactly NLoS conditions in the ground truth.
- Recall: the probability that all NLoS conditions in the ground truth are identified as NLoS conditions.
- False positive rate (FPR): the probability that a LoS condition is identified as a NLoS condition.

We first evaluate the learning results under different scenarios. For more robust model evaluation, we adopt the cross-validation scheme to validate training models. Specifically, for each data set, that is, $\mathcal{H}$ with 16425 samples, $\mathcal{S}$ with 16033 samples and $\mathcal{U}$ with 27439 samples, we first split them into 10 subsets, then cross validate the learning models by using the i-th subset, for $i \in \{1, 2, ..., 10\}$, as the validation set and the remaining subsets together as training sets. Figure 5a shows the accuracy of NLoS detection under different scenarios with the NB and SVM methods. We have the following two main observations. First, both the NB and SVM methods can achieve superb accuracy values. For instance, with the NB method, accuracy can reach about 92.5 percent, 96.9 percent and 97.4 percent in highway, suburban and urban, respectively, while with SVM, the values can be about 93.7 percent, 98.3 percent and 98.3 percent, respectively. Second, the performance of SVM can slightly out-

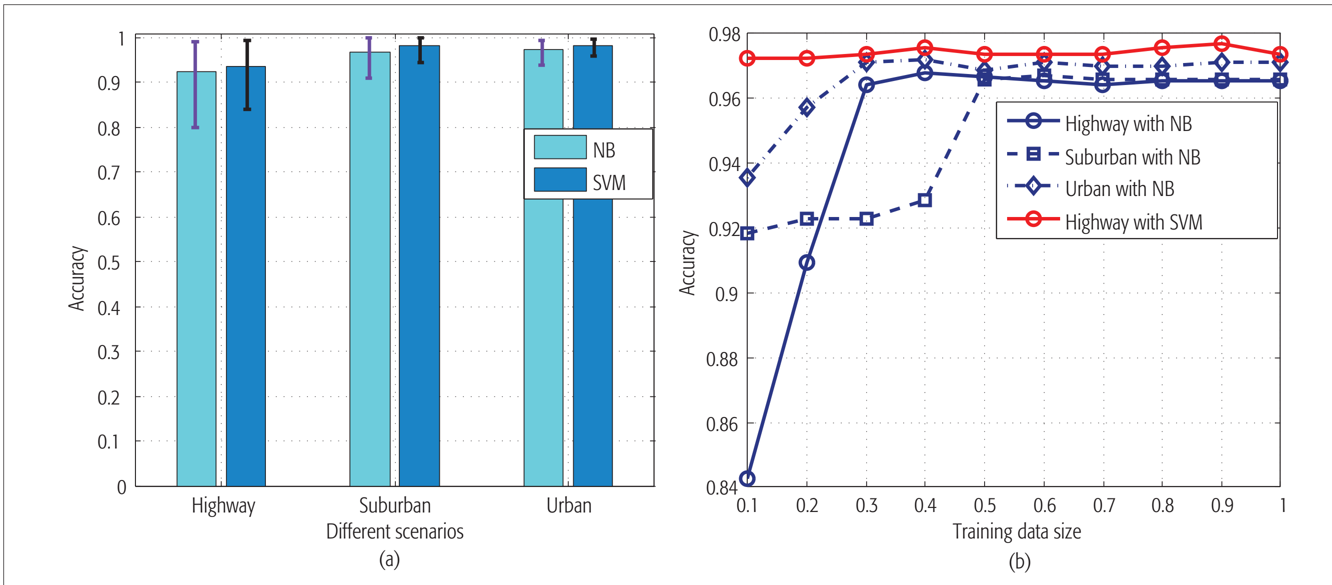

FIGURE 5. NLoS detection accuracy: a) accuracy under different scenarios; b) accuracy v.s. traing data size.

