Multi-Dimensional Affinity Propagation Clustering Applying a Machine Learning in 5G-Cellular V2X

TAKASHI KOSHIMIZU©, (Member, IEEE), SHI GENG TIAN, HUAN WANG, ZHENNI PAN©, (Member, IEEE), JIANG LIU©, (Senior Member, IEEE), AND SHIGERU SHIMAMOTO©, (Member, IEEE)

Graduate School of Fundamental Science and Engineering, Waseda University, Tokyo 169-8050, Japan

Corresponding author: Takashi Koshimizu (koshimizu@nttdocomo.com)

This work was supported by the Department of Computer Science and Communications Engineering, Faculty of Science and Engineering, Waseda University.

ABSTRACT Cellular systems are facing the ever-increasing demand for vehicular communication aimed at applications such as advanced driving assistance and ultimately fully autonomous driving. Cellular Vehicle to Anything (C-V2X) has become more applicable with the release of the first sets of 5G (5th Generation) system specifications. The highly capable 5G systems will therefore support even a larger number of moving objects. This study aims to present a sophisticated clustering mechanism that enables cellular systems to accommodate a massive number of moving Machine Type Communication (MTC) objects with a minimum set of connections while maintaining system scalability. Specifically, we proposed Normalized Multi Dimension-Affinity Propagation Clustering (NMDP-APC) scheme and applied it for Vehicular Ad hoc Network (VANET) clustering. For VANET clustering formation, our study employed Machine Learning (ML) to determine the granularity, i.e., the size and span of clusters desirable for use in dynamic motion environments. The study achieved a sufficient level of prediction accuracy with fewer training data through a learned prediction function based on the selected key criteria. This paper also proposes a system sequence designed with a series of procedures fully compliant with C-V2X systems. We demonstrated substantial simulations and numerical experiments with theoretical analysis, specifically applying soft-margin-based Support Vector Machine (SVM) algorithm. The simulation results confirmed that the granularity parameter we applied fairly controls the size of VANET clusters although vehicles are in motion and that the prediction performance has been adjusted through controlling of key SVM parameters.

INDEX TERMS 5G mobile communication, vehicular ad hoc networks, machine-to-machine communications, machine learning, mobile computing, clustering methods, heterogeneous networks.

I. INTRODUCTION

The concept of connecting vehicles to anything (V2X) opens a new paradigm in vehicle transportation, leveraging the power of wireless communications [1]. It is easy to imagine that the way of our daily life will be greatly changed even with realization of one of the potential scenarios, e.g., autonomous driving. For realizing such highly demanded applications and use cases of V2X, aggressive inter-industry discussions have been underway. 5G Automotive Association (5GAA), a newly created cross-industry organization of automotive and telecommunication, has been working on next-generation connected mobility and automated vehicle solutions, showing strong initiative and collaboration toward V2X service realization [2]. IEEE 802.11p-based Dedicated Short Range Communication (DSRC) technology is already available for inter-vehicular communication [3]. In response to more advanced demands and requirements and various scenarios of V2X [4], a global standard called C-V2X has been recently specified by 3GPP [5], [6] with the 5G network systems [7], [8]. C-V2X has the capability to provide both short- and long-range communication modes that work interactively to fulfill various scenarios. Scenarios of short-range direct connectivity include vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I) and vehicle-to-pedestrian (V2P) communications. The short-range mode works independently of
cellular systems in the dedicated ITS 5.9 GHz spectrum band [9]. Long-range network connectivity (V2N) of C-V2X provides broadband and inter-system communication capability through a Public Land Mobile Network (PLMN), i.e., a cellular network via LTE and advanced 5G access [10], [11]. Typical demand for low latency and broadband connectivity to V2N stems from the demand for vehicular edge computing. Automotive Edge Computing Consortium (AECC) [12] has been established specifically to drive vehicular computing infrastructure for forthcoming automotive big data applications, intelligent drive management, acquisition of 3D map data and much safer driving assistance through reliable and real-time constant connectivity between the edge cloud system and vehicles [13].

For consistent system connectivity and quality, scalable access management is essential as it helps systems to avoid access storms caused by a massive number of MTC objects. Widely available cellular networks are susceptible to access stagnation caused by spontaneous access attempts of a large number of vehicles, moving drone objects and dynamic IoT devices. The clustering capability of VANETs largely contributes to reducing the number of direct connections to cellular networks. Such clusters formed by a set of vehicles connected via short-range communication can mitigate access storms by grouping member vehicles and eventually provide further access stability for PLMNs through such aggregation effects [14].

Our study aims to identify a valid VANET clustering scheme which able to meet such demands effectively. In the process of VANET cluster formation, a number of considerations need to be taken into account since vehicles are in motion. Fig. 1 depicts an image of VANET cluster formation and its reformation due to motion dynamics. Once a cluster is formed through message exchanges via inter-vehicular communication, the identified Cluster Head (CH) becomes a proxy toward the PLMN as shown in Fig. 1(a). At time $t + n$ in Fig. 1(b), $t$ and $n \in \mathbb{R} > 0$, the form of the original cluster has changed due to further dynamics. A clustering scheme used in such dynamic environments therefore should be able to effectively adapt to the dynamics of such constantly changing traffic. To address this requirement, we selected Affinity Propagation (AP), a mathematically well-proven and highly reliable scheme and applied its logic to VANET clustering. By employing the fundamental concept of AP, this study extended its key decision function in order to find out a scheme highly adaptable to the motion dynamics of VANET in real time.

To identify an effective scheme, we had to determine a desirable VANET clustering size by exploring appropriate criteria and conditions. Specifically, we employed a concept of ML to dynamically determine the cluster size based on minimum criteria. Namely, the aim of this study is to propose a novel vehicle clustering scheme applying AP to adapt to real-time motion dynamics of target vehicles by extending previous work [15]. Moreover, this study employed a ML scheme for deducing adequate clustering granularity in consideration of the key attributes for 5G Cellular V2X.

The contributions of this paper are summarized as follows:
- Vehicle clusters are autonomously formed through NMDP-APC process in consideration of real-time motion dynamics;
- The cluster granularity is designed to be derived from a ML algorithm with minimum criteria;
- Traffic density and cellular congestion states are specifically considered for granularity determination;
- A fully distributed system is designed with its entire procedure and sequence in line with 5G C-V2X; and
- Numerical evaluation and substantial simulations are performed by reflecting theoretical analysis.

The proposed scheme is applicable not only to vehicular communication but also to a wide range of other moving MTC objects including a large number of mobile robots and emerging drone-like objects aiming for future applications [16]. This study therefore must have a large potential to be usefully applied to a wide variety of mobile MTC applications.

The rest of this paper is organized as follows. Section II provides a review of the related works. Section III explains the proposed VANET clustering scheme applying NMDP-APC to message sequences. Section IV introduces our granularity determination scheme employing soft-margin-based SVM machine learning to find an appropriate cluster size. This section also provides mathematical analysis of SVM. Finally, Section V provides sets of simulation results of NMDP-APC clustering and the prediction accuracy of granularity with different key SVM parameters. Finally, Section VI concludes this study by summarizing the findings.

II. RELATED WORKS
A. VANET CLUSTERING TECHNIQUES
Identifying a valid clustering technique is an important research subject for VANETs not only to organize and manage VANETs but also to efficiently integrate a short-range vehicle network with a wide-range network. Vehicular clustering offers various benefits such as dynamic topology stabilization, improvement of routing efficiency and minimization of control overhead, and specifically provides scalability by limiting the number of accesses to a PLMN [14], [16]. Clustering improves system stability by localizing the control
targets and reducing signals for avoiding global propagation. Additionally, as VANET clustering reduces the impact of dynamic topology changes, the structure of network system becomes more manageable and stable [16]. An overall taxonomy of clustering techniques and protocols in various VANET applications in each category are well summarized in [17]. An extensive survey on clustering techniques is conducted by Cooper et al. [18], which analyses a variety of VANET clustering techniques to address such points as cluster head selection, cluster affiliation, and cluster management.

The selection metrics of VANET clustering scheme observed vary: some schemes use single-metric and others multi-metrics. Such variations can be seen in [17], [18]. Specifically, mobility-based clustering which focuses on inter-vehicular mobility dynamics is identified as a key direction in contemporary study. A reference [19] outlines a mechanism of VANET-UMTS integration and proposes VANET dynamic clustering mechanism using specific multi-metrics for finding a minimum number of vehicular CHs. Although it proposes a comprehensive mechanism targeting a dynamic mobility environment, its realization would be challenging as it requires the inclination angles toward eNBs and between vehicles. Studies on a multi-hop clustering scheme with IEEE 802.11p and 4G hybrid systems are introduced in [20], [21].

