Vocation Identification for Heavy-duty Vehicles: A Tournament Bracket Approach

Daniel Kobold Jr.[1], Andy Byerly[2], Rishikesh Mahesh Bagwe[1], Euzeli Cipriano dos Santos Jr.[1][a] and Zina Ben Miled[1][b]

[1]*Department of Electrical and Computer Engineering (IUPUI), Indianapolis, IN 46202, U.S.A.*

[2]*Allison Transmission, Inc., One Allison Way, Indianapolis, IN 46222, U.S.A.*

Keywords: Heavy-duty Vehicles, Vocation, Classification.

Abstract: The identification of the vocation of an unknown heavy-duty vehicle is valuable to parts' manufacturers. This study proposes a methodology for vocation identification that is based on clustering techniques. Two clustering algorithms are considered: K-Means and Expectation Maximization. These algorithms are used to first construct the operating profile of each vocation from a set of vehicles with known vocations. The vocation of an unknown vehicle is then determined by using one-versus-all or one-versus-one assignment. The one-versus-one assignment is more desirable because it scales with an increasing number of vocations and requires less data to be collected from the unknown vehicles. These characteristics are important to parts' manufacturers since their parts may be installed in different vocations. Specifically, this paper compares the one-versus-one bracket and the one-versus-one round-robin tournament assignments to the one-versus-all assignment. The tournament assignments are able to scale with an increasing number of vocations. However, the bracket assignment also benefits from a linear time complexity. The results show that despite its scalability and computational efficiency, the bracket vocation identification model has a high accuracy and a comparable precision and recall. The NREL Fleet DNA drive cycle dataset is used to demonstrate these findings.

1 INTRODUCTION

The ability to identify the vocation of a heavy-duty vehicle from drive cycle data collected during the vehicle's daily operation is valuable to many parts' manufacturers in order to track the end-use of the vehicle. Electronic components and sensors are becoming increasingly pervasive in vehicles. This led to new sources of data. In fact, both OEMs and parts' manufacturers now have access to a large stream of operational data that can be acquired during maintenance or configuration updates, with the vehicle owner's consent. However, as opposed to OEMs, parts' manufacturers do not typically have knowledge of the actual use, application or vocation of the vehicle. Moreover, the same part can be deployed in a large number of varying vocations. The classification of vehicle usage and underlying parts into vocations benefits the component designers who likely do not have complete awareness of how the vehicle will be used. This classification may be obvious in the OEM service channel where direct physical interaction with the vehicle allows the identification of the vocation of the vehicle and consequently of the components and chassis (e.g., refuse truck, cement mixer, dump truck, coach bus, transit bus, etc.) However, this information is not directly accessible to the parts' manufacturer. Vocation classification can be used to: 1) detect when the component is not being used in a manner consistent with the intended vocation (e.g., a coach bus operating as a transit bus) or 2) identify field issues that are specific to the operational use of a given component in the vehicle. These issues can then be addressed via field action (e.g., part configuration updates) or future design improvements to the component.

This paper presents a methodology for identifying the vocation of an unknown heavy-duty vehicle using features collected from the vehicle's daily drive cycles. The methodology follows a two-step approach. First, the profile of each vocation is established using a set of vehicles with known vocations. Second, the daily drive cycles collected from the unknown vehicle are compared to all vocation profiles and the

[a] https://orcid.org/0000-0003-1584-8082

[b] https://orcid.org/0000-0001-5401-0368

Kobold Jr., D., Byerly, A., Bagwe, R., Santos Jr., E. and Ben Miled, Z.
Vocation Identification for Heavy-duty Vehicles: A Tournament Bracket Approach.
DOI: 10.5220/0010298702590266
In *Proceedings of the 7th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2021)*, pages 259-266
ISBN: 978-989-758-513-5

most likely vocation is selected. The methodology is demonstrated using two widely used clustering algorithms K-means (KM) (Chakraborty et al., 2020) and expectation maximization (EM) (Shin et al., 2019). However, it can be extended to other clustering algorithms such as particle swarm optimization (PSO) (Kennedy and Eberhart, 1995).