Scenarios	Accuracy (%)		Precision (%)		Recall (%)		FPR (%)	
	NB	SVM	NB	SVM	NB	SVM	NB	SVM
Highway	0.9247	0.9367	0.9578	0.9393	0.9359	0.9832	0.0606	0.0189
Suburban	0.9690	0.9831	0.9958	0.9867	0.9715	0.9943	0.0280	0.0035
Urban	0.9735	0.9828	0.9925	0.9851	0.9773	0.9971	0.0216	0.0037

TABLE 1. Learning results.

perform the performance of NB. Table 1 shows other metric values and similar observations can be obtained.

With the accuracy promise, we then investigate the robustness of the learning models, that is, the performance of the models with different sizes of training data. We first split each sample set into two subsets, one subset (occupying 10 percent proportion) as the validation set and the other subset (occupying 90 percent proportion) as the training set. The training set is evenly split into 10 subsets and for j-th training, for $j \in \{1, 2, ...10\}$, the union of the first to j-th subsets behave as the training set. Figure 5b shows the accuracy of NLoS detection with different sizes of training data under different scenarios. We have the following two main observations. First, for the NB method, to achieve very high accuracy, it requires high diversity training data to cover all situations in the validation set. When the accuracy performance reaches a supreme value (about 96.5 percent in the figure), increasing the training size cannot further improve the performance. For instance, in the highway scenario, the accuracy increases from 84.3 percent to 90.9 percent then to 96.4 percent with training data size 0.1, 0.2 and 0.3, respectively; with more training data, the accuracy will oscillate around 96.4 percent. It is noted that different subsets of training data may have varied impacts on the model performance, which explains why in the highway scenario, the accuracy increases significantly with the second and third subsets of training data, while in the suburban scenario, the accuracy increases more obviously with the fifth subset of training data. Second, the SVM method is not as sensitive to the training data size as the NB method is. For instance, the accuracy of SVM in the highway scenario oscillates around 97 percent regardless of the training data size. Similar observations can also be obtained in suburban and urban scenarios. As the results are tightly close to the highway results, which may confuse the figure, they are not shown in the figure.

Conclusions

In this article, we have discussed two important issues in VANETs in the big data era, that is, efficiently supporting big data through VANETs, and employing big data to improve VANETs. For the former, a framework combining a 5G cellular network and alternative opportunistic data pipes is introduced, and is envisioned to provide efficient, reliable, and flexible support of VANETs big data. For the latter, the mechanisms that analyze and learn typical big data for characterizing VANETs and designing intelligent protocols for VANETs are discussed. Furthermore, we have presented a case study in which urban VANETs measurement data is used to detect NLoS conditions through machine learning schemes.

Acknowledgment

This work is sponsored in part by the National Natural Science Foundation of China (NSFC) under Grant No. 91638204, and the Natural Sciences and Engineering Research Council of Canada.

References

[1] S. Al-Sultan *et al.*, "A Comprehensive Survey on Vehicular Ad Hoc Network," *Elsevier J. Netw. Comput. Appl.*, vol. 37, 2014, pp. 380–92.

[2] N. Lu *et al.*, "Connected Vehicles: Solutions and Challenges," *IEEE Internet Things J.*, vol. 1, no. 4, 2014, pp. 289–99.

[3] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Springer Mobile Netw. Appl.*, vol. 19, no. 2, 2014, pp. 171–209.

[4] P. Bedi and V. Jindal, "Use of Big Data Technology in Vehicular Ad-Hoc Networks," *Proc. IEEE ICACCI*, Delhi, India, 2014, pp. 1677–83.

[5] M. Maurer *et al.*, *Autonomous Driving: Technical, Legal and Social Aspects*, Springer, 2016.

[6] M. Hadded *et al.*, "TDMA-Based MAC Protocols for Vehicular Ad Hoc Networks: A Survey, Qualitative Analysis, and Open Research Issues," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, 2015, pp. 2461–92.

[7] W. Quan *et al.*, "Enhancing Crowd Collaborations for Software Defined Vehicular Networks," *IEEE Commun. Mag.*, vol. 55, no. 8, 2017, pp. 80–86.

[8] J. G. Andrews *et al.*, "What Will 5G Be?" *IEEE JSAC*, vol. 32, no. 6, 2014, pp. 1065–82.