For CH selection, this scheme proposes a relative mobility metric calculated from the averaged relative speed with respect to the neighbor vehicles. However, the proposed procedure will require a large state management effort using a specific state transition matrix to address the needs of considering multi-hop inter-vehicular relations.

Generally, VANET clustering schemes can be categorized as either centralized- or distributed-clustering schemes. In centralized clustering schemes, cluster formation is performed via a NB or R ode Side Unit (RSU) with periodical message exchanges to the target vehicles. Qi et al. [22] proposed an SDN-based centralized clustering scheme by exploiting a social pattern, i.e., knowing the vehicles’ traffic pattern to deduce an expected route, where the metrics of inter-vehicular distance, relative speed and vehicle attributes are used. In distributed clustering schemes, cluster formation and CH selection are performed only between the vehicles through message exchanges without involving managements of central node in a fully distributed manner. The centralized clustering approach is advantageous in that it has more computational resources and managements available in the central entity. On the other hand, from the standpoint of nomadicity of vehicular mobility, the distributed clustering approach has advantages that it can avoid signaling concentration and a single point of failure; specifically, it allows autonomous decision making among widely spanned vehicular targets.

In contrast to those conventional clustering schemes, an advanced, fully distributed VANET clustering scheme was proposed [23] by Hassanabadi et al. [24] and by Shahwani et al. [25] based on a mathematically well-proven concept called Affinity Propagation, which was originally published in Science by Frey and Dueck [26]. In these initial Affinity Propagation Clustering (APC) applications [24], however, the metrics used in the similarity function are vehicles’ distances in the future prediction time point set to 30 sec, which means individual vehicles’ positions are forecasted in 30 sec later. It is obvious that it may be subject to large prediction errors as the vehicles are in motion. The scheme described above uses a message to exchange APC information every second. This messaging process need to be preliminarily iterated around 10 times, which requires 10 sec for clustering formation. This value is not trivial as vehicles are in motion and their locations could change by more than 200 meters during the elapsed time in a high mobility environment such as on a highway.

In contrast to the conventional application of APC, our proposed NMDP-APC in C-V2X is designed to target the following three key advantages. Firstly, our proposed scheme uses the vehicle’s current position and velocity that are not associated with any predicted value. Secondly, our scheme autonomously deduces a desirable clustering granularity not only by knowing vehicles’ motion dynamics but also by incorporating the traffic density and access congestion status by applying ML. Thirdly, it leverages the power of C-V2X technology for more improved performance, i.e., a short uplink transmission interval of 0.1 sec [27] with higher reliability and lower latency [9]. In a system-wide perspective, we designed our scheme by applying mobility-based metrics, specifically the position and velocity, in single-hop clustering formation, taking advantage of C-V2X wireless, which spans enough in one hop, and a distributed-clustering scheme, which totally fits vehicles’ nomadicty.

Reviewing traffic models is an additional field that we surveyed to understand the behavior of automotive traffic dynamics. Studies in this field have been traditionally performed mainly in civil engineering and traffic science for analysis and design of safe and efficient real traffic mechanisms. Gazis-Herman-Rothery (GHR) car following model [28], which we applied in our work, is a well-studied and sufficiently proven model. We applied the traffic model in our simulation by referring to the mathematical explanation shown in [29]. This model is particularly useful as it enables us to obtain details of vehicles’ motion dynamics, such as acceleration, velocity and position information of every moment. Since the data acquisition of dynamically moving objects is sometimes limited, the aforementioned GHR car following model is significantly helpful and the details obtained through the use of the model are applied onto real traffic data, which will be demonstrated in Section V.

B. ML FOR VEHICULAR COMMUNICATION

Key salient features of vehicular communication to be applied to succeed in 5G systems are summarized by Shah et al. [30]. The importance and mechanism of 5G Mobile Edge Computing (MEC) with VANETs are analyzed in [31] and Ning et al. [32]. Low latency is an essential key capability for VANET networks, especially when the target application
involves unmanned vehicles that generally require ultra-low latency capabilities of the system. MEC can process and deliver data quite efficiently by pushing the computing resources to the Mobile Vehicular Cloud (MVC) located at the edge of the network as shown in [33] and Ma et al. [34]. It is also notable that a game theory for VANET clustering is employed by Khan et al. [35] to find a balance of total throughput capacity in consideration of the clustering granularity, i.e., population share. In [35], the RSU is used as a centralized controller to signal an overall observation, compute the average payoff of the entire population of clusters and then broadcast it to all clusters.

Potential application of ML over vehicular networks was investigated by Ye et al. [36]. A framework for ML on vehicular networks was reviewed by Liang et al. [37]. Tian et al. proposed an application of a Hebb neural network, which intelligently learns the topology and forms vehicle clusters [38]. An application of Q-learning for cluster head identification is introduced by Khan et al. in the proposal of two-level clustering [39]. Level-1 CHs are selected by a fuzzy logic algorithm using three metrics: relative velocity, multiple-connectivity factor and link-reliability factor. Level-2 CHs are determined by an improved Q-learning theory as an optimal gateway to the 5G node-B. However, no clear proposal has been presented as to how an identical number of clusters can be identified in these studies.

Support Vector Machine (SVM) is one of the well-known classification schemes in ML, which enables categorization of data into several classes in a sophisticated manner. An organized research on classification mechanism of SVM was provided by Junli and Licheng [40]. The fundamental logic and mathematical model of SVM are well explained in [41]–[43]. The application of SVM for the purpose of data-offloading in a VANET was provided by Wu et al. [44], proposing the use of the SVM classification scheme for deciding divided tasks to be executed either locally or for offloading. Application of ML to VANET clustering still seems to be in the advent phase. Hence in this study, we employed the power of SVM specifically for the determination of the VANET granularity in fitting the characteristics of nomadic vehicle dynamics, which represents our novelty.

III. PROPOSED VANET CLUSTERING SCHEME

A. AFFINITY PROPERGATION FUNDAMENTAL

We applied AP to VANET clustering in knowing the vehicles’ short-range communication capability that enables information exchange for clustering. Such inter-vehicular communication is an essential function of APC as clusters are formed through iterative message exchanges among the surrounding neighbors. Eventually, the scheme identifies the best set of clusters by recursively exchanging specific messages until a set of exemplars, i.e., a CH is identified through evaluation of similarities via a specific function.

In APC, a similarity measurement of each data pair between the points \( k \) and \( l \), denoted \( s(k, l) \), is iteratively evaluated, and thereby exemplars emerge, where \( k = (1, 2, \cdots, K) \), \( l = (1, 2, \cdots, L) \), \( k \) and \( l \) \( \in \mathbb{N} \). This process allows each data point to be a possible exemplar, which will exchange messages with other exemplars until a set of high-quality exemplars emerges. Two kinds of messages are exchanged between the points, i.e., vehicle \( k \) and \( l \). First, a responsibility message \( r(k, l) \) sent from \( k \) to \( l \) represents how adequate vehicle \( l \) is as an exemplar when other competing exemplars are taken into account. This is deduced from similarity evaluation (1). The responsibility message \( r(k, l) \) and the self-responsibly message \( r(l, l) \) in each iteration are given by:

\[
r(k, l) \leftarrow s(k, l) - \max_{l' \neq l} \{ a(k, l') + s(k, l') \} \quad \forall k, l \tag{1}
\]

\[
r(l, l) \leftarrow s(l, l) - \max_{l' \neq l} \{ s(l, l') \} \quad \tag{2}
\]

Second, an availability message \( a(k, l) \) reversely sent from the exemplar \( l \) back to the point \( k \) shows \( l \)'s suitability to be an exemplar for \( k \), considering the feedback from other competing potential exemplars. The availability message \( a(k, l) \) and the self-availability message \( a(l, l) \) in each of the iterations are given by:

\[
a(k, l) \leftarrow \min \left\{ 0, r(l, l) + \sum_{k' \neq k, l} \text{max} \{ 0, r(k', l) \} \right\} \quad \forall k, l \tag{3}
\]

\[
a(l, l) \leftarrow \sum_{k' \neq l} \text{max} \{ 0, r(k', l) \} \quad \tag{4}
\]

In the iterative process, the responsibility and availability messages are processed with a damping factor \( \lambda \) to avoid computational oscillations. This is done by \( \text{message}_{\text{new}} = \lambda \cdot \text{message}_{\text{old}} + (1 - \lambda) \cdot \text{message}_{\text{new}} \), where \( \lambda \) is a damping factor whose value is assigned between 0 to 1 [45]. Once the value of net-similarity converges, the APC process terminates its iteration. Eventually, a specific data point is identified as a CH among the cluster members, and this is determined when the data point’s self-responsibility and self-availability meets the following condition for the point \( k \) if \( r(k, k) + a(k, k) > 0 \), then the point \( k \) is an exemplar, or for the point \( l \) if \( r(l, l) + a(l, l) > 0 \), then the point \( l \) is an exemplar. By obtaining the result values of responsibilities and availabilities from the iterations, exemplars are identified. Upon convergence, each node \( k \)'s CH is identified by:

\[
CH_k = \arg \max_l \{ a(k, l) + r(k, l) \} \tag{5}
\]

The identified CH acts as a gateway for the VANET cluster members, which greatly contributes to the reduction of the number of connections to a PLMN.