Most classification algorithms are best at handling two classes (Athimethphat and Lerteerawong, 2012): a positive and a negative class. These binary classifiers have been extended to multiclass models using the one-versus-all (Scholkopf and Smola, 2001) and the one-versus-one methodology (Daengduang and Vateekul, 2017). The one-versus-all consists of an ensemble of classifiers where each classifier is trained to correctly predict one positive class while considering all the remaining classes as negative. This method has a linear complexity with respect to the number of vocations. The one-versus-one is also an ensemble of classifiers. However, a classifier is developed for each pair of classes leading to a quadratic complexity.

Vocation identification for heavy duty-vehicles is a multiclass application and the proposed methodology was inspired by the one-versus-one classification approach which can accommodate a large number of vocations. The daily measurements of the unknown vehicle are compared to two vocations at a time in a tournament bracket. However, as opposed to the traditional one-versus-one classification, a vocation is eliminated in each round making the approach linearly scalable with respect to the number of vocations. The proposed methodology is applied to 5 vocations from the NREL fleet DNA (NREL, 2019) dataset.

2 RELATED WORK

The purpose of a classifier is to assign a given data record to one of the pre-defined classes. Since vocations are known a-priori in our study, using a classifier with supervised learning would be expected. Some of the widely recognized classification algorithms include support vector machine (SVM) (Scholkopf and Smola, 2001), random forest (RF) (Breiman, 2001), and neural networks. Most of these algorithms are inherently two-class (binary) classifiers. However, they have been extended to accommodate multiclass applications. For instance, SVM was applied to multiclass classifiers using one-versus-one and one-versus-all ensemble learners (Scholkopf and Smola, 2001). Similarly, neural networks can use multiple nodes in the output layer where each node corresponds to a class (Sagi and Rokach, 2018). RF can also support multiple classes if multiway trees are used instead of binary decision trees (Murphy and Pazzani, 1991).

The purpose of a clustering algorithm is to: a) identify clusters with similar records, b) select a representative member for each cluster and c) adequately assign a record to a cluster. These three aspects vary from one clustering algorithm to the next. As opposed to a classifier, the first step is performed using unsupervised learning. For example, KM defines the similarity between two records according to a distance measure. The smaller the distances the more similar are the records. Other similarity criteria that are optimized to specific applications are proposed in (Kanemaru et al., 2013), and (Wang et al., 2020).

Once a cluster is identified, a representative member, called the centroid is selected and refined iteratively as members are added to or removed from the cluster. The centroid is typically calculated by averaging across all the members of the cluster. Other clustering algorithms, such as PSO, derive their efficiency from the selection of appropriate centroids. Centroids are mapped to particles in PSO. Each particle moves in the feature space and its velocity is updated based on the best position that the particle has found so far and the current global best position across all particles.

The assignment of a record to a cluster also varies from one clustering algorithm to the next. For KM, each record is assigned to exactly one cluster based on the distance between the record and the centroid of the cluster. This assignment is referred to as a "hard" assignment. EM uses a "soft" assignment (Wahba, 2002). That is, each record has a probability of belonging to each cluster.

Other important aspects of clustering algorithms include the relationship among the clusters and the appropriate number of clusters. Most clustering algorithms assume that all clusters are at the same level. This type of clustering is referred to as partitioning (Ester et al., 1996). This is also the type of clustering being used in this paper. In contrast, hierarchical clustering (McInnes and Healy, 2017) allows some clusters to be a subset of others.

Clustering has been used in several vehicular applications. For example, it was used in (Kanemaru et al., 2013) for sharing of traffic congestion information. Each cluster of vehicles was used to represent a given traffic flow thereby allowing the vehicle at the head of the flow to inform the vehicle at the tail of the flow of any traffic congestion. In (Wang et al., 2020), clustering was used to detect anomalous cab trajectories. Each of the above applications innovate by proposing a customized similarity measure for the target application.