[9] 3GPP, "3GPP TS 22.186: Enhancement of 3GPP support for V2X scenarios," *3GPP*, Tech. Rep., June 2017.

[10] V. Bychkovsky *et al.*, "A Measurement Study of Vehicular Internet Access Using In Situ Wi-Fi Networks," *Proc. ACM MobiCom*, Los Angeles, USA, Sept. 2006, pp. 50–61.

[11] N. Cheng *et al.*, "Opportunistic Spectrum Access for CR-VANETs: A Game-Theoretic Approach," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, 2014, pp. 237–51.

[12] H. Zhou *et al.*, "TV White Space Enabled Connected Vehicle Networks: Challenges and Solutions," *IEEE Netw.*, vol. 31, no. 3, 2017, pp. 6–13.

[13] C. Celes *et al.*, "Improving VANET Simulation with Calibrated Vehicular Mobility Traces," *IEEE Trans. Mobile Comput.*, vol. 16, no. 12, 2017, pp. 3376–89.

[14] C. F. Mecklenbrauker *et al.*, "Vehicular Channel Characterization and Its Implications for Wireless System Design and Performance," *Proc. IEEE*, vol. 99, no. 7, 2011, pp. 1189–1212.

[15] F. Lv *et al.*, "An Empirical Study on Urban IEEE 802.11p Vehicle-to-Vehicle Communication," *Proc. IEEE SECON*, Lon don, UK, June 2016, pp. 1–9.

Biographies

Nan Cheng [S'12, M'16] received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo. He is currently working as a post-doctoral fellow with the Department of Electrical and Computer Engineering, University of Toronto, and with the Department of Electrical and Computer Engineering, University of Waterloo. His research interests include performance analysis and opportunistic communications for vehicular networks, unmanned aerial vehicles, and cellular traffic offloading.

Feng Lyu received the B.S. degree in software from Central South University in 2013. He is pursuing his Ph.D. degree in the Department of Computer Science and Engineering at Shanghai Jiao Tong University. From October 2016 to October 2017, he was a visiting Ph.D. student at the Broadband Communications Research (BBCR) group in the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada. His research interests include vehicular ad hoc networks and cloud computing.

Different subsets of training data may have varied impacts on the model performance, which explains why in the highway scenario, the accuracy increases significantly with the second and third subsets of training data, while in the suburban scenario, the accuracy increases more obviously with the fifth subset of training data.

Jiayin Chen received the B.E. degree and the M.S. degree from the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China, in 2014 and 2016, respectively. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. Her research interests are in the areas of vehicular communication networks, with a current focus on intelligent transport systems, reinforcement learning, and big data.

Wenchao Xu received the B.E. and M.E. degrees from Zhejiang University, Hangzhou, China, in 2008 and 2011, respectively. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. In 2011, he joined Alcatel Lucent Shanghai Bell Co. Ltd., where he was a software engineer for telecom virtualization. His interests include wireless communications with an emphasis on resource allocation, network modeling, and mobile data offloading.

Haibo Zhou [M'14, SM'18] received the Ph.D. degree in information and communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. From 2014 to 2017, he was a post-doctoral fellow with the Broadband Communications Research Group, ECE Department, University of Waterloo. Currently, he is an associate professor with the School of Electronic Science and Engineering, Nanjing University. His research interests include resource management and protocol design in cognitive radio networks and vehicular networks.

Shan Zhang [S'13, M'16] received her Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2016. She is currently an assistant professor in the Department of Computer Science and Technology, Beihang University, Beijing, China. She was a post doctoral fellow in Department of Electronical and Computer Engineering, University of Waterloo, Ontario, Canada, from 2016 to 2017. Her research interests include mobile edge computing, wireless network virtualization and intelligent management.

Xuemin (Sherman) Shen [F] is a university professor, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on resource management, wireless network security, social networks, smart grid, and vehicular ad hoc networks. He is an IEEE Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, and a Royal Society of Canada Fellow. He was a Distinguished Lecturer of the IEEE Vehicular Technology Society and the IEEE Communications Society.