B. NMDP-APC CLUSTERING

Having the fundamental concept of APC, our proposed VANET clustering scheme NMDP-APC enhances its similarity function by combining the Euclidean special distance, i.e., the vehicle’s position information along with its velocity.
information in normalizing each term. We identified the most critical metrics that represent motion dynamics of a VANET are the vehicle’s current position and velocity [19], [21]. Some studies consider the link strength for a clustering criterion with multiple wireless hops. Instead, we designed our system in considering the inter-vehicle distance and relative velocity in the formation of single wireless-hop clusters. This results in a firm VANET cluster stability and the inter-vehicular distance that can be considered proportional to the radio strength of a pair of vehicles. Therefore, we designed our similarity function by associating these two essential terms into a single formula as shown in (6). It is formed by a combination of the negative Euclidean distance of each vehicle’s current position with the negative difference of each vehicle’s current velocity. In order to associate the two terms in a single similarity function, each term has been normalized from 0 to 1, as denoted “nor” in (6). We also assigned the weighted parameters $\varphi$ and $\psi$ on the normalized components as the control nobs of each term. This capability provides significant extendibility since it allows multiple dimensional terms to fit into a single similarity function with adjustment of the proportion given by the weight parameters.

We also assume that this scheme is mainly applied to the uplink connection to a PLMN. Eventually, our enhanced similarity function is shown as:

$$s(k, l) = (\varphi \|x_k - x_l\|_{nor} + \psi \|v_k - v_l\|_{nor})$$  \hspace{1cm} (6)

where $\varphi + \psi = 1$

$$x_k = \begin{bmatrix} x_k \\ y_k \end{bmatrix}, \quad v_k = \begin{bmatrix} v_{x,k} \\ v_{y,k} \end{bmatrix}$$  \hspace{1cm} (7)

where $x_k$ is a vector of the node $k$’s current position in the $x$ and $y$ coordinate and $v_k$ is a vector of the node $k$’s current velocity in the $x$ and $y$ coordinate. Note that in contrast to the previous study [24], our proposed similarity function in NMDP-APC does not compose any future prediction value. This helps avoid any associated prediction errors, which is a key difference from related previous works. The following subsection explains the sequence of clustering scheme we enhanced over the PC5-based C-V2X message sequence.

### C. PROPOSED CLUSTERING SEQUENCE

Through message exchanges between the points, the similarity function identifies the homogeneity of vehicles’ motion dynamics and then decides the formation of sets of clusters. In the C-V2X direct communication mode, the members use the PC5 interface designed for direct V2V communication. The vehicles communicate with the neighbors of C-V2X PC5 radio range. All required APC messages are sent on the application layer over the short-range interface, which is considered approximately a 500 m to 1 km radius in reliable communication [27]. As previously shown in Fig. 1, cluster formation takes place at Time $= t$ and at an interval of time later. At Time $= t$, a cluster is formed based on the APC clustering scheme and identified as a CH with the members denoted as vehicle-B and C in the diagram.

Reflecting the vehicles’ motion dynamics, the associated members are updated at Time $= t + n$ as in Fig. 1(b). In this case, although the CH has not changed, some members have been changed: vehicle-C has disappeared from the original wireless coverage and vehicle-D has become a new member because of its similarity evaluated by the clustering scheme. In this diagram, vehicle-E stays independent from the cluster as it has an intention to be a stand-alone or the dynamics makes it less likely to be a member of the cluster.

Now, Fig. 2 shows the enhanced signaling sequence exchanged in the above scenario. VE represents the Vehicle-UE that complies with 3GPP C-V2X scheme. Step 1 indicates the provisioning process for one-to-many and PC5-based communication between the VEs. The provisioning includes three important aspects: a) identifying authorization policies and parameters; b) setting up V2X direct communication parameters; and c) setting up radio parameters for not-served by 5G-RAN. The authorization policies and parameters include UE’s authorization parameter from a PLMN used to prevent malicious UEs from engaging. The V2X direct communication parameters include V2X-Layer2-ID, IPv4/v6 preferences and Application Layer IDs, which are used by the VEs to perform one-to-many communications. Radio parameters for VEs not-served by 5G-RAN provides the capability for one-to-many V2X communication in considering geographical dependencies of frequency bands. In Step 2, each VE starts to exchange messages to surrounding vehicles by looking at the destination and source V2X L2-IDs on a specific C-V2X short-range communication link. Step 3 is the message exchange of NMDP-APC process using one-to-many PC5 direct communication among the neighbors. Clustering Request is a specific message sent from a VE intending to form a cluster. Such request message can be sent from any of the neighbors such as VE-A, B and C. Triggering reception of the Clustering Request message, these vehicles initiate the message exchange of NMDP-APC. Step 4 indicates the results of clustering formation and the CH identified after the APC process for the cluster indicated in Fig. 1(a). When a VE becomes a member of the cluster,
the VE will enter In-Clustering State and then inform neighbors that it has become a member of the cluster. Sometime later as in Fig 1(b), VE-D initiates Step 6 and 7, which are identical to Step 1 and 2 respectively, and it also intends to form a cluster. After the provisioning process, an additional Clustering Request messages is sent from VE-D as shown in Step 8, and then an updated cluster is formed via additional NMDP-APC process considering the dynamics of the neighbors. Note that VE-E remains stand-alone as it has independent dynamics or it has an intention to be in standalone communication mode. All parts shown colored in-red are the specific enhancements that we proposed over the C-V2X original sequence [46]. The Fig. 2 sequence covers most of the scenarios by which a new VE joins in and evacuates from a cluster. When CH needs to be changed or replaced, V2X-to-NW Relaying function can be applied to CH relocation process between the new and old VEs, which will be explained later in Sec IV-C with its the procedures.

We have so far explained the fundamental mechanism of NMDP-APC, referring to a scenario that takes account of VE’s motion dynamics. We have yet to offer explanation, however, for the question as to how the clustering granularity should be determined. We employed a ML scheme to challenge the question while vehicles are dynamically in motion. The following section will explain the ML scheme, starting from its concept.

IV. GRANULARITY DETERMINATION BY SVM ML

A. OVERVIEW OF VARIATIONS OF ML

ML enables computers to identify hidden insights through iterative learning of given data sets. Table 1 summarizes the variations of ML in terms of categories, objectives, algorithms and application examples in communication fields [36]. In a large perspective, each ML application is classified as either a supervised or unsupervised learning scheme. In this study, we chose a supervised scheme because the training data are fairly obtained and it enables us to evaluate the prediction results by comparing with those given from the decision tree. In comparison with SVM, Neural Network (NN) is generally considered more applicable when the dimension of the explanatory variable is large. In addition, SVM is known to enables the tuning of prediction accuracy by the limited number of parameters, which are $\gamma$ and $C$ parameters. Reflecting these properties, we decided to employ SVM to identify the clustering granularity in this study.

In general, the process of supervised ML consists of two stages: training and testing. In the training stage, a model is learned based on a set of pre-prepared training data. Once the function is educated by the training data, the trained function is applied for providing predictions in the testing stage. The next sub-section explains the details of the task vector and the decision tree we propose, both of which are the essential elements of SVM.