The fleet DNA dataset used in this study was in-

troduced and extensively analyzed in (Duran et al., 2018). Indeed, dimension reduction was performed on the dataset using principal component analysis (PCA) and cross-correlation to identify the eight most expressive features in the dataset. These were found to be aerodynamic speed, characteristic acceleration, percent of total cycle distance accumulated at speeds below 55 mph, percent of total cycle time duration accumulated at vehicle speeds of 0 mph, number of vehicle stops per mile, mean (nonzero) driving speed, maximum driving speed and standard deviation of (nonzero) driving speed. Using these eight features, the study found that the first 6 components of PCA were able to describe 99% of the variance in the data. KM was also used to cluster all the drive cycles in the fleet DNA dataset into three clusters.

The above study by NREL helped guide the methodology proposed in this paper. That said, the present paper addresses a different problem. The NREL study (Duran et al., 2018) aims at identifying a limited number of representative drive cycles across all US commercial fleets. The aim of the present paper is to identify the specific vocation of an unknown vehicle. The methodology is also different since it demonstrates the use of a clustering algorithm for vocation identification. In fact, while targeting a different application, the methodology proposed in this paper shares this aspect with the approach for the detection of anomalous cab trajectories proposed in (Wang et al., 2020). The algorithm proposed in this paper enhances this methodology by showing that a one-versus-one bracket assignment can be efficiently applied to a large number of classes.

3 METHODOLOGY

The proposed methodology creates a model that identifies the vocation of an unknown vehicle. In the next subsections, we describe the dataset, the training phase of the model which establishes the operating profile of each vocation, and the three vocation assignment algorithms.

3.1 Dataset

Each vehicle in the Fleet DNA dataset is represented by a set of records where every record is an aggregation of the drive cycle measurements over a single day. The features of the records used in this study are shown in Table 1. Their definitions are available in (Duran et al., 2018) and references therein. For convenience, some of these definitions are reproduced below:

- Total Average Speed: Average speed over the trip (including zero speed points).
- Driving Average Speed: Average speed over the trip not including the zero speeds.
- Zero Seconds: Number of seconds at zero speed.
- Average Kinetic Power Density Demand: Mean of the kinetic power density demand (with respect to mass).

Table 1: Feature list.

	Feature
1	Max Speed (mph)
2	Total Average Speed (mph)
3	*Total Speed Standard Deviation (mph)
4	Driving Average Speed (mph)
5	Driving Speed Standard Deviation (mph)
6	Zero Seconds (s)
7	Distance Total ($miles$)
8	Total Stops ($count$)
9	*Average Kinetic Power Density Demand (W/kg)
10	*Cumulative Instantaneous Kinetic Energy Density (J/kg)
11	*Characteristic Acceleration (m/s^2)
12	*Aerodynamic Speed (m/s)
13	Max Acceleration (ft/s^2)
14	Average Acceleration (ft/s^2)
15	*Max Deceleration (ft/s^2)

The list of the 15 features shown in Table 1 was selected among the 350 available variables in the original dataset using dimension reduction. Some of the variables in the original data identify the vehicle, the deployment or the vocation. These were used to label the data. A large number of variables were removed because they had a linear or an inverse relationship with another variable (e.g., Characteristic Acceleration and Characteristic Deceleration, Average Acceleration and Average Deceleration). Variables related to potential energy (e.g., Cumulative Instantaneous Potential Energy Density and Average Potential Power Density Demand) were also removed because they are more dependent on the road elevation than on the vocation of the vehicle. Moreover, daily records with Zero Seconds $>$ 18,000s were also removed from all the vehicles because this is an indication that the vehicle was not in operation for more than 5 hours in the given day.

The Fleet DNA dataset includes eight vocations: Bucket Trucks, Class 8 Tractors, Delivery Vans, Delivery Trucks, Transit Buses, Refuse Trucks, School Buses, and Service Vans. The latter three vocations were eliminated because they did not include suffi-

cient data. For the remaining vocations, the vocation identification model followed a training/testing split at the vehicle level. This prevents information leakage that may result from allowing records from the same vehicle to participate in both the training and the testing of the model. After assigning a vehicle to either training or testing, 13 records were randomly sampled without replacement from each vehicle. Each random selection was considered as a separate vehicle. This effectively allows a given vehicle to appear multiple times in either the training or testing vehicle pools. However, the underlying drive cycle will always be unique as per the sampling policy. Moreover, to keep the training records balanced across vocations, 10 vehicles were selected per vocation for training. The remaining vehicles were used for testing. This split approach led to variations in the number of vehicles available for testing across the vocations (Table 2). In total, 50 vehicles are used for training and 81 are used for testing across the 5 vocations.