B. PROPOSED TASK VECTOR AND DECISION TREE

It is generally known that obtaining an applicable trained decision function $f(X)$ requires a certain volume of training data. In this study, a predefined decision-tree is applied for identifying the level of clustering granularity, i.e., an objective variable $y$ from a set of explanatory variables, which consists of the task-vector $X$. SVM is known to be well applicable for its sufficient volume of data and a clear relation of explanatory variables to the objective variable [47].

With the mapping decision of the task-vector $X$ to an objective variable $y$, the classification is provided as $D = \{ (X, y) | y = \text{sign}(f(X)) \}$. In this study, we decided to classify the clusters in four levels based on the size: large, medium, small-medium, and small. In consideration of the obtained real traffic data and the span of C-V2X radio propagation, we set a four-level classification, although theoretically the number of levels can be changed to any. Numerically it can be shown as $D \in \{0, 1, 2, 3\}$ corresponding to the above four levels of cluster size. We designed the task-vector $X$ consisting of the following minimum set of explanatory variables for deciding the cluster size. We kept the number of elements to a minimum, otherwise it would have led to a longer computational time and potential deterioration of prediction performance because the target vehicles are in motion.

$$X = (I_v, F_p, T_d, Nb_c)$$

In (8), $I_v$ is the communication volume, which we assumed to be 100 kbps uplink data randomly generated in a certain interval from each vehicle. $F_p$ represents an initial preference as to which mode is used to establish a session: clustered mode or individual mode, which can be denoted as $F_p \in \{0, 1\}$. The preference value of $F_p = 0$ represents an intention of independent session establishment, in which case a session is established directly and connected to a cellular network. A broad-band data upload could be a use case. On the other hand, $F_p = 1$ represents an intention to form a cluster, in which case a session will be established via a CH. The third attribute constituting the task vector is the traffic density of vehicles shown as $T_d$. We obtained real traffic data from a highway and applied them in our simulation. According to the obtained traffic data, the maximum number of observed vehicles in a 2 km span, i.e., in a 1 km radius, was almost 60 in a single direction. The vehicle density was obtained from the number of vehicles in the observed range. The last

| Category | Objective | Algorithm | Application Ex. |
|----------|-----------|-----------|-----------------|
| Supervised | Classification | Neural Network, SVM | Anomaly detection, Server selection |
|          | Regression | Logistic Regression, SVR | Throughput prediction, channel parameter regression |
| Unsupervised | Clustering | K-means, Spectrum clustering | Congestion control, hierarchical routing |
|          | Dimension Reduction | Manifold learning | Data aggregation |
attribute constituting the task vector is the congestion status of a PLMN access, denoted as $Nb_c$, the set of which can be denoted as $Nb_c \in \{0, 1\}$, where $Nb_c = 0$ represents a normal access state and $Nb_c = 1$ represents a congestion access state. When a particular access is in the state of congestion, the number of clusters is made smaller, as aggregation effect contributes to reducing the number of the cellular access connections. This parameter is therefore factored in our task-vector.

By incorporating these explanatory variables in the task-vector $X$, we developed the following decision tree to process a set of training data that enables deduction of an appropriate size of clusters. Fig. 3 shows our proposed four-layer decision tree with an essential and minimum set of decision criteria to deduce the level of clustering granularity. The aforementioned NMDP-APC enables the formation of a desired clustering granularity by assigning a granularity parameter-$px$. The following sub-section introduces the overall clustering procedure and shows how the granularity parameter-$px$ is applied in the system perspective.

**FIGURE 3. Decision tree for clustering granularity determination.**

### C. PROCEDURE WITH ML DEDUCED px-VALUE

Once a $px$ value is given, NMDP-APC enables the control of the desired clustering granularity in the process. We employed a ML scheme to obtain the desired clustering size from the trained decision function $f(X)$ as this process identifies the desired $px$ value. Table 2 explains how NMDP-APC process uses the ML-deduced $px$-value segmenting three procedures parts. As Part-A has been mostly explained with the sequence in Section III-C, let us focus on the procedure in Part-B and C for the sake of paper space.

Part-B covers the procedure of an NMDP-APC re-clustering case for a member change. Step 06 and 07 is the provisioning steps identical with step 01 and 02, respectively. Step 08 initiates re-clustering process triggered by a Clustering Request Message (CRM) from a VE which has engaged in a clustered V2X wireless range. Step 09 starts another NMDP-APC process with a ML driven $px$-value to form a new cluster, aiming for cluster maintenance upon receiving the CRM. Step 10 is a process in staying original CH and the CH updates the associated members’ context information. Step 11 is a process triggered by a VE about to leave the cluster because of wireless range-out or re-clustering request reception from a member. Then, step 12 starts another NMDP-APC process with a ML-driven $px$-value to form a new cluster in the same manner as in step 09. The CH only maintain the context of updated cluster members. Part-C provides a set of procedures when initial CH replacement is required. In step 13, a CH receives a CRM message from a VE. Step 14 initiates Cluster Maintenance by executing another NMDP-APC process. In step 15, an original CH is recognized to be replaced and $CH_{NEW}$ is identified from the APC. Then, step 16 $CH_{OLD}$ and $CH_{NEW}$ exchange the contexts of associated members information via V2X-to-NW relay. $CH_{OLD}$ replaced by $CH_{NEW}$ by transferring the context of VE (k=1…K), $k \in V_{INST} \cap ACP$. Step 17 is an annex procedure that enables a recovery from an

**TABLE 2. Procedure of NMDP-APC with machine learning derived $px$-value.**

| Step | Part-A, Initial Provisioning & Cluster Formation |
|------|--------------------------------------------------|
| 01   | Initiate Provisioning request to V2X function in PLMN from those VE desired to cluster $V_{des} (des=0 \rightarrow 2)$ des $\in VVE$. (Authorization, DC parameters, Radio parameter configuration) |
| 02   | Exchange neighbors in setting 2VX L2 IP via C-V2X wireless link with those VEs are in the wireless rage. $V_{des} (inf=1 \rightarrow 2)$ inf $\in Z$. |
| 03   | If a Clustering Request Message is received from a neighbor. (any wireless propagation in coverage $V_{des}$ are potential to initiate) |
| 04   | Then initiate Cluster Establishment by executing NMDP-APC with trained $f(x)$ provided a $px$ value, $CH_{j}$ associated with the member $V_{E_{n}(u=1 \rightarrow 2)}$ j, and $u \in V_{INST} \cap ACP$. (Determined CHs and the members enter In-cluster mode) |
| 05   | Communicate with PLMN via identified CH. |
| 06-07| Identical provisioning process as in step 01-02 respectively. |
| 08   | If a Clustering Request Message is received from a neighbor. (any wireless propagation in coverage $V_{des}$ are potential to initiate) |
| 09   | Then initiate Cluster Maintenance by executing NMDP-APC with trained $f(x)$ provided a $px$ value, as in step 04. |
| 10   | If original CH stays, incoming VE becomes a cluster member. The CH updates the member, $V_{E_{n}(u=1 \rightarrow 2)}$ u, $\in V_{INST} \cap ACP$. (The CH relashes the member status and associated context.) |
| 11   | If an original member goes out the wireless range or received a Re-Clustering Request. |
| 12   | Then initiate Cluster Maintenance by executing NMDP-APC with trained $f(x)$ provided a $px$ value from the CH. $CH_{j}$ only has the instance of revised member $V_{E_{n}(u=1 \rightarrow 2)}$ u, $\in V_{INST} \cap ACP$. |
| 13   | If a CH received a Clustering Request Message from a VE. |
| 14   | Then initiate Cluster Maintenance by executing NMDP-APC with trained $f(x)$ provided a $px$ value. |
| 15   | If an original CH is recognized to be replaced and $CH_{NEW}$ is identified from the APC. |
| 16   | Then $CH_{OLD}$ and $CH_{NEW}$ exchange the contexts of associated members information via V2X-to-NW relay. $CH_{OLD}$ replaced by $CH_{NEW}$ by transferring the context of VE (k=1…K), $k \in V_{INST} \cap new ACP$. |
| 17   | If this process failed, enters step 03 then restart initial clustering, evacuating from abnormal state and returning to normal procedure. |
abnormal state to a normal state. Then, the following section introduces mathematical analysis of the soft-margin-based SVM classification model and the kernel operation which are specifically employed in this study.

D. SVM CLASSIFICATION MECHANISM
1) MECHANISM OF SOFT-MARGIN-BASED SVM
The fundamental concept of SVM originates from a neural model applying a linear-threshold logical unit, which eventually provides class discrimination. When a set of training data is given in \((X_i, y_i), \cdots, (X_M, y_M)\), \(1 \leq i \leq M, X \in \mathbb{R}^d, y \in \{\pm 1\}, M \in \mathbb{N}\), with the corresponding correct class-labels \(y_M\), the linear discriminate function can be denoted as

\[
y = f(X) = \text{sign} \left[ (W^T X - h) \right]
\]

\(X\) is a formerly mentioned task-vector consisting of explanatory variables. \(W\) is the corresponding synapse weight-factor, and \(h\) represent a threshold value. This model provides an output \(y \in \{\pm 1\}\). It returns \(+1\) when the inner product of the task vector \(X\) and weight-factor \(W^T\) exceeds the threshold, and it returns \(-1\) when it is less than the threshold. Geometrically, this concept is applicable as a classification system of input data, which segments data into two fields as depicted in Fig. 4(a). The figure shows two types of class-regions \(R_1\) and \(R_2\). They are distinguished by the corresponding hyper planes \(H_1\): \(W^T X - h = 1\) and \(H_2\): \(W^T X - h = -1\), which are formed by a limited number of double-circled support vectors. Notably, no data exist within the margin-area formed by the two hyper planes [42]. An optimal hyperplane \(H^*\) can be identified by maximizing the margin between the hyperplanes given by \(1/ ||W||\), which is an equivalent question to one for maximizing \(||W||^2\). This leads to the following constrained optimization problem:

\[
\min L(W) = \frac{1}{2} ||W||^2 \\
s.t. y_i \left( W^T X_i - h \right) \geq 1 \ (i = 1, 2, \cdots, M)
\]

This formula has been called soft-margin SVM as it has a specific cost-parameter \(C\). This parameter introduces additional control capability which enables adjustment of the balance between the amount of margin allocation and the amount of slack allowance [41]. In a graphical representation on Fig. 4(a), when a smaller \(C\) is assigned, the indicated margin becomes wider. It has been known that the solution of this optimization problem can be obtained by the saddle point of the Lagrangian [40]. Introducing two Lagrange multipliers \(\alpha_i \geq 0\) and \(\beta_i \geq 0\), reformulate the objective function into:

\[
L(W, h, \alpha, \beta) = \frac{1}{2} ||W||^2 + C \sum_{i=1}^{M} \xi_i \\
- \sum_{i=1}^{M} \alpha_i \left( y_i \left( W^T X_i - h \right) - (1-\xi_i) \right) \\
- \sum_{i=1}^{M} \beta_i \xi_i
\]

Lagrangian duality enables this primal problem to be transformed to the Wolfe-dual problem, which is given by:

\[
\max_{W, h, \xi, \alpha, \beta} \left\{ \min_{W, h, \xi} L(W, h, \xi, \alpha, \beta) \right\}
\]

By minimization with respect to \(W, h,\) and \(\xi_i\) of the Lagrangian \(L\), a partial derivative is taken on each of them for the following dual Lagrangian formula [48]:

\[
\max_{\alpha} W(\alpha) = \max_{\alpha} \left[ \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \alpha_i \alpha_j y_i y_j X_i^T X_j \right]
\]

\[
s.t. \left\{ \sum_{i=1}^{M} \alpha_i y_i = 0 \right. \\
| \alpha_i | \leq C \ (i = 1, \cdots, M)
\]
Then, the optimal weight vector $W^*$ can be given by:

$$W^* = \sum_{i=1}^{M} \alpha_i^* y_i X_i$$  \hspace{1cm} (20)

Optimal threshold $h^*$ can be given by:

$$h^* = y_j - \sum_{i=1}^{M} \alpha_i^* y_i (X_i^T X_j)$$  \hspace{1cm} (21)

Finally, the decision function is provided by:

$$f(X) = \text{sign} \left( \sum_{i=1}^{M} \alpha_i^* y_i (X_i^T X) - h^* \right)$$  \hspace{1cm} (22)

### 2) SVM KERNEL OPERATION

The aforementioned soft-margin SVM provided additional solution by introducing cost and margin allowance to solve non-linear separable test data. This approach still faces a limitation however, when the data distribution is fundamentally linearly inseparable, i.e., stars from dots, as shown in Fig. 4(b). To cope with such cases, a kernel operation has been introduced for further enhancement of SVM. In a case where a linear boundary is inappropriate, the kernel operation maps the input vector into a higher dimensional feature space. By choosing a non-linear mapping, SVM becomes capable of constructing an optimal hyperplane in a higher dimensional space $\Phi$, of which image is shown in Fig. 4(b).

$K(X_1, X_2)$ represents a kernel function which performs such non-linear mapping into a higher dimension feature space $\Phi$ as in $\Phi (X_1)^T \Phi (X_2) = K(X_1, X_2)$. Such kernel offers an efficient and less expensive way to transform data into a higher dimension. It allows operation in the original feature space without computing the data in the new dimension space [42], [49]. Polynomial, gaussian and sigmoid kernel have been known for practicality and computational advantages [40], [41]. Especially, Gaussian Radial Basis Function (GRBF) is one of the widely used SVM kernel types, which is given by following form:

$$K(X_1, X_2) = \exp \left( -\frac{\|X_1 - X_2\|^2}{2\sigma^2} \right)$$  \hspace{1cm} (23)

where the newly introduced $\gamma$ parameter controls the degree of linearity, i.e., the curvature of the hyperplane which is more flexibly able to fit to the distribution of sample data. Therefore, applying a leverage of kernel operation, it gives SVM control capability of the prediction performance by adjusting the both $C$ and $\gamma$ parameters. This is a large rationale of its wide application of SVM to solve various classification issues. With the kernel operation, the optimization problem of (18) can be written as:

$$\max_{\alpha} W(\alpha) = \max_{\alpha} \left[ \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j K(X_i, X_j) \right]$$  \hspace{1cm} (24)

with same constrains provided in (19). Once (24) is solved with constraints (19) by applying the same approach, it becomes possible to determine the Lagrange multipliers and a classifier with an optimal hyperplane in the transformed feature space.

$$f(X) = \text{sign} \left[ \sum_{i=1}^{M} \alpha_i^* y_i K(X_i, X) - h^* \right]$$  \hspace{1cm} (25)

These are the fundamental mechanisms which we employed in our classification scheme. The following section will discuss the performance evaluation describing sets of simulations conducted. We used GRBF kernel with SVM and observe the prediction performance by adjusting formerly explained parameter $C$ and $\gamma$.

### V. SIMULATION RESULTS

This section describes extensive sets of simulations we conducted in order to evaluate and verify the proposed VANET clustering scheme. Before explaining simulation results, let us introduce the concept and mechanism of the GHR car following model employed although we briefly touched upon it in Section II. A well-organized reference [29] shows how the GHR model predicts a follower’s acceleration $a_c$ using the following formula:

$$a_c^{n+1} = \frac{\epsilon_{l,m} \left( v_{n+1}^p - v_{n-1}^p \right)^m}{\left( v_n^p - v_{n-1}^p \right)} [v_n^p - v_{n+1}^p]$$  \hspace{1cm} (26)

for the $n^{th}$ vehicle ($n = 1, 2, \ldots$) at discreet time $t$, where $l$ represents a headway exponent, $m$ represents a speed exponent and $\epsilon_{l,m}$ represents a sensitivity coefficient of the GHR. From the derived acceleration, both $v_n^p$ and $v_{n-1}^p$ are identified from the Newton’s law of motion. The follower’s acceleration can be observed as if a spring has been placed between the ahead vehicle as in Fig. 5 depicted. According to [28], [29], we applied specifically the following parameters in the GHR model: $l = 1.0$, $m = 0$, $\epsilon_{l,m} = 18$ in order to perform simulations.

**FIGURE 5. Concept of GHR car following model.**

For this simulation, we captured real traffic data on a four-lane Inter City highway (I-294) in Chicago, Illinois, USA on a day in August 2017. From the sampled real traffic data, we obtained leading vehicles’ position and velocity information and then identified following vehicles’ traffic information via the aforementioned GHR car following model. A leading vehicle is identified with a sufficient distance from a vehicle in front, which eventually tends to drive at its own pace. In contrast, a following vehicle tends to follow the leading car’s driving pace as it has a limited headway distance. Using the data obtained with the above process, we traced each vehicle’s traffic data in four lanes at 5 sec intervals from 5 to 60 sec in one-minute. Then, the data were
processed in our NMDP-APC MatLab simulation to analyze how the proposed scheme identifies clusters with the CHs in the real traffic data.