Table 2: Number of test vehicles in each vocation.

Vocation	Total num. of vehicles	Num. of test vehicles
Bucket Truck (BT)	12	2
Class 8 Tractor (CT)	43	33
Delivery Truck (DT)	29	19
Delivery Van (DV)	26	16
Transit Bus (TB)	21	11

Each vocation represents a group of vehicles that perform similar tasks. A detailed description of each vocation in the fleet DNA is provided in (NREL, 2019). Some of the vocations (e.g., Transit Bus) have a distinct operational profile while others have an operational profile that can be confounded with the remaining vocations. Delivery Vans (DV) and Delivery Trucks (DT) are expected to have similar operating profiles since the main difference between these two vocations is the vehicle weight, with DT vehicles being typically heavier than DV vehicles. Bucket Trucks (BT) perform tasks at the job site and will possibly spend less time driving from one point to another. Thus, compared to DT, DV and TB vehicles, the operational profile of BT vehicles should show lower distances traveled and lower average speeds. Class 8 Tractors (CT) are typically used to haul a trailer from a source (e.g., distribution center) to a destination (e.g., customer site). Therefore, CT vehicles are expected to travel long distances over highways compared to DT or DV vehicles. However, according to the vocation characteristics in (NREL, 2019), the CT vocation consists of various types of class 7 and 8 vehicles that can be used for different tasks ranging from food delivery to long-hauling tasks. This variation explains some of the results discussed in Section 4.

3.2 Model Development

The training is executed for each vocation independently. It starts by randomly selecting a set of initial centroids for the target vocation from the available training data. During each iteration, records from the training data are compared to each centroid of the vocation. After processing all records, the centroids are updated and a new training iteration begins. The records and centroids are denoted as follows:

- $\mathbf{r}_i = (r_i[1], r_i[2], ..., r_i[n])$ represents a record where each element $r_i[.]$ of the input vector $\mathbf{r}_i$ is the value of one of the input features and n is the total number of features.
- $\mathbf{C}_v = \{\mathbf{cv}_1, \mathbf{cv}_2, ..., \mathbf{cv}_m\}$ is the set of centroids of vocation $\mathbf{v}$ where each centroid represents a cluster of the vocation $\mathbf{v} \in \mathbf{V} = \{BT, CT, DV, DT, TB\}$. The total number of centroids, m, for each vocation in this study is fixed.

Under the KM algorithm, each record is assigned to exactly one cluster which is selected according to the minimum distance between the record and the centroids of all clusters. In the case of EM, the assignment of a record to a cluster follows a probabilistic measure. This measure is derived using Bayes' rule, with the assumption that each feature has a normal distribution and that all the features are independent. At the end of each training iteration of either the KM or EM algorithms, the centroids of the clusters are updated according to the record assignment derived during the iteration.

3.3 Feature Reduction

Even though the starting dataset was manually reduced from 350 parameters to 15 features as described in Section 3.1, a minimalist model is desirable in order to limit the deployment cost of the vocation identifier and promote its applicability in production. This minimalist model should only include the features that are necessary and practical for vocation identification. Feature reduction was performed using the wrapper induction method (Khalid et al., 2014). During each iteration of the feature reduction process, the standard deviation of each target feature is evaluated for each cluster and the feature is removed if the resulting value is below a certain pre-set threshold across all the clusters. One feature was considered per iteration until none of the features had a standard deviation below this threshold. In addition, features

that are easier to collect (e.g., vehicle speed) were favored over features that may not be readily available (e.g., characteristic acceleration and kinetic energy density). In the remainder of the paper, the model with the full feature set is labeled FFmodel and the reduced feature model is labeled RFmodel.