**A. GRANULARITY PARAMETER px AND CLUSTERS**

The number or granularity of clusters can be controlled by varying the granularity parameter $px$ in NMDP-APC process. Fig. 6 shows the relation of the number of clusters formed with the changing the values in our simulation. The $px$ represents the degree of the multiplexity to the median of the distribution of similarity function in APC. The graph indicates how the cluster granularity can be changed when the value is adjusted from 0.1, 0.2, 0.5, 1.0, 1.5 and 2.0. The result clearly shows that the clusters are controllable through the adjustment of the value in the NMDP-APC. As mentioned in Fig. 3, when a small $px$-value is applied, a large number of small clusters are formed; in contrast, when a large $px$-value is applied, a small number of large clusters are formed. Now again, the $px$-value is deduced from the trained SVM prediction function and it is applied during the NMDP-APC according to the procedure shown in the Table 2. The following section shows more comprehensive sets of simulation results of NMDP-APC clustering at different observation time points.

*FIGURE 6. Number of clusters by changing px-value.*

**B. CLUSTERING FORMATION IN SIMULATION**

This sub-section shows the sets of results of NMDP-APC simulation conducted on Matlab to observe cluster formation for different granularity parameter $px$ values which derived from the proposed ML scheme. With one-minute real traffic data available at hand, we evaluated the clustering performance based on the data collected at different time points. Fig. 7-9 display the clustering results we obtained until the observation field is filled by using the traffic data collected at 35 sec, which is the middle part of the entire data. Fig. 10-12 display the clustering results with the data collected at 35 sec for different values of the granularity parameter $px$ where the observation space is entirely filled. Additional reason we chose the data applied at 35 sec is our consideration of the practical radio propagation span of C-V2X. We set the observation range to within 1 km, i.e., a 500 m radius, in this simulation as the reliable transmission range is considered to be a 500 m to 1 km radius [1], [27]. Fig. 13-15 display the clustering results applying the data at 60 sec for different granularity parameter $px$ values where the observation space is entirely filled. The 2 km observation
range, i.e., 1 km radius, seems to be a critical distance for reliable communication on PC5 based C-V2X [27].

Fig. 7 provides the result at $t = 15$ sec, the snapshot data from the overall observation in 35 sec. The top row in Fig. 7(a) shows the original vehicle distribution with the number of vehicles being indicated as $N = 16$. Fig. 7(b) shows the NMDP-APC clustering result with $px = 2.0$, which was assigned to produce a large group of fewest clusters. The proportion of NMDP-APC weight parameter $\varphi$ represents a position component and $\psi$ represents a velocity component, which are assigned as $\varphi = 0.8$ and $\psi = 0.2$ respectively, and the same values were applied throughout this simulation. The number of clusters formed in this condition is $Clusters = 3$. CHs are located mostly in the middle of each group with the
associated members linked by dotted lines. From the results, it is easy to understand that few large-sized clusters were produced because a large $px$-value has been assigned. In this granularity assignment, it can be observed that vehicles are forced to be clustered due to the extreme $px$-value applied. Fig. 7(c) shows the Net Similarity value in the APC iteration process. By reaching the stable net similarity value, the APC process identifies the final shape of stable sets of clusters. Fig. 8 and 9 provide the result sets applying the same $px$-value, then the NMDP-APC clustering status is observed at $t = 25$ and $35$ sec respectively during the overall observation time of $35$ sec. It can be observed that a higher number of vehicles come into the observed space, and the NMDP-APC is producing the same levels of granularity as the same $px$-value has been applied.

Fig. 10(b) displays the NMDP-APC clustering result obtained with the granularity parameter $px = 1.4$ assigned to form medium sized clusters on the data at $t = 35$ sec where the observation space is filled. The number of clusters has increased to Clusters $= 5$ and it clearly shows that the number of members in a cluster tends to be smaller. Similarly, it is observed that CHs are located mostly at the center of the group and linked with the members in dotted lines. The rest of parameters are assigned in the same manner. Fig. 10(c) provides the evidence that the clusters have reached stability. Fig. 11(b) displays the clustering result obtained from assigning the granularity parameter $px = 0.5$, which is targeted at forming small-to-medium sized clusters on the same data at $t = 35$ sec. As a smaller value of granularity parameter is assigned, the number of clusters has changed to Clusters $= 6$ and it is observed that members are more inclined to form independent clusters compared to Fig. 10(b). Fig. 12(b) displays the clustering result obtained by assigning the smallest granularity parameter value $px = 0.1$ at $t = 35$ sec on the same data used in Fig. 9-11. Now, the number of clusters largely has increased to Clusters $= 13$ and it is easily observed that the members are more isolated to other clusters as the extreme $px$ value has been applied.

Fig. 13-15 display NMDP-APC clustering results at $t = 60$ sec, the time when the entire traffic data are used. Now the observation span has exceeded $2$ km and the number of vehicles has reached $N = 61$. In the same manner, NMDP-APC is applied for different values of clustering parameter $px = 1.4$, $0.5$ and $0.1$ respectively to observe the clustering status on the traffic data. Fig. 13(b) displays the clustering result obtained by applying $px = 1.4$ to form medium sized clusters. The number of clusters in this condition is shown as Clusters $= 6$. The same values of NMDP-APC weight parameter $\varphi = 0.8$ and $\psi = 0.2$ are applied respectively. Comparing this result with Fig. 10(b), although medium sized clusters are formed, Fig. 13(b) tends to consist of more members per cluster. We attribute this increase to the observation range that spans $2$ km wide and the number of target vehicles that becomes almost double. The number of target vehicles performing NMDP-APC directly relates to the wireless propagation range of C-V2X. As NMDP-APC process will be performed for only those vehicles able to reach the wireless propagation, we assume a $1$ km radius to be a critical span. Fig. 14(b) displays the clustering result obtained by applying $px = 0.5$, which aims to form small-to-medium sized clusters on the same data at $t = 60$ sec and the number of clusters identified as Clusters $= 9$. Finally, Fig. 15(b) displays the clustering result obtained by applying $px = 0.1$, which aims to form smallest sized clusters on the same data at $t = 60$ sec and the number of clusters made in this condition rises to Clusters $= 18$ as an extreme $px$ value applied, and the same tendency has been constantly observed through the simulations.

### C. Prediction Performance in Simulation

In this part, we present sets of simulation results on the prediction performance of the soft-margin-based SVM-ML with GRBF kernel operation. We conducted these simulations in a Python coding environment as it has a series of well-prepared functions and libraries specifically for ML. As shown in Table-1, SVM and NN are the major candidate algorithms for classification. Simulation results obtained in using NN however provided an inferior performance compared to those in SVM in this simulation. We recognized that this is because our task vector $X$ has only four dimensions and that algorithms using NN would be more applicable to a large dimensional problem case. Thus, we conducted further simulation focusing on the SVM by varying the formerly explained $C$ and $\gamma$ parameters.

Fig. 16 shows the prediction accuracy obtained by applying constant $C = 1.0$ and then varying two types of $\gamma$-parameters, which are SVM-default $\gamma$ and SVM-scaled $\gamma$. The SVM-default uses $\gamma = 1/\text{feature number}$, i.e., $0.25$ as our task vector is 4 dimensional features. The SVM-scaled $\gamma = 1/\text{feature number multiplying by the variance of the task vector}$ $X$. As the result indicates, SVM-default $\gamma$ provided better prediction performance compared to SVM-scaled $\gamma$ parameter. It is notable that reached 100% of production accuracy with the sampling data given around at $140$. It is remarkable that such high performance was obtained even with the small number of training data. This is important specifically in a dynamic vehicular application scenario. We recognize that this is attributable to the compact structure of proposed decision tree, which only consists of essential criteria with a four-dimensional task vector.

Now in Fig. 17, we demonstrated prediction performance evaluation comparing default-$\gamma$ with $C = 1.0$ where best performance obtained in above, versus scaled-$\gamma$ by changing parameter-$C$ to observe how the value influence the result. As the theoretical analysis reviewed, the cost-parameter $C$ controls the balance of margin allowance and miss-classifications. In this simulation, by increasing $C$ value, improvements in the prediction performance have been observed in SVM scaled-$\gamma$. However, 100 % accuracy has not been reached in SVM scaled-$\gamma$ although increasing the $C$ value and even large volume of training data applied. From these results, it can be noticed that applying combination of
SCM default-$\gamma$ with $C = 1.0$ potentially provides desirable prediction performance in the simulation condition.

Finally, we observed the statistical average access latency in 5G compared with LTE through a simulation, as in Fig. 18. The access latency was statistically measured from a CH to gNB and eNB, i.e., the Uu interface by assigning the target latency as the median value, i.e., $\mu_{5G} = 4$ ms and $\mu_{LTE} = 11.5$ ms [3], [50]. The results are obtained by applying a truncated normal distribution [51] to avoid extreme and unrealistic values being produced. The simulation applied 10,000 times of sampling by changing different values of latency distribution variance $\sigma^2 = 1.0, 5.0$ and 10.0 then, we observed how the system stability influenced the average access latency in this simulation. The outcomes showed that the average access latency observed in 5G was 3.9 ms at $\sigma^2 = 1.0, 5.8$ ms at $\sigma^2 = 5.0$ and 7.8 ms at $\sigma^2 = 10.0$ respectively. In LTE access, the average access latency observed was 11.0 ms at $\sigma^2 = 1.0, 12.0$ ms at $\sigma^2 = 5.0$ and 15.4 ms at $\sigma^2 = 10.0$ respectively. A clear tendency observed was that when a large variance is applied, i.e., the system is unstable, the average access latency has deteriorated. The result can provide an understanding that the more stable access is provided, the less access latency can be achieved. Therefore, our objective of minimizing the number of accessing objects can also contribute to provide the reduction of access latency, contributing to the system scalability for avoiding access storms from individual vehicles and mobile MTC terminals potentially to be deployed in a large number in the future.

VI. CONCLUSION

This paper has proposed an advanced VANET clustering scheme called NMDP-APC to form stable single-hop clusters in a distributed control manner. Specifically, in consideration of the real-time motion dynamics of vehicles, two parameters, i.e., inter-vehicular Euclidian distances and its velocities, are assigned as the metrics without associating them with any prediction values. In the NMDP-APC process, we controlled the clustering granularity level by adjusting the granularity parameter. Particularly, this study deduced the desirable clustering size, applying a ML scheme which employs soft-margin-based SVM-ML with Gaussian Radial Basis Kernel Function.

As ML has been known as a powerful tool in decision making, we applied it to find an ideal VANET granularity incorporating minimum sets of decision criteria. Relative to the limited and essential criteria, the ML prediction performance achieved satisfactory results with fewer training data. In addition, we designed a message sequence and procedure of NMDP-APC by associating them with the ML deduced clustering granularity. The procedure and sequence are enhanced on the existing 3GPP C-V2X specifications. With the PC5 interface, therefore, the proposed scheme can be easily implemented with emerging 5G cellular systems. Especially, the proposed scheme is designed through a distributed control approach to adapt nomadic vehicles’ mobility. Through the simulations, the cluster formation and granularity control has been clearly observed through different values of the $px$-parameter. For improving ML prediction performance, two key parameters $C$ and $\gamma$ were adjusted to tune the prediction performance. The simulation results indicated that a particular selection of parameters resulted in a better performance. We also observed the performance with lesser access latency in a stable PLMN system via the simulation, which explains that the clustering capability is contributing...
to providing aggregation effects and thus it provides a stable system even when a large number of objects are connecting to a PLMN.

Although this study was originally targeted to vehicular applications, it has potential applicability to any machine type moving objects expected to increase reflecting the growing demand for such applications in 5G and beyond. Future work should focus on enhancing the proposed scheme by applying more complicated scenarios and/or using different types of ML algorithms.

ACKNOWLEDGMENT

The authors would like to express sincere thanks to the anonymous referees for their invaluable comments and constructive suggestions for the improvement and refining of this work.

REFERENCES

[1] Qualcomm. (Sep. 2017). Accelerating C-V2X Commercialization. [Online]. Available: https://www.qualcomm.com/documents/path-5g-cellular-vehicle-everything-c-v2x
[2] About Us. Bridging the Automotive and ICT Industries. Accessed: Oct. 21, 2019. [Online]. Available: https://5gaa.org/about-5gaa/about-us
[3] G. Naik, B. Choudhury, and J.-M. Park, “IEEE 802.11bd & 5G NR V2X: Multi-Dimensional Affinity Propagation Clustering Applying a ML in 5G-Cellular V2X” in IEEE Access, vol. 8, pp. 559–570, Mar. 2021.
[4] About Us, Bridging the Automotive and ICT Industries
[5] T. Koshimizu, H. Wang, Z. Pan, J. Liu, and S. Shimamoto, “Normalizing multidi-#multi-choice
dimensional parameter based affinity propagation clustering for cellular V2X,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Apr. 2018, pp. 1–6.
[6] F. Aftab, A. Khan, and Z. Zhang, “Hybrid self-organized clustering scheme for drone based cognitive Internet of Things,” IEEE Access, vol. 7, pp. 56217–56227, 2019.
[7] A. Malathi and N. D. Sreenath, “A comparative study of clustering protocols in VANET,” Int. J. Emerg. Trends Technol. Comput. Sci., vol. 6, no. 1, pp. 225–229, Apr. 2013.
[8] C. Cooper, D. Franklin, M. Ros, F. Safaei, and M. Abolhasan, “A comparative survey of VANET clustering techniques,” IEEE Commun. Surveys Tuts., vol. 19, no. 1, pp. 657–681, 1st Quart., 2017.
[9] A. Benslimane, T. Taleb, and R. Sivaraj, “Dynamic clustering-based adaptive mobile gateway management in integrated VANET—3G heterogeneous wireless networks,” IEEE J. Sel. Areas Commun., vol. 29, no. 3, pp. 559–570, Mar. 2011.
[10] S. Ucar, S. C. Ergen, and O. Ozkasap, “Multi-hop-cluster-based IEEE 802.11p and LTE hybrid architecture for VANET safety message dissemination,” IEEE Trans. Veh. Technol., vol. 65, no. 4, pp. 2621–2636, Apr. 2016.
[11] S. Ucar, S. C. Ergen, and O. Ozkasap, “VMaSC: Vehicular multi-hop algorithm for stable clustering in vehicular ad hoc networks,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Apr. 2013, pp. 2381–2386.
[12] W. Qi, Q. Song, X. Wang, L. Guo, and Z. Ning, “SDN-enabled social-aware clustering in 5G-VANET systems,” IEEE Access, vol. 6, pp. 28213–28224, 2018.
[13] C. Shea, B. Hassanabadi, and S. Valaei, “Mobility-based clustering in VANETs using affinity propagation,” in Proc. GLOBECOM-IEEE Global Telecommun. Conf., Nov. 2009, pp. 1–6.
[14] B. Hassanabadi, C. Shea, L. Zhang, and S. Valaei, “Clustering in vehicular ad hoc networks using affinity propagation,” Ad Hoc Netw., vol. 13, pp. 535–548, Feb. 2014.
[15] H. Shahwani, T. D. Bui, J. P. Jeong, and J. Shin, “A stable clustering algorithm based on affinity propagation for VANETs,” in Proc. 19th Int. Conf. Adv. Commun. Technol. (ICACT), 2017, pp. 501–504.
[16] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” Science, vol. 315, pp. 972–976, Feb. 2007.
[17] Qualcomm. (2018). C-V2X Trial in Japan. [Online]. Available: https://www.qualcomm.com/media/documents/file/c-v2x-trial-in-japan.pdf
[18] K. Bevrani, E. Chung, and M. Miska, “Evaluation of the GHR car following model for traffic safety studies,” in Proc. 25th ARRB Conf., Perth, WA, Australia, 2012, pp. 1–11.
[19] T. V. Mathew, “Car following models,” in Transportation System Engineering, vol. 35. Mumbai, India: IIT Bombay, Mar. 2017, ch. 35, pp. 35.2–35.6.
[20] S. A. A. Shah, E. Ahmed, M. Imran, and S. Zeedally, “5G for vehicular communications,” IEEE Commun. Mag., vol. 56, no. 1, pp. 111–117, Jan. 2018.
[21] H. El-Sayed, S. Sankar, M. Prasad, D. Puthal, A. Gupta, M. Mohanty, and C.-T. Lin, “Edge of things: The big picture on the integration of edge, IoT and the cloud in a distributed computing environment,” IEEE Access, vol. 6, pp. 1706–1717, 2018.
[22] Z. Ning, X. Wang, and J. Huang, “Mobile edge computing-enabled 5G vehicular networks: Toward the integration of communication and computing,” IEEE Veh. Technol. Mag., vol. 14, no. 1, pp. 54–61, Mar. 2019.
[23] L. Li, Y. Li, and R. Hou, “A novel mobile edge computing-based architecture for future cellular vehicular networks,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Mar. 2017, pp. 1–6.
[24] Z. Ma, J. Zhao, Y. Gong, and Y. Wang, “Key technologies of MEC towards 5G-enabled vehicular networks,” in Quality, Reliability, Security and Robustness in Heterogeneous Systems. QShine (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), vol. 234, L. Wang, T. Qu, and W. Zhao, Eds. Cham, Switzerland: Springer, 2017.
[25] A. A. Khan, M. Abolhasan, and W. Ni, “An evolutionary game theoretic approach for stable and optimized clustering in VANETs,” IEEE Trans. Veh. Technol., vol. 67, no. 5, pp. 4501–4513, May 2018.
[26] H. Ye, L. Liang, G. Ye Li, J. Kim, L. Lu, and M. Wu, “Machine learning for vehicular networks: Recent advances and application examples,” IEEE Veh. Technol. Mag., vol. 13, no. 2, pp. 94–101, Jun. 2018.
[27] L. Liang, H. Ye, and G. Y. Li, “Toward intelligent vehicular networks: A machine learning framework,” IEEE Internet Things J., vol. 6, no. 1, pp. 124–135, Feb. 2019.
[28] D. Tian, Y. Wang, H. Xia, and F. Cai, “Clustering multi-hop information dissemination method in vehicular ad hoc networks,” IET Intell. Transp. Syst., vol. 7, no. 4, pp. 464–472, Jan. 2013.
[29] Z. Khan, P. Fan, F. Abbas, H. Chen, and S. Fang, “Two-level cluster based routing scheme for 5G V2X communication,” IEEE Access, vol. 7, pp. 16194–16205, 2019.
[30] C. Junli and J. Licheng, “Classification mechanism of support vector machines,” in Proc. 5th Int. Conf. Signal Process. Proc. 16th World Comput. Congr, Aug. 2000, pp. 1556–1559.
[31] A. Ben-Hur and J. Weston, “A user’s guide to support vector machines,” in Methods in Molecular Biology. Bethesda, MD, USA: National Center for Biotechnology Information, Jan. 2010. [Online]. Available: https://www.researchgate.net/publication/41896604_A_User’s_Guide_to_Support_Vector_Machines
[42] T. Kurita, “An introduction to support vector machine,” Neuroscience Res. Inst., Nat. Inst. Adv. Industral Sci. Technol., Tokyo, Japan, Tech. Rep., Jul. 2002. [Online]. Available: https://home.hiroshima-u.ac.jp/kturita/lecture/svm.pdf