3.4 Vocation Assignment

Once the model is trained, it is exposed to a record $\mathbf{r}_i$ from an unknown vehicle. That is, for each vocation $\mathbf{v}$ and centroid $\mathbf{cv}_j$ of $\mathbf{v}$, the conditional probability $P(\mathbf{cv}_j|\mathbf{r}_i)$ is calculated. In the case of KM, this probability measure is binary. The record is then assigned to the vocation - $v^T(\mathbf{r}_i)$ - with the largest probability according to the following equation:

$$v^T(\mathbf{r}_i) = \underset{\mathbf{v}\in\mathbf{V}}{\operatorname{argmax}} \left\{ \underset{1\le j\le m}{\operatorname{argmax}} \{P(\mathbf{cv}_j|\mathbf{r}_i)\} \right\}. \quad (1)$$

Equation 1 is used for a single daily record from an unknown vehicle. When the unknown vehicle has multiple records, each record can be assigned to a different vocation and a consensus is needed to select the winning vocation. Let $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_p\}$ represent the set of records of the unknown vehicle. The winning vocation of the unknown vehicle is the vocation that is assigned the highest number of records. This process is defined by the following equation:

$$voc^T(\mathbf{R}) = \underset{\mathbf{v}\in\mathbf{V}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{p} v^T(\mathbf{r}_i) = \mathbf{v} \right\}. \quad (2)$$

Equations 1 and 2 show the *traditional one-versus-all* (T) assignment where all the vocations compete for the same vehicle at once. This assignment has an important limitation since the wrong vocations may weaken the chances of the correct vocation by acquiring several of the records of the unknown vehicle. This aspect is particularly important for the current application because the number of vocations can be large and the number of daily records available for each unknown vehicle is small.

$$v^R(\mathbf{r}_i, \mathbf{a}, \mathbf{b}) = \underset{\mathbf{v}\in\{\mathbf{a},\mathbf{b}\}}{\operatorname{argmax}} \left\{ \underset{1\le k\le m}{\operatorname{argmax}} \{P(\mathbf{cv}_k|\mathbf{r}_i)\} \right\} \quad (3)$$

$$voc^R(\mathbf{R}) = \underset{\mathbf{v}\in\mathbf{V}}{\operatorname{argmax}} \left\{ \sum_{\mathbf{a}\neq\mathbf{v}} \sum_{i=1}^{p} v^R(\mathbf{r}_i, \mathbf{a}, \mathbf{v}) = \mathbf{v} \right\} \quad (4)$$

In order to mitigate this potential limitation, the one-versus-one *round-robin tournament* (R) assignment was investigated. This assignment consists of multiple rounds where each vehicle is exposed to every combination of two vocations. The vocation of choice is the one that is assigned the most records across all of the rounds for a given vehicle as defined in (3) and (4).

Unfortunately, the round-robin assignment has a quadratic time complexity with respect to the number of vocations. The *tournament bracket* (B) also follows the one-versus-one assignment and consists of multiple rounds where the unknown vehicle is only exposed to two vocations in each round. However, in the bracket assignment, a vocation is eliminated in each round. The vocation that is retained is the one that collects the highest number of records from the unknown vehicle in the round and this vocation proceeds to the next round. The assignment concludes when only one vocation remains.

Equation (5) shows the selection between two vocations $\mathbf{a}$ and $\mathbf{b}$ for one round of the bracket assignment. This equation is applied recursively in order to determine the winning vocation as shown in (6).

$$w^B(\mathbf{R}, \mathbf{a}, \mathbf{b}) = \underset{\mathbf{v}\in\{\mathbf{a},\mathbf{b}\}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{p} v^R(\mathbf{r}_i, \mathbf{a}, \mathbf{b}) = \mathbf{v} \right\} \quad (5)$$

$$voc^B(\mathbf{R}) = w^B\left(\mathbf{R}, \mathbf{v}_c, w^B(\mathbf{R}, \mathbf{v}_{c-1}, \mathbf{v}_{c-2})\right) \quad (6)$$

where c is the number of vocations in $\mathbf{V}$. As opposed to the round-robin assignment, Equation (6) is only executed $c-1$ times allowing the bracket assignment to have a linear time complexity with respect to the number of vocations.