[43] S. Akaho, “Support vector machine,” Adv. Ind. Science Technol. (AIST), Tokyo, Japan, Tech. Rep. AIST Open Lecture Jul 6-7, 2007. [Online]. Available: https://www.ism.ac.jp/~fukumizu/ISM_lecture_2006/svm-ism.pdf

[44] S. Wu, W. Xia, W. Cui, Q. Chao, Z. Lan, F. Yan, and L. Shen, “An efficient offloading algorithm based on support vector machine for mobile edge computing in vehicular networks,” in Proc. 10th Int. Conf. Wireless Commum. Signal Process. (WCSP), Oct. 2018, pp. 1–6.

[45] Affinity Propagation Frequently Asked Questions. Accessed: Jan. 15, 2018. [Online]. Available: http://www.psi.toronto.edu/affinitypropagation/faq.html

[46] Proximity-Based Services (ProSe), Stage 2, Release 15, document TS 23.303, 3GPP, Jun. 2018.

[47] sklearn.svm.SVC. Support Vector Classification. Accessed: May 11, 2019. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[48] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, pp. 273–297, Sep. 1995.

[49] S. Bhattacharyya. (Dec. 2018). Support Vector Machine: Kernel Trick; Mercer’s Theorem Understanding SVM Series: Part 2. [Online]. Available: https://towardsdatascience.com/understanding-support-vector-machine-part-2-kernel-trick-mercers-theorem-e1e6848e6c4d

[50] A. Khan, A. Minokuchi, K. Tsubouchi, G. Kunito, S. Iwashina, “Technology and standards accelerating 5G commercialization,” IEICE Trans. Commun., vol. 102, no. 3, pp. 1–10, 2018.

[51] Wikipedia. Truncated Normal Distribution. Accessed: Jul. 10, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Truncated_normal_distribution

TAKASHI KOSHI Mizu (Member, IEEE) received the M.S. degree from the Graduate School of Electrical and Computer Science, Oklahoma State University, OK, USA, in 1998. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Communications Engineering, School of Fundamental Science and Engineering, Waseda University. He has worked for NTT DOCOMO R&D Center, as a Core Network Standardization Expert. He dedicated the work specifically in signaling and system architecture in SA2 WG from the 3rd Generation to 5th Generation cellular network. His research interests include vehicular communication, machine learning, and beyond 5G cellular networks.

SHI GENG TIAN received the B.S. degree in software engineering from the Harbin Far East Institute of Technology, China, in 2013. He is currently pursuing the M.S. degree in wireless communications engineering with Waseda University, Tokyo, Japan. His research interests include intelligent wireless communication, D2D, and dynamic power control systems.

HUAN WANG received the B.S. degree from the Nanjing University of Information Science and Technology, China, in 2013, and the M.S. degree in wireless communication engineering from Waseda University, Tokyo, Japan, in 2017, where he is currently pursuing the Ph.D. degree in wireless communication engineering. His research interests include intelligent transportation systems, V2X, and dynamic traffic signal control.

ZHENNI PAN (Member, IEEE) received the B.S. degree in computer science engineering from China Agricultural University, Beijing, China, in 2007, and the M.S. and Ph.D. degrees in information and telecommunications from the Graduate School of Global Information and Telecommunication Studies (GITS), Waseda University, Japan, in 2011 and 2018, respectively. She was a Research Associate at Waseda University, from 2013 to 2018, where she is currently an Assistant Professor at the Global Center for Science and Engineering, Faculty of Science and Engineering. Her research interests include green wireless communications, mobile communications, wireless sensor networks, and optical communication.

JIAN CHENG received the B.S. degree in electronic engineering from the Chongqing University of Technology, China, in 2001, and the M.S. and Ph.D. degrees in information and telecommunications from Waseda University, in 2006 and 2012, respectively. From 2009 to 2012, she was a Research Associate at Waseda University, where she joined the Faculty of Science and Engineering, in 2012, as an Assistant Professor, and since 2017, she has been an Associate Professor affiliated with the International Center for Science and Engineering Programs. Her researches focus on the optical wireless communications (OWC), wireless network systems, near field communication, and their applications on 5G network design and healthcare industry. She is a Committee Member of the Institute of Image Electronics Engineers of Japan (IEEE) and the Institute of Electronics, Information and Communication Engineers (IEICE). She also serves as the Secretary of the Japan Division for the Institution of Engineering and Technology (IET).

SHIGERU SHIMAMOTO (Member, IEEE) received the B.E. and M.E. degrees in communication engineering from The University of Electro-Communications, Tokyo, Japan, in 1985 and 1987, respectively, and the Ph.D. degree from Tohoku University, Japan, in 1992. From 1987 to 1991, he joined NEC Corporation. From 1991 to 1992, he was an Assistant Professor with The University of Electro-Communications. From 1992 to 1993, he was an Assistant Professor with the Department of Computer Science, Gunma University, where he was also an Associate Professor, from 1994 to 2000. Since 2002, he has been a Professor with the Graduate School of Global Information and Telecommunication Studies, Waseda University. In 2008, he served as a Visiting Professor with Stanford University, USA. In 2014, he became a Professor with the Department of Communications and Computer Engineering, Waseda University. His research interests include satellite and mobile communications, access scheme, and unmanned aerial vehicle for disaster networks.