4 RESULTS AND DISCUSSION

The one-versus-all, round-robin and bracket assignments are applied to the dataset described in Table 2. During training, the centroids of each vocation are determined using 130 daily records from the vocation. The model is then exposed to the test vehicles.

4.1 One-versus-All Assignment

Table 3 shows the confusion matrix of the one-versus-all FFmodel with KM and EM clustering. The results are presented in this manner in order to facilitate the analysis of confounding vocations and the identification of vocations with unique profiles. At the end of the section, the aggregated accuracy, precision and recall of the models are discussed.

The assignment of a vehicle to a vocation follows (2). Each row in Table 3 represents a vocation. The entries are the number of vehicles of the target vocation (row) that are assigned to a given vocation (column). The numbers in between parenthesis represent the number of ties for each vocation. For example, the

CT vocation has a total of 33 test vehicles (Table 2). Using the KM algorithm, 20 out these vehicles were correctly assigned to the CT vocation. The remaining 13 vehicles were incorrectly assigned as follows: 6 to BT, 3 to DT, 2 to DV, 1 to TB and one vehicle was a tie between DT and TB. The KM FFmodel was able to correctly classify 51 out of the 81 test vehicles whereas the EM FF model shows 59 true positives.

Table 3: Vocation assignment of the test vehicles using the traditional one-versus-all KM and EM FFmodels.

		BT	CT	DT	DV	TB
KM	BT	2	0	0	0	0
	CT	6	20	3(1)	2	1(1)
	DT	2	0	10	4(1)	2(1)
	DV	4	1	1	8	2
	TB	0	0	0	0	11
EM	BT	2	0	0	0	0
	CT	2	21(1)	4	1	4(1)
	DT	0	0	15	4	0
	DV	4	1	1	10	0
	TB	0	0	0	0	11

None of the BT vehicles were assigned to a different vocation under the two FFmodels. Despite the low number of test vehicles in this vocation (Table 3), this is still an indication of the unique BT profile. TB is another vocation with a distinct operational profile with no vehicles incorrectly classified under both FFmodels. The large number of DT vehicles that are assigned to the DV vocation indicates that the two vocations may be similar as discussed in Section 3.1.

Table 4: Vocation assignment of the test vehicles using the one-versus-all KM and EM RFmodels.

		BT	CT	DT	DV	TB
KM	BT	2	0	0	0	0
	CT	4	22	2(1)	3(1)	1
	DT	2(1)	1	11(1)	1(1)	2(1)
	DV	4(2)	1	1(1)	7(1)	1
	TB	0	0	0	0	11
EM	BT	2	0	0	0	0
	CT	3	21(1)	4(1)	0(1)	3(1)
	DT	1	1	14	3	0
	DV	4(1)	0(1)	0	10(2)	0
	TB	0	0	0	0	11

Feature reduction as described in Section 3.3 was performed on the models. The features that were eliminated include Total Speed Standard Deviation and Average Kinetic Power Density Demand. The eliminated features are indicated by a '*' in Table 1. The reduced feature model (RFmodel) includes only 9 features which can all be derived from two readily available parameters: speed and distance traveled.

Table 4 shows the confusion matrix of the RFmodel under KM and EM. The model generated 53 and 58 true positives with KM and EM, respectively. The number of true positives for the reduced and full feature models are similar. However, the number of ties is higher for the reduced feature model. This is expected as fewer parameters are available to distinguish among all the vocations at the same time. The one-versus-one assignment was introduced to help address this limitation.

4.2 Round-robin Assignment

The KM and EM round-robin FFmodel models correctly classified 50 and 57 test vehicles, respectively (Table 5). The number of true positives is comparable to that of the corresponding traditional one-versus-all model. However, the round-robin assignment does not suffer from ties. The numbers of true positives for the KM and EM RFmodels with round-robin assignment are 55 and 58, respectively (Table 6).

Table 5: Vocation assignment of the test vehicles using the round-robin KM and EM FFmodels.

		BT	CT	DT	DV	TB
KM	BT	2	0	0	0	0
	CT	5	17	7	2	2
	DT	1	0	12	5	1
	DV	5	1	1	8	1
	TB	0	0	0	0	11
EM	BT	2	0	0	0	0
	CT	2	19	4	0	8
	DT	0	0	15	4	0
	DV	4	0	2	10	0
	TB	0	0	0	0	11

As in the case of the one-versus-all assignment, EM performs better than KM for the round-robin models. Moreover, compared to the one-versus-all assignment, the round-robin assignment has higher number of true positives for all vocations except for the CT vocation. As discussed in Section 3.1, this exception may be due to the fact that the CT vocation is actually a combination of two or more vocations.

4.3 Bracket Assignment

Tables 7 and 8 show the bracket assignment for the FFmodel and RFmodel, respectively. Similar to the round-robin assignment, the bracket assignment does not suffer from ties and the number of true positives generated by the respective models is nearly the same. In fact, the model with the highest number of true pos-

itives is the bracket RFmodel. While the difference in performance may be marginal, the bracket RF model offers several advantages: It scales linearly with respect to the number of vocations; it is less susceptible to an increasing number of vocations since only two vocations are compared at a time; and it uses a reduced feature set that is readily available.

Table 6: Vocation assignment of the test vehicles using the round-robin KM and EM RFmodels.

		BT	CT	DT	DV	TB
KM	BT	2	0	0	0	0
	CT	4	19	6	3	1
	DT	0	1	13	2	3
	DV	2	1	2	10	1
	TB	0	0	0	0	11
EM	BT	2	0	0	0	0
	CT	4	20	6	1	2
	DT	0	1	15	3	0
	DV	5	0	0	11	0
	TB	0	0	1	0	10

Table 7: Vocation assignment of the test vehicles using the bracket KM and EM FFmodels.

		BT	CT	DT	DV	TB
KM	BT	2	0	0	0	0
	CT	5	19	6	1	2
	DT	1	0	12	5	1
	DV	5	1	1	8	1
	TB	0	0	0	0	11
EM	BT	2	0	0	0	0
	CT	2	19	4	1	7
	DT	0	0	15	4	0
	DV	4	1	1	10	0
	TB	0	0	0	0	11

Table 8: Vocation assignment of the test vehicles using the bracket KM and EM RFmodels.

		BT	CT	DT	DV	TB
KM	BT	2	0	0	0	0
	CT	4	20	5	3	1
	DT	1	1	13	1	3
	DV	2	2	1	10	1
	TB	0	0	0	0	11
EM	BT	2	0	0	0	0
	CT	4	21	4	1	3
	DT	2	1	13	3	0
	DV	5	0	0	11	0
	TB	0	0	0	0	11

The above results focus on the true positive assignments generated by each model. They show that the bracket model delivers the same or higher number of correct assignments compared to the other models while being computationally more efficient than the round-robin model and more scalable than the one-versus-all model. In the remainder of this section, we also show that these benefits do not come at the expense of a significantly lower precision or recall.

The average accuracy of the models across all vocations is 85% or higher. The precision and recall of these models are included in Table 9. This table shows that for each of the three assignments, EM has higher precision and recall than KM. The results also show that the reduced feature models have higher precision and recall compared to the full feature models. Finally, the model with the highest precision and recall (i.e., 75.3%) is the one-versus-all RF model. The bracket RFmodel has a higher precision and recall (71.6%) than all round-robin models.

5 CONCLUSIONS

This paper introduced a methodology for vocation identification of heavy duty vehicles when the number of vocations is expected to be large and the number of records available for each unknown vehicle is small. The profile of the vocation is first developed using a set of training vehicles. This profile consists of a set of centroids that represent the operating modes of the vocation. The unknown vehicle is then assigned to a vocation using a tournament bracket. In each round, two vocations are compared to the unknown vehicle and the unlikely vocation is eliminated. This assignment was compared to the one-versus-all and round-robin assignments. Two models were considered. The first was based on 15 features. Some of these features included complex variables which may not be accessible to the parts' manufacturer. The second model is more practical and was limited to 9 features that can be derived solely from speed and distance traveled. Compared to the full feature model, the reduced feature model had higher precision and recall.

Table 9: Precision (P) and recall (R) of the models with the three different assignments.

	One-versus-all		Round Robin		Bracket	
	KM	EM	KM	EM	KM	EM
	FFmodel					
P	62.2	73.8	61.7	70.4	64.2	70.4
R	63.0	72.8	61.7	70.4	64.2	70.4
	RFmodel					
P	66.3	75.3	67.9	67.9	69.1	71.6
R	65.4	75.3	71.6	71.6	69.1	71.6

With the exception of the CT vocation, the number of true positives for each vocation using the bracket

assignment is also either the same or higher than the true positives obtained using the one-versus-all and the round-robin assignments. The bracket assignment was introduced to avoid some of the drawbacks of the one-versus-all assignment for this application. The latter assignment inherently implies the availability of a large number of records for the unknown vehicles as these records are exposed to all the clusters of all the vocations at once. The bracket assignment overcomes this limitation by comparing two vocations at a time and was shown in this study to have a comparable performance to that of the one-versus-all assignment. The bracket assignment was also compared to a round-robin assignment which scales with an increasing number of vocations. The results show that the bracket assignment has higher precision and recall but most importantly has lower time complexity.

There are several directions that are being considered for future work including exploring the possibility of reducing vocation confounding by applying weights to specific features. In addition, the proposed vocation identification algorithm relies on features aggregated daily from the duty cycle of the vehicle over a period of 13 days. Using data points collected over shorter sample periods will enhance the applicability of the algorithm to a wide range of vehicles.

ACKNOWLEDGEMENTS

This research was supported in part by Allison Transmission, Inc.

REFERENCES

Athimethphat, M. and Lerteerawong, B. (2012). Binary classification tree for multiclass classification with observation-based clustering. In *9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, pages 1– 4.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chakraborty, A., Faujdar, N., Punhani, A., and Saraswat, S. (2020). Comparative study of k-means clustering using iris data set for various distances. In *10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 332–335.

Daengduang, S. and Vateekul, P. (2017). Applying one-versus-one svms to classify multi-label data with large labels using spark. In *9th International Conference on Knowledge and Smart Technology*, pages 72 – 77.

Duran, A., Phillips, C., Perr-Sauer, J., Kelly, K., and Konan, A. (2018). Leveraging big data analysis techniques for us vocational vehicle drive cycle characterization, segmentation, and development. Technical report, SAE Technical Paper.

Ester, M., Kriegel, H., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, 96(34):226–231.

Kanemaru, Y., Matsuura, S., Kakiuchi, M., Noguchi, S., Inomata, A., and Fujikawa, K. (2013). Vehicle clustering algorithm for sharing information on traffic congestion. In *13th International Conference on ITS Telecommunications*, pages 38–43. IEEE.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE.

Khalid, S., Khalil, T., and Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In *Science and Information Conference*, pages 372–378. IEEE.

McInnes, L. and Healy, J. (2017). Accelerated hierarchical density based clustering. In *International Conference on Data Mining Workshops (ICDMW)*, pages 33–42. IEEE.

Murphy, P. and Pazzani, M. (1991). Id2-of-3: Constructive induction of m-of-n concepts for discriminators in decision trees. In *Machine Learning Proceedings*, pages 183–187. Elsevier.

NREL (2019). Fleet dna project data.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *WIREs: Data Mining & Knowledge Discovery*, 8(4):1.

Scholkopf, B. and Smola, A. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shin, Y., Goh, Y., Lee, C., and Chung, J. (2019). Effective data structure for smart big data systems applying an expectation-maximization algorithm. In *Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4)*, pages 136–140.

Wahba, G. (2002). Soft and hard classification by reproducing kernel hilbert space methods. *Proceedings of the National Academy of Sciences*, 99(26):16524–16530.

Wang, J., Yuan, Y., Ni, T., Ma, Y., Liu, M., Xu, G., and Shen, W. (2020). Anomalous trajectory detection and classification based on difference and intersection set distance. *IEEE Transactions on Vehicular Technology*, 69(3):2487–2500.